K-12 Assessments and College Readiness: Necessary Validity Evidence for Educators, Teachers and Parents

A paper presented at the annual meeting of the National Council on Measurement in Education
New Orleans, LA

Catherine Welch

Stephen Dunbar

April 2011

K-12 Assessments and College Readiness:
Necessary Validity Evidence for Educators, Teachers and Parents

Abstract

The *Blueprint for Reform* places college and career readiness at the forefront of goals for
education reform, and positions growth as a critical aspect of assessment for accountability and
student learning. Growth information can provide families and educators with information they
need to help determine whether their students are "on track" for college readiness. Based on the
results of monitoring growth, interventions can be identified for the individual student, the
classroom or the school. This paper addresses the need for empirical validity evidence to
establish the connection between grade-to-grade progress of individuals and the concept of
readiness for postsecondary education. It also addresses how the appropriate form of evidence is
necessary to support the very different interpretations for parents and students, educators and
policy makers.

K-12 Assessments and College Readiness:
Necessary Validity Evidence for Educators, Teachers and Parents

An absolute priority of new large-scale assessment systems is the ability to measure student achievement against common college- and career-ready standards *(A Blueprint for Reform, 2010).* The Common Core State Standards (CCSS), which were the result of a voluntary effort to develop a set of evidence-based standards in English Language Arts and Mathematics essential for college and career readiness in a twenty-first century competitive society (CCSSO, 2010). The standards will help ensure that students are prepared for non-remedial college courses and will be prepared for training programs for career-level jobs. Given the high percentage of students who need remediation after high school to be ready for college-level courses (College Board, 2010), however, there is a need to identify students early on who are not on track to be college ready in order for the reform effort envisioned by the *Blueprint* to be successful.

Evidence related to student growth and trajectories targeting college- and career-readiness is critical in validating the assessment information that is proposed, for example, by the SMARTER Balanced Assessment Consortium (SBAC) and the Partnership for Assessment of readiness for College and Careers (PARCC). The SBAC application for its assessment system suggests that an important component of the validity evidence will be the extent to which summative results for each content area accurately measure whether students are on track or ready for college or a career (SBAC 2010). Beginning at the content level, their proposed blueprint is intended to provide sufficient data across the clusters of the CCSS to measure achievement (i.e., obtained proficiency level) and growth (i.e., both progress toward meeting grade-level expectations and progress toward the grade 12 exit criteria). SBAC suggests that the

summative assessment will have the ability to provide student achievement and growth data to measure college- and career-readiness by using an adaptive engine to sample items within grade level and above or below as necessary to provide precise measurement of the student's achievement level. Results of assessments, as translated by the vertically articulated content and achievement standards, will be expressed on the same common scale. The proposal suggests that the consortium will conduct external validity studies to measure whether students who achieve particular scores are appropriately prepared for college. PARCC proposes an assessment system that will produce the required student performance data (student achievement data and student growth data) that can be used to determine whether individual students are college- and career-ready or on track to being college- and career-ready (PARCC 2010).

The *Blueprint* presupposes that the evidence to support uses and interpretations related to college and career readiness will be available when proposed new assessments become available. It is the responsibility of test developers and educators to ensure that a comprehensive approach to the collection and examination of validity evidence is an integral part of these new assessments. The validity of the interpretation and use of scores from educational and psychological measurement is "the most fundamental consideration in developing and evaluating tests" (AERA, APA, NCME, 1999, p. 9). An extensive framework has developed to support collecting evidence and marshaling rationales for a validity argument (Messick, 1989; Kane 2006). This framework extends to scores for college- and career-readiness measures. With the assessment imperative of college- and career-readiness at the forefront of efforts to reform education, a critical aspect of validity arguments for new assessments involves the validation of approaches to measuring and reporting of growth. Conceptual frameworks for understanding student growth

are evolving rapidly, and interpretations of growth that are both criterion-referenced and norm-referenced are being implemented in statewide testing programs (Betebenner, 2008).

Accurate tracking towards college and career readiness and the appropriate use of this information relies on various types of validity evidence. This paper considers the role of traditional K-12 assessments of achievement in the readiness discussion. Can these assessments be used to be used to identify students that are on track for college and career readiness? If so, what are the appropriate messages for students, parents and educators with respect to interpretation and use of readiness information?

A Framework for Consideration

Content Validity

Content-related validity evidence is tied to test development. The proposed interpretations of growth and readiness should guide the development of a test and the inferences leading from the test scores to conclusions about a student's readiness. For inferences related to college- and career-readiness, content validity evidence demands clear, well-articulated and easily understood expectations that show a progression of learning across time. Assuming that the CCSS will support evidence of these progressions over time, content validity evidence will be identified in the match between the standards and the resulting assessments. Content alignment studies will serve as the foundation for a trail of evidence needed for establishing the validity of readiness reporting (Loomis, 2011). Alignment studies such as those identified by Loomis will inform the interpretation of readiness research findings from the statistical relationship studies and shape assessments that are making the claim to identify students who are college ready.

Webb's four criteria (2006) for specifying content of assessments designed to measure college and career readiness can be used to study the alignment between the CCSS and existing or yet-to-be-developed assessments. These criteria include categorical concurrence, depth-of-knowledge consistency, range-of-knowledge correspondence and balance of representation. These criteria provided guidelines for the creation of the specifications for alignment activities.

- Categorical concurrence is demonstrated by the occurrence of the same content categories between the content standards and the reportable categories in the specifications.

- Depth-of-knowledge consistency is demonstrated by the inclusion of items that represent the varying levels of complexity of the content standards.

- Range-of-knowledge correspondence is demonstrated by having an adequate number of items associated with each of the content standards.

- Balance of representation is demonstrated through a consistent number of items by standard within grade.

Criterion-related Validity

Beyond the content perspective, there are at least three basic requirements necessary to address criterion-related validity evidence. These include a scale to support tracking, a target to aim for, and the collection of validity evidence to support the resulting inferences.

Scales

Scales or linking studies that allow for the longitudinal estimation and tracking of growth towards readiness are a necessity for measuring growth in the present context. Vertical scales facilitate the estimation and tracking of growth over time, as repeated measures on individual students using different, grade-appropriate test forms becomes possible. This helps to determine how much growth has occurred over time. Vertically-scaled assessments also allow comparisons

of one grade level to another and one cohort of students to another at any point in time (Patz, 2007).

Vertical scales that have been well constructed for use in large-scale educational testing programs would include a number of defining technical characteristics (Patz, 2007), including:

1. an increase in difficulty of associated assessments across grades,

2. an increase in scale score means with grade level, and

3. a pattern of increase that is regular and not erratic.

Kolen and Brennan (2004) include a discussion of ways to evaluate the appropriateness and integrity of vertical scaling results. Such criteria and guidelines will be necessary to ensure that the vertical scales being designed to measure growth possess the psychometric characteristics necessary to accomplish the task.

Targets for Growth and Readiness

Targets must exist that quantify the level of achievement where a student is ready to enroll and succeed in credit-bearing, first-year postsecondary courses. To date, these targets are currently defined by the ACT Benchmarks (ACT, 2010), by the College Board Readiness Index (College Board, 2010), or by individual institutions of higher education. Regardless of the research base to support these targets, they are typically predicated on a criterion measure of success (e.g. grades of B or C in entry-level college courses) and an empirical relationship between prior assessment information, an admissions test score, for example, and success in college. Higher education institutions have long considered these connections in their admissions policies. The extent to

which empirical evidence for the connection between college success and assessment information can be established early in a student's education is of interest.

Collection of Evidence

Many tests will claim to measure college readiness, but a plan must be in place for validating that claim.   For example, validity studies may be conducted to determine the placement accuracy of students into entry level college credit coursework and remedial courses. Conley (2007) emphasizes the importance of academic behaviors that influence graduation and persistence as another source of validity evidence.   Camara (2010) suggests that the most immediate concern should be performance in specific courses such as Western Civilization or Chemistry. Suffice it to say that the validity of college readiness metrics will rely on a variety of sources of evidence that demonstrate the relationship between the measure and subsequent performance.

<div align="center">

An Empirical Example

</div>

This section considers the role of a traditional K-12 assessment of achievement in the readiness discussion. Can an assessment, such as the Iowa Tests, be used to be used to identify students that are on track for college and career readiness? If so, what evidence is there to support such an inference. If there is adequate evidence to support this inference, what are the appropriate messages for students, parents and educators with respect to interpretation and use of readiness information?

Evidence of Content Validity

The Iowa Tests represent a continuum of achievement that measures student progress from kindergarten to grade 12. The tests measure achievement in core academic areas important for success in college including reading, language arts, mathematics and science. The most recent

forms of these assessments have been carefully designed using the *Common Core State Standards Initiative*, individual state standards, surveys of classroom teachers, reviews of curriculum guides and instructional materials and response data from students during field testing. For example, the five content domains in the *Common Core Standards* for 6th grade mathematics are consistent with five of the content strands found in the 6th grade mathematics assessment of the Iowa Tests. The content devoted to each of these domains on The Iowa Tests is illustrated in Table 1.

Table 1. Summary of Categorical Alignment between CCSS Domains and Iowa Tests.

Iowa Tests										
	Grades									
Domains of the Common Core	K	1	2	3	4	5	6	7	8	High School
Counting and Cardinality	✔									
Operations and Algebraic Thinking	✔	✔	✔	✔	✔	✔				
Number and Operations in Base 10	✔	✔	✔	✔	✔	✔				
Number and Operations—Fractions				✔	✔					
Measurement and Data	✔	✔	✔	✔	✔	✔				
Geometry	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
Ratios and Proportional Relationships							✔	✔		
The Number System							✔	✔	✔	
Expressions and Equations							✔	✔	✔	
Statistics and Probability							✔	✔	✔	✔
Functions									✔	✔
Number and Quantity										✔
Algebra										✔
Modeling										

Evidence of Criterion-related Validity

Scale

The Iowa Tests were developed using standard scores that describe a student's location on an achievement continuum, much like a learning progression for a broadly defined content domain. Expectations for a student's annual growth (beginning at any point on the scale) can be established based on intervention and instructional strategies. The Iowa scale tracks year-to-year growth and compares student expectations to achieved growth. The score scale is a vertical scale

that quantifies and describes student growth over time. The current vertical scale, developed by Iowa Testing Programs in 1992, is psychometrically sound, has been used extensively at the district and state level and meets the technical requirements of large scale assessment (APA, AERA, NCME Standards, 1999).

Target for Growth and Readiness

Scores on individual Iowa Tests were mapped to defined targets of readiness to determine preparedness in English, mathematics, reading and science. Validity studies such as those described below have been completed to map these indicators to the ACT Benchmarks (2010).

Collection of Evidence

Evidence of a strong relationship between Iowa test scores and a key indicator of college readiness (ACT Composite) suggests that the Iowa tests and college readiness measure the highly correlated if not comparable achievement domains. Based on a matched cohort of over 25,000 students who tested annually from 2003 to 2008, this relationship continues and strengthens from 5th to 11th grade. Figure 1 summarizes the correlation for this matched group of students from grades 5 to 11.

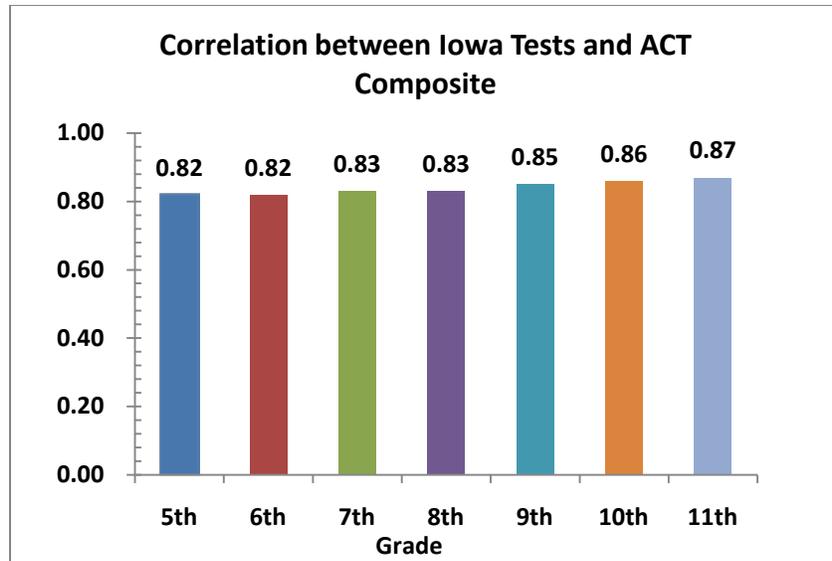**Correlation between Iowa Tests and ACT Composite**

Figure 1. Correlation between Iowa Tests and ACT Composite

Table 2 presents correlations between the corresponding Iowa test and *ACT* subtests in grades 8 to 11 in the matched sample referenced above. Each correlation is based on the number of students who have both an *ACT* score in the appropriate content area and Iowa test score in both the content area and grade of interest (Furgol, Fina, & Welch, 2011). These correlations are generally highest in grade 11, ranging from .68 (Science) to .76 (English and Math), providing supporting evidence for the use of the grade 11 Iowa tests to predict whether students are likely to meet or exceed the *ACT* College Readiness Benchmarks. The correlations adjusted for restriction of range (using the variances for the *ACT* subsample) are higher at .81, .82, .83, and .76 for Reading, English, Math, and Science, respectively, and thus, further support the use of these data for determining college readiness benchmarks. Moreover, even the unadjusted correlations between the grade 8 Iowa content area tests and the corresponding *ACT* tests are in the same neighborhood as those between corresponding content area tests on *EXPLORE* and

ACT, which are .75 for English, .73 for Math, .68 for Reading, and .65 for Science (ACT, 2007,

p. 45).

Table 2. Observed Correlations between *ACT* and Iowa Tests, by Content Areas

Grade	Reading	English	Math	Science
11	.75	.76	.76	.68
10	.72	.79	.75	.67
9	.75	.76	.74	.65
8	.74	.72	.75	.60

To help illustrate how these relationships can be useful in predicting college readiness, the

bivariate distribution of Iowa and ACT content area scores was used to establish a benchmark

score on the Iowa vertical scale that corresponded to the ACT college readiness benchmark. The

criterion for defining the Iowa benchmark score was based on the concept of error rates and

classification rates corresponding to equal sensitivity and specificity rates in decision making.

Sensitivity here refers to the estimated probability of correctly judging a student to <u>not</u> be

college-ready (i.e. falling short of the ACT benchmark) on the basis of the Iowa achievement test

score. Likewise, specificity refers to the estimated probability of correctly judging a student to be

college ready (i.e. exceeding the ACT benchmark). Figure 2 illustrates the determination of the

Iowa standard score points that correspond to the ACT college readiness benchmarks in four

content areas, assuming an application of an equal error rate method (Furgol, Fina & Welch,

2011). For example, the resulting cut point on the Iowa scale corresponding to an ACT college

readiness benchmark in mathematics of 22 is 312. Given the bivariate distribution of the Iowa

and ACT assessments in mathematics, 312 is the Iowa scale score that equalizes the proportion

of false positive and false negative judgments of college readiness.

Table 3 (Furgol, Fina, & Welch, 2011) provides the classification rates using these cut scores. This table shows that all of the cuts correspond to sensitivity and specificity rates of about 80 percent and false positive and false negative error rates of about 20 percent.

Table 3. Classification Rates for the Grade 11 Iowa Content Areas Using the Cutscores from the Equal Error Rate Method

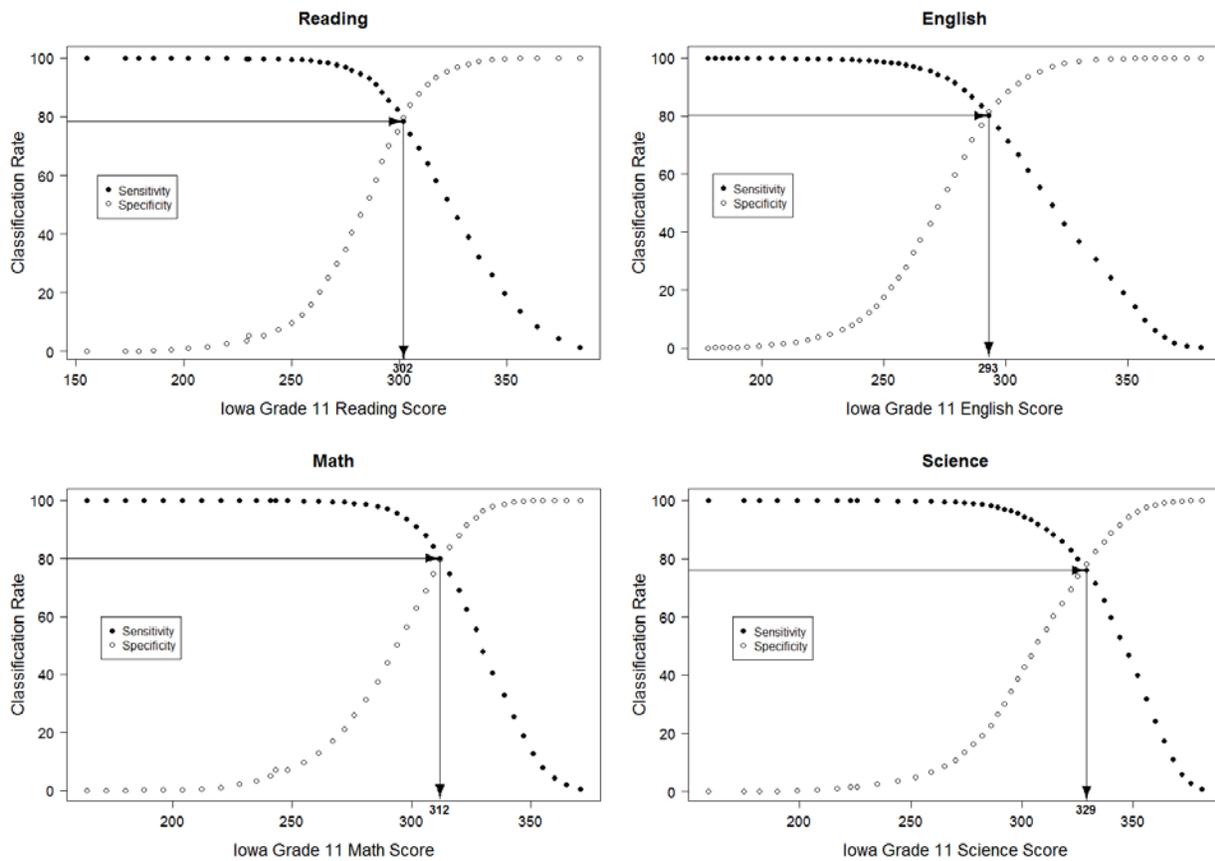| Content	Specificity	False Positive Rate	Sensitivity	False Negative Rate
Reading	79.69	20.31	78.51	21.49
English	81.49	18.51	80.19	19.81
Math	79.56	20.44	80.10	19.90
Science	78.15	21.85	76.06	23.94

Figure 2. Application of an equal error rate to determine cutscores on Iowa Tests (Grade 11), by Content Area.

Combining the Vertical Scale and the Correlations to Produce "On Track" Indicators

One approach to using this information is to take advantage of the information in the vertical scale of The Iowa Tests to make predictions about being on track to college readiness. The properties of the vertical scale underpinning The Iowa Tests afford a unique method to determine cut scores on the grade level achievement tests prior to 11th grade. The National Percentile Ranks (NPRs) for the grade 11 Iowa cut scores can be used in linking back to earlier grades to convey "on track to college readiness" messages.

The "linking back" procedure through the Iowa vertical scale is easily executed. The grade 11 Iowa content area cut scores found using the equal error rate method are 293 for English, 302 for Reading, 312 for Math, and 329 for Science. These scale scores correspond to national percentile ranks (NPRs) of 64 for English, 74 for Reading, 81 for Math, and 87 for Science (Forsyth et al., 2003). It is interesting to note that the order of these NPRs is the same as the order of the *ACT* college readiness benchmarks of 18, 21, 22, and 24 for English, Reading, Math, and Science, respectively. Thus, it is not surprising that students who score at or above only 64 percent of the national norming sample are predicted to be college ready in English, whereas students have to score greater than that of 87 percent of the national norming sample to be predicted to be college ready in Science.

Figure 3 illustrates how the linking back procedure can identify the appropriate cut scores to generate "on track" messages back to grade 5. The figure shows scale scores required at each grade level to achieve the same relative standing within grade as the Iowa scale score corresponding to the ACT college readiness benchmark in mathematics.
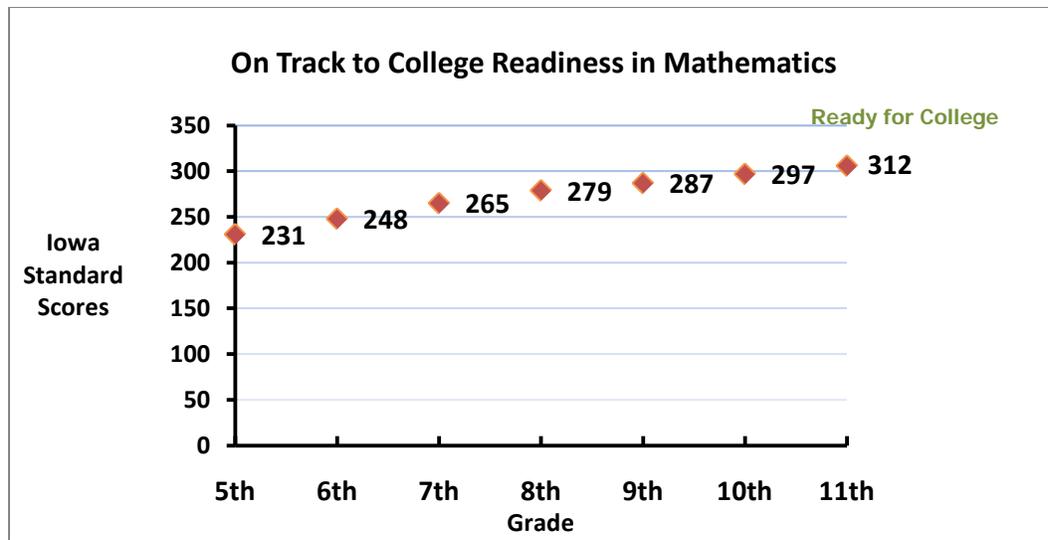
On Track to College Readiness in Mathematics

Figure 3. Standard Score Equivalents of Readiness Indicators from Grades 5 to 11 on the Iowa Tests.

Providing Information to Parents, Teachers and Students

Readiness information gives families and educators information to determine whether students are on track and where additional coursework and preparation should be considered. It allows families and educators to monitor student progress from elementary school through high school and allows flexibility to determine the appropriate improvement and support strategies for students. Information such as that provided in the Iowa example should be combined with other available information to help students develop realistic goals and educators plan relevant educational interventions.

The target could be a college readiness benchmark expressed in terms of the vertical scale of the assessment, a goal defined by expected growth given the student's reading score on the first measurement occasion, or any other performance standard that could be associated with the vertical scale of the assessment. Assessment results each year are depicted with confidence bands. If the band covers the trajectory, the implication is that the student is "on track" toward

reaching the target; otherwise, the band appears in red and a "not yet on track" message is indicated. In terms of the validity argument, the trajectory toward the target must be validated as well as the intermediate assessment results that are used to indicate whether the student is "one track" or "not yet on track" with respect to the target.

References

ACT. (2010a). *College readiness standards: For EXPLORE, PLAN, and the ACT.* Retrieved July 5, 2010 from http://www.act.org/standard/pdf/CRS.pdf

ACT. (2010b). *The condition of college and career readiness 2010.* Retrieved November 9, 2010 from http://www.act.org/research/policymakers/cccr10/pdf/ConditionofCollegeandCareerReadiness2010.pdf

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

Betebenner, D.W. (2008). Toward a normative understanding of student growth. In K.E. Ryan & L.A. Shepard (Eds.), *The future of test-based accountability* (pp. 155-170). New York: Taylor & Francis.

College Board. (2010). *The development of a multidimensional college readiness index*. College Board Research Report No. 2010-3. Retrieved June 29, 2010 from http://professionals.collegeboard.com/profdownload/pdf/10b_2084_DevMultiDimenRR_WEB_100618.pdf

Council of Chief State School Officers [CCSSO] & National Governors Association [NGA] Center for Best Practices. (2010a). *The standards: English language arts standards*. Retrieved from: http://www.corestandards.org/the-standards/english-language-artsstandards.

Council of Chief State School Officers [CCSSO] & National Governors Association [NGA] Center for Best Practices. (2010b). *The standards: Mathematics*. Retrieved from: http://www.corestandards.org/the-standards/mathematics.

Forsyth, R.A., Ansley, T.N., Feldt, L.S. and Alnot, S.A. (2003). *Norms and score conversions: Iowa Tests of Educational Development*. Itasca, IL: Riverside Publishing.

Furgol, K. Fina, A. and Welch, C. (2011, April). *Establishing validity evidence to assess college readiness through a vertical scale*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Kane, M. T. (2006). Validation. In R.L. Brennan (Ed.), *Educational Measurement (4[th]ed.).*Westport, CT: American Council on Education/Praeger.

Loomis, S.C. (2011, April). *Toward a validity framework for reporting preparedness of 12th graders for college-level course placement and entry to job training programs*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational Measurement (3rd ed.).* New York: America Council on Education and Macmillan.

Partnership for Assessment of Readiness for College and Careers. (2010). *The Partnership for Assessment of Readiness for College and Careers (PARCC) application for the Race to the Top Comprehensive Assessment Systems Competition*. Retrieved from http://www.fldoe.org/parcc/pdf/apprtcasc.pdf.

Patz, R.J. (2007, January). *Vertical scaling in standards-based educational assessment and accountability systems*. Washington, DC: Council of Chief State School Officers. Retrieved April 4, 2011 from http://www.ccsso.org/Documents/2007/Vertical_Scaling_in_standards_2007.pdf

SMARTER Balanced Assessment Consortium. (2010). *Race to the Top Assessment Program Application for New Grants Comprehensive Assessment Systems.* Retrieved from http://www.k12.wa.us/SMARTER/pubdocs

U. S. Department of Education. (2010). *A Blueprint for Reform. The Reauthorization of the Elementary and Secondary Education Act.* Retrieved June 26, 2010 from http://www2.ed.gov/policy/elsec/leg/blueprint/blueprint.pdf