Defining and Measuring College and Career Readiness and Informing the Development of Performance Level Descriptors (PLDs)

Wayne Camara

The College Board

and

Rachel Quenemoen

National Center on Educational Outcomes

Paper Commissioned by PARCC

January 5, 2012

The authors would like to thank members of the PARCC technical advisory committee for their comments and review. We are especially grateful to Scott Marion, Jeff Nellhaus, Barbara Plake, Kris Ellington, Ronald Hambleton, Kristen Huff and Mary Ann Snider for comments and advice on earlier versions of this paper.

Executive Summary

PARCC is a consortia of states that is developing a next-generation assessment system in English and math anchored in what it takes to be ready for college and careers. To accomplish this goal the consortia must determine whether individual students are college-and career-ready, or are on track. A direct empirical relationship between PARCC assessment scores and subsequent success in college and career training provides the strongest form of evidence.

This paper reviews many criteria that can be used to gauge college and career success but argues that student academic performance (e.g., grades, GPA) in credit bearing courses is the most relevant and available criteria for defining success. College and Career Readiness can be conceptually defined as including multiple factors, but consortia assessments should be more narrowly tailored to a definition which is based on the cognitive skills and content knowledge required in the common core standards and types of learning which occurs in schools and classrooms.

There are alternative approaches to establishing performance level descriptors (PLDs), cut scores and metrics that will be used to determine if students are college-and career-ready. Judgmental approaches have generally been used in state assessment programs, but since scores from these assessments will primarily be used to make inferences about future performance, empirical methods (e.g., prediction models, regression, linking scores across assessments) traditional used in admissions and placement testing programs are of greater relevance (Kane, 2001).

The paper describes a general validation approach and the required evidence to conduct predictive studies between PARCC secondary assessments postsecondary success. The progression and coherence of PLDs should be based on empirical data from the statistical links between high school assessments and college and career outcomes, as well as, educators' judgments from content-based standard setting

approaches. The paper provides examples of PLDs based on statistical data and postsecondary outcomes, and nine major recommendations for establishing a validity argument for consortia assessments.

College and Career Readiness: Informing the Development of Performance Level Descriptors (PLDs)

Wayne Camara and Rachel Quenemoen

The purpose of this paper is to assist PARCC in developing a working definition of college and career readiness (CCR) which can be used to: (a) establish an interpretative argument for the primary inferences that will be made from test scores; (b) determine CCR in high school and ascertain whether students are 'on track' toward CCR at lower grades; (c) aid in collecting validation evidence of CCR metrics, PLDs and cut scores; (d) determine the criterion associated with CCR; (e) guide the development of performance level descriptors (PLDs) early in the design of assessments; and (e) clarify expectations for item and test development.

Determining whether individual students are college- and career-ready, or on track, is a central objective of PARCC's assessment design. The PARCC Governing Board has established that test scores will be used to make inferences about the CCR of students, and validation evidence is required to support these inferences and resulting decisions (Kane, 2001). PARCC has also chosen an evidenced centered design (ECD) approach for the design and development of assessments. It is important to establish performance level descriptors (PLDs) at the initial stages of the assessment design and development work, because in ECD the PLDs will drive the validity and interpretative arguments (Kane, 1994). Developing PLDs early in this process is also required to ensure coherence of PLDs across grades. Given the intended purposes of the PARCC assessments (i.e., CCR and on track for CCR), the PLDs should be anchored in both the definition of CCR and the Common Core State Standards (CCSS). PLDs should are an important component of the validity argument and influence standard setting (Kane 2001).

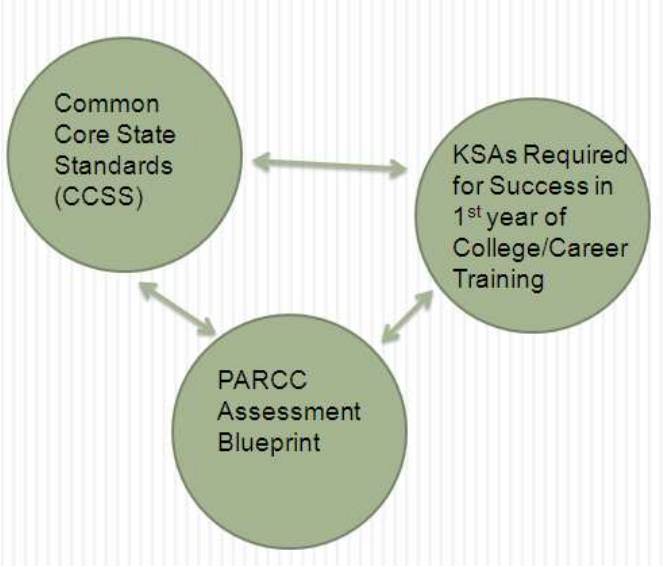
In determining whether students are prepared or ready to successfully undertake college or career training programs, direct evidence between test scores and subsequent college and career training

success may be the strongest form of evidence. PARCC assessments will be used for multiple purposes, but a primary purpose is to determine if students who score above a cut score are ready or prepared to succeed in post-secondary education. The explicit implication is that there should be a strong statistical relationship between performance on PARCC assessments, particularly high school assessments, and subsequent postsecondary success. PARCC assessments will be used to determine proficiency and readiness in a similar way that placement and certification tests are currently being used. Certainly, a weak relationship between scores on CCR assessments and subsequent postsecondary success would be reason to question the validity of inferences associated with the assessments[1]. Therefore, this paper begins with a brief review of college- and career-training success.

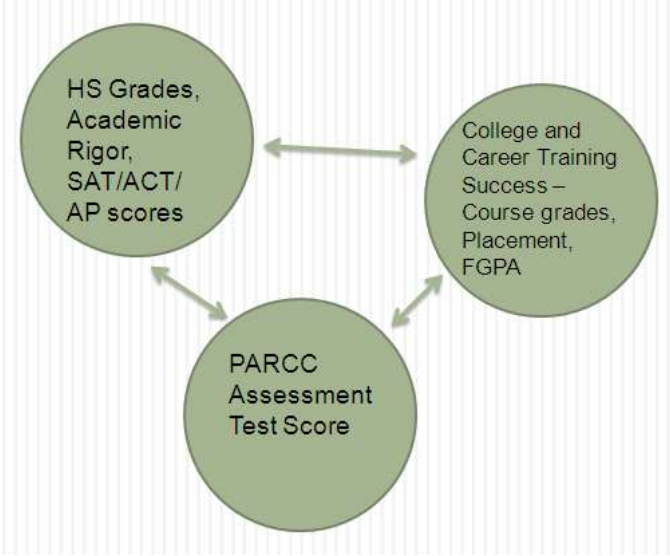
Figure 1 illustrates the implicit validity argument for PARCC assessments which are grounded in both the CCSS and empirical data which is a direct measure of student success in college or career training programs. The assumption is that the CCSS capture the prerequisite content and skills for entry level credit bearing courses in college and postsecondary career training programs. Figure 1 further illustrates that empirical evidence is critical in providing a validation argument to support the intended interpretations of test scores in the PARCC high school assessment. PARCC's state leadership must specify the criteria for college success in terms of the construct (e.g., course grades, FGPA, placement in credit bearing courses), performance level (e.g., 2.0) and probability (e.g., 50%, 70%). Ultimately the test benchmark and cut score should be related to the criteria as well as the CCSS. A validation strategy employing multiple methods and a variety of evidence (some of which is empirical, but also including judgments) is developed to support implied relationship between these three pillars of the College and Career Readiness conceptual argument.

[1] For example, if the majority of students scoring above a CCR cut score failed or a majority of students scoring below a CCR cut score succeeded, there would be significant concern and caution about using scores to support CCR inferences and decisions.

Figure 1 Validation framework for PARCC



Conceptual framework for validation argument



Empirical pathway for validation argument

College Success: Construct and Criteria

Much of the interest and attention with college- and career-readiness has resulted from the relatively high remediation and low completion rates in post secondary education reported for students with a high school diploma who have succeeded in high school courses and on state assessments. A variety of criteria[2] have been proposed as measures of post secondary success:

1. Persistence and successive completion of courses resulting in a certificate or degree (e.g., persistence to second year)
2. Graduation or completion of a degree or certification program
3. Time to degree or completion of a certification program (e.g., six-year graduation for a Bachelor's degree).
4. Placement into college credit courses
5. Exemption from remediation courses
6. Grades (and performance) in specific college courses, typically taken in freshman year (e.g., college algebra, freshman composition).
7. Grade-point average (GPA) in college which can also be described as successful performance across a range of college courses

[2] Other criteria have been used to evaluate the effectiveness of postsecondary education and may include both economic factors (e.g., starting salary), employment, life events (e.g., divorce), and self reported attitudes and behaviors (e.g., interest in lifelong learning, community service) which do not appear relevant as primary criteria of cognitive assessments.

Clearly, these criteria are highly related, but not mutually exclusive. It is quite possible for students to require remedial courses but still graduate and attain good grades, just as it is possible for students to attain high test scores, require no remediation, succeed academically and drop out of college. So what definition of college- and career-success should PARCC adopt? This is not a trivial question, because the assessment design and validation argument must be tied to the definition of college- and career-success. The following section will briefly review each of these potential outcomes. It is also important to distinguish between definitions of CCR which are constrained to academic elements and those definitions which extend beyond academics. Since CCR assessments will only measure mathematics and English Languages Arts (ELA) the relationship to non-academic criteria may be less defensible and a variety of individual and institutional factors may more influential as a moderator (in a statistical sense and from a validity argument).

Persistence, graduation and time to degree

Several criteria such as persistence, graduation, and time to degree are extremely important when evaluating an educational system, yet they are heavily influenced by a host of nonacademic factors that are generally not directly measured by cognitive ability tests and not included in the CCSS. While such broad criteria are often attractive to policymakers they are not a direct outcome of academic preparedness alone. Many factors have been shown to relate to such outcomes which would threaten the validity of statements we may wish to make about test scores and their impact on college success.

Financial factors. Finances are highly related to persistence, graduation and time to degree. Despite increases in enrollment rates among all racial, ethnic, and income groups, participation gaps between affluent student and those from less privileged backgrounds have persisted. Also, gaps in degree attainment are larger than gaps in enrollment because lower-income students who are able to

overcome the financial barriers to enter college are less likely to complete degrees (Camara, 2009). "Fewer than 40% of the academically high scoring/low SES students who do enroll in college earn bachelor's degrees" according to Baum and McPherson (2008, 5).

Institutional factors. "By itself, Carnegie classification level (e.g., institutional selectivity and prestigious as measured by external funding and research) has a profound effect on graduation rates" (Hamrick, Schuh and Shelley, 2004, 10). Institutional demographic characteristics, geographic region, institutional financial assistance, and per student spending are also related to graduation rates. **Social adjustment**, whether students integrate socially and academically in their institutions and their level of engagement, are also strong predictors of persistence (Tinto, 1987). Students are expected to devote time to their education and assume responsibility for investing time and energy in learning. Student motivation, attendance, and engagement in learning are related to outcomes of persistence and graduation (Tinto, 1987). A wide range of **psychological and social factors** have also been shown to impact persistence, graduation and time to degree. These factors include maturation, roommate conflicts, dating problems, health, nutrition, fitness, and time management (Harmston 2004; Purcell and Clark 2002; Robbins, Lauver, Le, Davis, Langley, and Carlstrom, 2004).

Research has consistently shown that **cognitive measures** of academic performance, such as high school grades and test scores, are highly predictive of grades earned in college, but less so of retention and graduation (e.g., Robbins, Allen, Casillas, Peterson & Le, 2006; Robbins, Lauver, Le, Davis, Langley, and Carlstrom, 2004; Schmitt, Billington, Keeney, Oswald, Pleskac, Sinha, et al. 2009). In fact, Robbins et al. (2004) found that the correlations between cognitive measures and first-year GPA were roughly two to three times larger than the correlations between cognitive measures and retention. Burton and Ramist (2001) reported that the combination of admission test scores, grades and academic rigor offer the best predictors of graduation, but the correlations are generally about half as large as those found in predicting college grades. Grades on the Advanced Placement Exams have been shown to be a superior

predictor of graduation when compared to admission tests, but they still placed second to high school grades even when the quality of high school was considered (Bowen, Chingos and McPherson, 2009). This research also reported that non-cognitive factors such as academic discipline, commitment, and persistence are related to graduation and persistence (Camara, 2005b; Robbins et al., 2004; 2006; Schmitt et al., 2009).

Persistence, graduation and time to degree are important outcomes of educational success, but there are influences other than academic preparation and cognitive ability that present challenges for PARCC. Many students attending two-year colleges or career and technical training may not be enrolled full time or in degree or certification programs. The relationship weakens further when you look at six-year graduation rates and control for institutional transfers and part-time status.

Placement and Exemption from Remedial Courses

There are limitations in selecting other available criteria as a measure of college success. **Placement into college credit courses**, and the associated exemption from taking remedial courses, is frequently cited as a principal objective of college- and career-readiness. Setting a college readiness benchmark that is associated with the knowledge, skills, and abilities (KSAs) required for entry level courses such as college algebra and composition can be accomplished through content-based and judgmental approaches (e.g., surveys, standard settings), but validation through empirical approaches will present greater challenges. Remediation and placement decisions are typically made with the same instruments. That is, students are generally required to take placement tests prior to matriculation at an institution, and their performance on such tests are used to determine: (1) whether they are placed in remedial or credit bearing courses, and (2) to determine a specific course placement (e.g., college algebra, precalculus/trigonometry). However, it is less likely that PARCC assessments would be used for

the latter purpose by a majority of institutions. A similar scenario exists today with some institutions that view a high score from admission or state tests as sufficient to waive remediation, but not adequate for placement in advanced courses (e.g., precalculus).

Placement tests are not only used to determine whether students require remediation, but also to determine the best placement for students. This is more likely to occur in math[3] than composition. A single CCR benchmark score will not be equally effective for differentiated placement across math courses because the level of math proficiency will differ across these courses. In addition, there are likely to be significant differences across institutions and academic departments for the same course[4]. Colleges and universities currently have very different entry level courses, remedial courses, and requirements for entry into the same and different courses (NAGB, 2009; Shaw and Patterson, 2010).

Placement into college credit courses without remediation is dichotomous, and studies of classification accuracy would be more appropriate than more traditional linear regression studies. Second, this criterion may present a challenge in mathematics where there is significantly more variability in freshmen course taking behavior. For example, 28 percent of students taking the SAT and attending a 4-year college did not take a math course during their freshman year at four-year colleges, and this number is likely to be much higher if extended to two-year colleges and career-training programs (Shaw and Patterson, 2010). Even among students taking math, there is significant variability in math courses completed. ACT (2007) reported that college algebra was the most frequently taken math course by college freshmen across two- and four-year institutions. Table 1 illustrates that 36

[3] It also is true of placement in foreign languages and within course sequences in STEM areas for some selective colleges.

[4] The requirements and expectations for success in calculus in an Engineering department may differ from those associated with calculus taught in a Business or Social Science department within the same institution. Research demonstrates that the average college grades differ significantly across departments.

percent of students in ACT's college readiness benchmark study took college algebra, but 64 percent of students either took another math course or did not enroll in math. Adelman (2006) used longitudinal samples and reported on math courses completed by freshmen.

Table 1 illustrates that basing a CCR benchmark on specific courses could exclude nearly ¾ of college going students. Some universities have no remedial courses, but a significant percent of students will still receive a grade of C- or below in their freshman math course. These students met the existing requirement for the placement into a credit bearing course, but may not have been college ready.

It is quite appropriate to use multiple methods to arrive at an educationally defensible and empirically validated benchmark score that corresponds to a decision about remediation or college credit across institutions. PARCC assessments could serve this function for students transitioning directly from high school to post secondary education. In addition, PARCC assessments might serve as a placement tool for English composition where there are many fewer entry level courses than in mathematics. It is less likely that PARCC assessments would be a viable option to replace comprehensive placement test that are used to determine student placement across a variety of mathematical courses with differing requirements and rigor. Arriving at a consensus about a CCR benchmark that would be acceptable to 20 or more states and hundreds of different institutions will not be a simple task. In fact, institutions which conduct their own local validation studies may find that different cut scores offer the greatest utility and efficiency because of significant differences in their enrolled students.

A common CCR benchmark could produce significantly different remediation outcomes across different types of institutions, but should produce similar impact among institutions where student selectivity and characteristics are comparable. For example, if we used a common cut score on an admissions test (e.g., SAT math 440 or ACT math 18), we would find smaller differences in the

percentage of students below that score among comparable schools (e.g., state flagships, community college systems, the typical four-year public institution).

Many students who are placed into college credit courses may fail, and conversely, many students who are placed into remedial courses might have been successful in an entry level course. If placement is the primary criterion for CCR, what would happen if subsequent research demonstrated that a significant percentage of these students did not succeed in the course? Such research will certainly be conducted, and it is quite possible that such research would show that different cut scores are needed across different institutions, different courses, and different departments.

There are several risks adopting a single definition of CCR and cut score that is based primarily on decision consistency across post secondary education:

Table 1

Percentage of students taking the ACT or SAT and completing math courses during their freshman year in college

Math Courses During Freshman Year	ACT – ACT, 2007: Allen & Sconing, 2005 n=80,000 (90 institutions)	SAT - Wyatt et al., 2011; Shaw and Patterson, (2010 n= 164,331 (110 institutions)	Adelman (2006)
Any Math Course	Not reported	72%	55%
Calculus*	Not reported	34%	18%
College Algebra*	36%	18% (22%[5])	22%
Statistics*	Not reported	10% (13%[6])	5%
Pre-calculus*	Not reported	9%	19%

*percentages based on total students completing any math course

[5] An additional 3% took a course labeled as ”algebra/trigonometry.”

[6] An additional 3% took a course labeled as ”probability/statistics”; 1% took a course labeled as “business statistics.”

- Will higher education actually use a common cut score for placement out of remediation? Will the score be too low for some institutions and too high for other institutions? Will some institutions and departments insist on their own higher CCR score for PARCC assessments (especially in math courses)? If higher cut scores are still imposed by moderately selective institutions would that benchmark be credible?
- Will institutions waive placement tests for students attaining the CCR benchmark or will students be required to take a second institutional placement test?
- What would be the impact of a CCR benchmark if some institutions demonstrate that a significant percentage of students reaching that benchmark do not succeed in entry level courses?
- Is it possible to get state systems to accept one standard (one cut score on the same test) immediately when the test becomes operational or is a phased in approach more viable?
- Do career and vocational training programs have remedial courses and what parallel data could be used for such programs?

College Grades

ACT and the College Board have set CCR benchmarks using academic criteria as opposed to placement decisions or measures that also include student persistence. The advantages of using a purely academic criterion include:

- A strong statistical relationship between the predictor (test scores) and criteria (grades) (r=.50 to .62 adjusted for restriction of range)
- A criterion that minimizes construct irrelevant variance (e.g., higher educational decision consistency, many of the factors noted above)
- A criterion where data are more easily available

- An outcome that appears logical and rational
- Significant research investigating the relationship between test scores and this criterion exist.

There is an overwhelming body of research which uses college grades and GPA as the primary criteria of college success[7]. This paper will not review this literature, but whether one uses course grades or combined grades the relationship with cognitive ability tests is likely to be quite strong. The major limitation in using grades or GPA they are often considered a proximal measure of the ultimate criteria – graduation. But as noted above, grades are a more direct measure than other available criteria.

Other issues arise when using grades as a criterion. For example, restriction of range (in the predictor), criterion unreliability and differences in course taking patterns of students are issues that have to be addressed. While statistical adjustments have not been without controversy, failure to account for such issues can result in underreporting of validity results (Berry and Sackett, 2008; Sackett, Borneman and Connelly, 2009). Also, grades should be combined within an institution then aggregated to control for differences in grading practices and institutional samples (Kobrin, Patterson, Shaw, Mattern, and Barbuti, 2008).

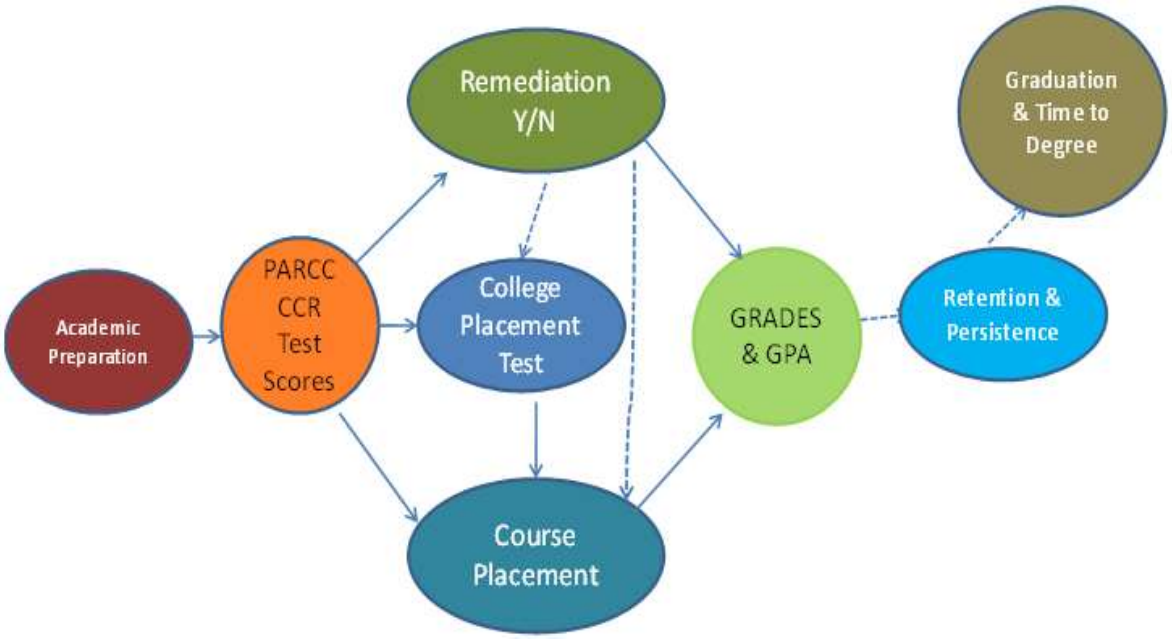
Figure 2 illustrates the temporal relationship between PARCC high school assessments and potential criteria of college success, as well as the relationship among criteria and placement tests. PARCC will likely be asked to provide evidence of the relationship of high school assessments and CCSS with each of the college success criteria discussed above (as well as many other criteria not addressed). However, it will be most important to specify, in advance, the primary criteria which should be used in

[7] FGPA is the most popular criterion measure for research on admissions, but other criteria used in studies include cumulative or final GPA, GPA in major, and grades in specific courses (Camara, 2005b).

establishing empirical validity evidence. If there is a weak relationship between the predictors (i.e., Common Core assessments) and the criteria it could be due to the standards, predictors or both.

Figure 2

Chronology of CCR Assessments and major milestones in college and career success



Available data for validation studies

The availability of data for validation studies should also be considered in developing an interpretative argument for CCR. Most states are now developing K-20 data systems that promise to allow researchers to track individual students who participate in the public K-12 system to public postsecondary education institutions within the state. PARCC should explore the feasibility and costs of several different options for gaining access to such data in the next year. Such an exploratory study could be done in cooperation with Smarter-Balanced and follow the process recently used to identify technology requirements across state K-12 schools. Specifically, such a study should determine the level and type of matched student data available in each PARCC state (e.g., course grades, full time status, transfer data), the technical and legal requirements to match PARCC student level assessment data and

external data (e.g., national test scores, outcome data from other states), and the general level of effort that might be required.

Public school data will come from participating states, but postsecondary data will be more problematic to collect and match to K-12 data. There are three types of sources of postsecondary data:

1. State K-20 data systems. State specific data will not be representative of undergraduates in the US. Nationally, 26 percent and 19 percent of students attending a four-year institution or two-year institution, respectively, are out-of state, and 9 percent of students attend a private college (NCES, 2010). However, even with such limitations, state specific data will still be the most comprehensive. States who have or will have integrated systems with student level data will have records from all students attending public colleges (two-year and four-year) and/or public school. Some states may also have data from other postsecondary career and training programs, students attending private or independent schools, and some home schooled students.
2. Cross state data. PARCC should explore the feasibility of matching data across states, especially where there are significant cross state college enrollments (e.g., New Jersey and New York, Massachusetts and Rhode Island). Certainly a centralized data warehouse across PARCC states would be ideal, but individual state level agreements are more feasible.
3. External data. There are a few potential sources of national data. Both the College Board and ACT collect college outcome data from some number of institutions, and PARCC should explore the feasibility of establishing cooperative agreements to access data for validation studies. The National Student Clearinghouse (NSC) maintains enrollment records for approximately 90 percent of postsecondary institutions. This data is limited to enrollment, status (full time or part time) and may have date of degree, degree type, and major for a subset of institutions. This data will not be useful in determining placement or grades, but can be of value in following

students and tracking enrollment, retention, transfers, and graduation. In addition, institutions listed as participants may not have submitted data in particular years[8]. PARCC should also begin to explore the requirements to establish data licenses with the largest career-technical training programs and schools in participating states. State officials may be of assistance in exploring such agreements.

Acquiring college and career outcome data will be complex and costly in many instances. There are four areas where missing data will present the greatest challenge to empirical validation studies:

1. State colleges where K-20 data systems cannot support the type of validation studies required
2. States where community college data are not incorporated or there are other flaws in such data
3. Private colleges
4. Career training programs

Potential strategies can be considered to acquire data in these areas. The absence of national and state outcome data on career training programs may present the greatest challenge. In selecting criteria, it is important to consider several factors such as feasibility (are data readily available and how difficult will it be to collect consistent data across institutions and states?), the relationship between the predictors and criterion measures (are there other important factors that would likely impact the relationship between test scores and the criteria, and can they be measured and accounted for in statistical models?) and demand (which criteria are of greatest concern to key stakeholders?). Preliminary recommendations about college- and career-readiness criterion measures:

[8] The College Board has often found some institutions do not provide NSC data annually. These institutions are listed as participating, but their data for the entering class of 2007 may not be provided until 2010 along with data from the 2008-2009 cohorts.

- Attempt to establish a PLD that corresponds to CCR ("on track to attain CCR" in lower grades) using multiple methods (described below). This could serve as the CCR benchmark[9]. An important criterion measure would be the percentage of students succeeding academically in their first college credit courses who score at or above the benchmark. Decision consistency and classification accuracy would need to be estimated for different types of institutions. Rather than specifying one course in math (e.g., college algebra) which could exclude the majority of students, an attempt could be made to determine if a common composite score could serve as a benchmark for decisions about remediation or college credit, assuming many institutions would still use a separate placement test.
- First year grades – grades are a direct measure of academic success and could be considered the most important criteria for PARCC tests. A weak relationship between test scores and grades could threaten adoption and continued use by higher education, while a strong relationship will encourage institutions that are undecided to reconsider their use. It is likely that many institutions will insist on validity and fairness data before using new test scores from PARCC, and this is the most important line of evidence. There are three potential grade metrics PARCC may consider: (a) grades in specific courses (algebra, composition), (b) GPA (across all grades or academic subject grades), or (c) an aggregate criteria across math courses (combining performance across different entry level math courses). PARCC should engage higher education in determining the most appropriate metrics for college success and the requirements to obtain such data annually on all enrollees.

Related validity issues for students with disabilities

[9] Whether a single benchmark is appropriate for all post secondary institutions or should differ for colleges and careers is discussed below.

Students with disabilities by and large require the same knowledge, skills, and abilities for college and career as their typical peers. A distinction from other special populations must be made, however, to note that the goal of special education services is not to cure students of their disabilities. The goal of special education is to provide the services, supports, and specialized instruction – including identification and use of appropriate accommodations – so that they are able to go around the barriers of disability and achieve the same goals and standards (in IDEA language) as all other students. Adults with disabilities are successful in college and career with appropriate accommodations, and their rights to use accommodations do not end with the end of special education. As Burgstahler and Cory (2008) demonstrate in their edited collection of universal design practices in higher education, inclusion has arrived in higher education. Two federal laws protect people with disabilities in college and the workplace in receiving appropriate accommodations: Section 504 of the Rehabilitation Act and the Americans with Disabilities Act (as reauthorized in 2010).

A few examples are in order. People with low vision and blindness are easily recognizable by the types of accommodations they use, whether as typically is permitted on state large-scale assessments at the present, in post-secondary settings, or in the workplace. Still, approximately 43% of all students with disabilities in K-12 schools have specific learning disabilities, sometimes called an "invisible" disability, and use of accommodations for these types of issues have been controversial in large scale testing. Corporate executives, poets, and governors with dyslexia don't always broadcast their use of accommodations, but recent stories featured Richard Branson, founder of Virgin Atlantic Airways; Charles R. Schwab, founder of the discount brokerage firm; John T. Chambers, chief executive and chairman of board of Cisco; and Paul Orfalea, founder of the Kinko's copy chain; Phil Schultz, Pulitzer prize winning poet with a new book *My Dyslexia*; Connecticut governor Dan Malloy, (Bowers, 2007; Reitz, 2011; Schultz, 2011). If a test predicting college or career readiness or preparedness does not allow accommodations that successful adults use and thus systematically predicts poor outcomes for all

students with certain types of barriers, it may threaten not only the validity but the legality of the test if it is used in diploma, college admissions, or course placement decisions. (See Walsh (2011) for a recent settlement related to accessibility of examinations for post-secondary educational applications, under the ADA.)

College and Career Readiness vs. Preparedness

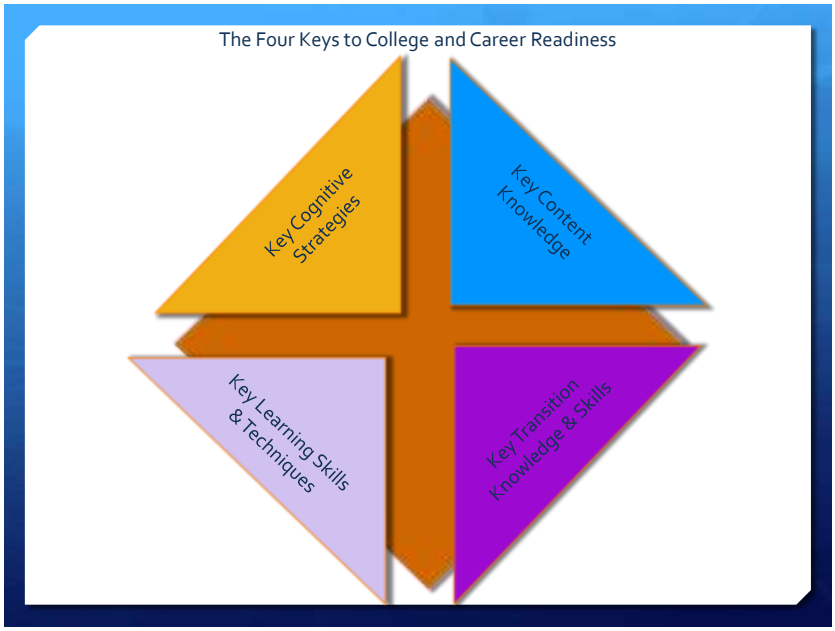
College and career success is ultimately the goal of current educational initiatives. Readiness or preparedness should help ensure greater success and improve the odds for students. Readiness is generally considered to be a synonym of preparedness in most disciplines. College and career readiness (CCR) should result in greater levels of success in both environments. Again, it is important to draw distinctions between definitions of CCR which focus solely on academic and cognitive factors from those which involve personality, predispositions, and other non-cognitive factors.

Conley (2011) defines four key dimensions of college readiness (see Figure 3). These dimensions are semi-independent, meaning a student may possess one or more dimensions and that to some extent they are all related to a successful transition to college. Most of what Conley describes is related to cognition, but some factors include elements which are not primarily academic or cognitive. **Key Cognitive Strategies** include cognitive strategies and higher order thinking skills (e.g., problem formation, interpretation, and analysis), and **key content knowledge** would most closely include the CCSS and other disciplinary documents which specify the concepts and knowledge in disciplines. The two remaining dimensions are not generally specified in state or disciplinary standards, but are related to college success. **Key learning skills** include time management, persistence, metacognition, goal setting, and self awareness. College environments generally require much more self management for students than high school courses. Students are increasingly responsible for accessing readings and submitting assignments through web interfaces, engaging in collaborative learning with other students,

and planning ahead to complete more complex assignments. **Key transition knowledge and skills** includes knowledge and awareness about the admissions and financial aid process, as well as how to interact with faculty and navigate college systems. Some of this information may be considered tacit knowledge and is often not obvious to students entering a new environment and culture. Some have referred to the latter two facets as *noncognitive* behaviors, but there are cognitive skills embedded in learning skills. Transition skills do include behaviors and skills which can be acquired but are less centered in cognition. For example, successful students must communicate and interact with a range of different types of individuals, as well as adjust to systems and organizations.

Figure 3

Conley's Keys to College and Career Readiness



D. Conley (personal communication, July 31, 2011)

These nonacademic Conley attributes of CCR are being discussed at length in the two federally funded projects to develop alternate assessments based on alternate achievement standards (AA-AAS), the Dynamic Learning Maps (DLM) Consortium at the University of Kentucky and the National Center and State Collaborative (NCSC) project at the University of Minnesota. Although both AA-AAS Consortia

are grappling with how to measure the CCSS in the large-scale assessment, reflecting the CCSS in claims, in high level PLD drafts, and test design choices, across consortia there is a commitment to be transparent on how the summative test relates to broader conceptions of CCR. These partners are discussing the use of high quality formative tools to augment the summative test for teacher evaluation and overall program improvement purposes throughout the year. Kearns et al. (2011) have done a careful job comparing and contrasting the Conley paper with best practice for students who participate in the AA-AAS, finding almost total overlap. It would be a good resource for a PARCC group working on participation policies, at a minimum, but could also inform working definition of CCR across the full assessment system, regular and alternate.

Many educators have used the terms college readiness or college and career readiness when only referring to academic or cognitive skills, which include both key cognitive strategies and content knowledge. The National Assessment Governing Board (NAGB) established a technical panel to determine how the results from the 12th grade National Assessment of Educational Progress (NAEP) could be used as a tool to report CCR.

NAGB and the technical panel defined preparedness as a subset of readiness (NAGB, 2009). College preparedness is defined as the academic knowledge and skills in required to qualify for placement into entry level college credit coursework without remediation. Preparation for workplace training refers to the academic knowledge and skills required to qualify for job training. This definition of preparedness does not mean the prepared student currently has the skills required to succeed in those entry level college courses or to be hired for a job. This is an important distinction. The panel is saying a student has the prerequisite knowledge and skills needed **to be placed** in a credit bearing course or training program (Loomis, 2011; NAGB, 2009). The assumption is also that such a student has the skills to access

the new knowledge and acquire the additional skills needed to succeed in the course or training program.

This definition would base CCR on skills students possess on Day 1 of college and not success during their first semester. This presents some challenges to in bringing empirical data from college to bear on the validity argument. There are no readily available measures of student knowledge as they enter college other than past performance in high school, scores on admissions tests (which are available for a self selected sample) and scores across a variety of placement tests (which again are available for a self selected sample and not easily obtained from high education or linked to other tests). Conceptually this definition may be attractive, but the lack of available empirical data is problematic. This is a particularly acute issue if we find a gap between the CCR benchmark set based on Day 1 readiness and ultimate success in college courses. NAGB which has adopted a similar definition of college preparedness appears to still employ college grades as a central criterion for predictive validity evidence.

This definition of preparedness appears consistent with Conley's Key Content Knowledge and Key Cognitive Strategies in reading and math, while excluding other facets not measured by NAEP. These first two sections have focused definitions of college and career readiness and preparededness, the relationship between these constructs and college success, and the types of criteria that have been employed across various efforts. The next section will provide a review of existing metrics of CCR.

Metrics of College and Career Readiness

Until recently, state standards and assessments have focused on 'challenging content standards' and performance standards which 'determine how well children are mastering the material in the State academic content standards' (No Child Left Behind Act, 2001). Current efforts to develop assessments which will be used to gauge the preparedness of students to succeed in college and career training

programs are fundamentally different and require strong empirical associations with relevant criteria. The major difference is in the interpretative argument and definition of the construct. CCR assessments are intended to be predictive of success and since college and career success can be measured there is an implicit assumption that performance standards should not be inconsistent with empirical evidence of success. Current state assessments have been developed to measure a state's content standards and no clear empirical measure has been established. Judgmental standard setting approaches are used to determine if the assessments measure the content standards and impact data may be used to help ensure consistency across grades and years. CCR is different because student success in college can be measured and empirical data are available. There is also an implicit expectation that there would be a common definition and common performance standard for CCR as opposed to allowing each state or post-secondary institution to define its own local standard for success and performance level (i.e., cut score on tests).

This paper does not argue that judgmental approaches cannot be employed, just that such approaches alone would be insufficient. If a cut score on assessments is established which contradicts empirical data of CCR it would not only lack credibility but would raise questions about the validity of the interpretative argument. Let's say that a typical judgmental standard setting process was used and resulted in a cut score of X for CCR. Now let's say that a meta-analysis was conducted which showed that 55% of students scoring one standard deviation above that cut score required remediation. In such a situation, the empirical data would undermine the validity argument for the cut score and CCR benchmark. However, determining CCR benchmarks based solely on empirical data at the high school level would be defensible if that data are representative across different types of postsecondary institutions and students. Judgmental processes could be used to determine the level of probability and

criterion level associated with performance levels on the high school assessments and to determine trajectory back to assessments in grades 3-8.

The challenge for PARCC will be to develop an approach to standard setting which is consistent and sensitive to empirical evidence related to CCR, to collect such evidence across a large and representative sample of institutions and to incorporate multiple approaches that also consider alignment to the CCSS. Today, many states have incorporated empirical data in standard setting and such approaches should have much greater emphasis and weight when the criterion is CCR. Wiley, Wyatt, and Camara (2010) describe several earlier efforts to define college readiness that evolved in one of two methods: (a) standards development approach using expert judgments and content based approaches, and (b) empirical approaches that attempt to link predictors and criterion of college and career success.

Judgmental methods

In *Understanding University Success* (Conley, 2003) over 400 faculty from 20 research universities identified the knowledge students needed to succeed in entry-level courses. Achieve Inc. developed a series of standards for both college and workplace success through partnerships with employers and faculty (American Diploma Project). ACT and the College Board also created similar standards for CCR. Each of these efforts used content-based approaches with experts in the discipline to develop and refine rigorous standards for specific disciplines. The Common Core State Standards (CCSS, 2010) were developed to provide a consistent, clear understanding of what students are expected to learn, so teachers and parents know what they need to do to help them progress toward college and career readiness and eventual success in postsecondary settings. The standards provide a vision of the skills and knowledge required for students to be college and career ready. The CCSS have largely supplanted previous documents developed with similar goals.

Prediction models

ACT and College Board employed empirical models to establish benchmarks. Each focused on a single academic predictor (test score) and criterion (college grades). Currently, several states are developing their own links to college outcomes either directly or through benchmarks set on admissions tests. For example, Texas used a variety of methods to determine the college readiness of students (Miller, Twing, and Meyers, 2008). They found strong correlations between scores on state and admissions tests, and they were able to develop predictive relationships and benchmarks to college success through admissions tests. Several states are currently undertaking similar studies and intend to establish direct links between state tests and college success criteria when students taking new tests complete college courses.

ACT and the College Board used similar methodologies based on linear and logistic regression, but selected different probabilities and criteria. ACT regressed specific subtest scores on grades in specific college courses (ACT, 2007; Allen and Sconing, 2005):

ACT English to Composition

ACT Math to College Algebra

ACT Science to Biology

ACT Reading to Social Science

The College Board used logistic regression to set the SAT College Readiness Benchmark (Wyatt, Kobrin, Wiley, Camara and Proestler, 2011). A third methodology using multiple predictors was conducted by Wiley et al. (2010) but not used operationally. Table 2 provides a brief comparison of these three methods. There have been several other attempts to develop multiple measures in defining CCR. Greene and Winters (2005) required students to have a high school diploma, read at the Basic level or above on NAEP, and to complete the minimum course requirements at less selective colleges.

Table 2

Comparison of ACT and SAT Benchmarks

	ACT (ACT, 2007; Allen and Sconing, 2005)	SAT (Wyatt et al. 2011)	Multiple Predictor (SAT) (Wiley, et al. 2010)
Criterion Variable	Grades in specific courses during freshman year (English Composition, College Algebra, Biology, Social Science)	FGPA	FGPA
Predictor	Individual ACT Test Score (e.g., English, Science)	Cumulative Score from a single administration (CR+M+W) and individual scores separately	(1) Cumulative Score from a single administration (CR+M+W); (2) HSGPA; (3) Academic Rigor[10]
Benchmark Cut Score (Percent at or above Benchmark[11])	Conjunctive –English 18 (62%), Math 22 (45%), Reading 21 (52%), Science 24 (30%), All 4 (25%) on a 1-36 scale in 2011	Compensatory – scores combined total 1550 (43% in 2011)	Conjunctive model – SAT scores are compensatory – total of 1550 (46%), HSGPA 3.33 (64%), Academic Rigor 10 on a 0-25 scale (53%), All 3 (32%) in 2009
Grade (Grades)	C (2.0) and B (3.0)	B- (2.65)	B- (2.65)
Probability	75% of C AND 50% of B	65%	65%
Institutions in Study	2 yr and 4 yr	4 yr (reported separately by selectivity)	4 yr (reported separately by selectivity)

Only 34 percent of all high school graduates in 2002 met all criteria. Citing hundreds of studies showing multiple predictors produce the highest validity coefficients, Wiley et al., (2010) proposed a CCR prediction model that included: (a) SAT scores, (b) HSGPA, and (c) a quantitative measure of

[10] See Wiley et al. (2010) and Wyatt, Wiley, Camara, and Proestler (In press) for computation of academic rigor index.

[11] Based on a college bound population of students taking each admissions test. Note the population of SAT and ACT test takers is not identical and percentages would differ with either a total college bound population or high school graduate population.

academic rigor of high school courses. They employed logistic regression to derive a cut score on each metric that corresponded to a 65 percent probability of a B- FGPA. Thirty-two percent of SAT test takers met all three benchmarks, while 23 percent met none of the benchmarks. Another study used four subject matter experts to set cut scores on admissions tests which would correspond to five categories of readiness from not qualified to very highly qualified. Students were moved up a category if they completed college core curriculum and down one category if they did not. High school grades, class rank, and test scores from the 1992 National Educational Longitudinal Study (NELS) were also used to compute a college qualification index. About 65 percent of 1992 high school graduates were minimally qualified (Berkner and Chavez, 1997). Finally, several districts have attempted to define college readiness. For example, Montgomery County in Maryland developed an index of college readiness that included seven key indicators ranging from advanced levels on K-8 reading tests, completion of grade 6 math in 5th grade, success in algebra courses and an AP examination, as well as a high minimum score on admissions tests (Wiley et al., 2010).

NAGB undertook a program of research which forms the validation framework for college and career preparedness. Four broad types of studies have been designed (NAGB, 2009):

- Content alignment – NAEP with ACT, ACCUPLACER, SAT and Workeys reading and mathematics. Overall, there was considerable overlap between ACT, SAT and NAEP. Elements of the NAEP domain were present in nearly all of the standards documents developed by both ACT and the College Board. There were similarities, but important differences, between items on NAEP and ACT. There were similar levels of depth of knowledge across NAEP and SAT, but stronger alignment with SAT Math than Critical Reading. Alignment results were somewhat weaker with ACCUPLACER and substantially weaker with Workeys (NAEP, 2011a).

- Linkages to other assessments and postsecondary outcome – NAGB (2011a) noted that the "highest priority is generally placed on empirical studies" (p.7). One study focused on the statistical relationship between NAEP and SAT, finding correlations of .91 and .74, respectively with math and critical reading. Results indicated that the SAT readiness benchmark of 500 for critical reading and math is very close to the NAEP Proficient cut scores. The SAT benchmark for math is slightly lower that the Proficient cut score while the benchmark for critical reading is slightly higher than the Proficient cut score. Judgmental standard setting–These studies focused on colleges as well as job training programs for five exemplar jobs with high potential for future employment requiring at least 3 months of training but not a bachelor's degree (Loomis, 2011). Many items on NAEP 12th grade reading and mathematics tests were judged to be irrelevant for job training programs in these five occupational areas. For example, between 56% and 100% of NAEP items in the four major math domains were deemed irrelevant by panelists representing three of the occupational clusters. Judgments about NAEP item relevance were less critical in reading, yet the emphasis of literary text over informational text was viewed as a major weakness in the NAEP framework if applied to job training programs. Overall, results suggested that panelists viewed career-readiness as less relevant to the NAEP frameworks. However, panelists still completed the standard setting process and selected cut scores for NAEP. In math, cut scores for job training programs were below those for college readiness and differed somewhat across occupations (with licensed professional nurse requiring more mathematical proficiency than other domains). In reading, cut scores were more consistent across college and career-clusters, yet there was substantial variation across panels in deriving final cut scores (NEAP, 2011b). At the end of the day, these studies are examples of the type of ongoing efforts required to establish career-readiness benchmarks and determine their level of generalizability

across job training programs. There is also a need to tie empirical outcome data from such programs to the CCR assessments in order to support the validity of cut scores.

- Postsecondary surveys – A national survey of two-and four-year higher education institutions is underway to collect information about assessments and cut scores used for course placement (remedial and credit bearing courses).

Benchmarks for Which Colleges?

When describing CCR benchmarks to higher education officials the first question is typically about the types of colleges used in the study. Clearly colleges differ greatly in their selectivity[12]. Clearly colleges differ widely in admissions test scores and high school grades of their admitted class and the same cut score will have different impact on at each institution. The fact that different institutions may require different levels of skill for entrance and success in entry level classes does not threaten the validity of the PARCC assessments, but could pose a problem in validation of a common CCR benchmark. There are logical arguments to favor different cut scores (or benchmarks) by institutional type (e.g., two year college, four year college, and postsecondary vocational training program), selectivity or academic major. Yet, wide variability would still be expected across institutions within these classifications and such a system would be inconsistent with the broader policy goals to increase the transparency between K-12 and higher educational systems in terms of requirements and readiness.

It is for those reasons that studies to date have relied on large representative samples in setting benchmarks. In addition, local institutional studies are most relevant for admissions decisions, but broader cut scores can be used across institutions if there is a level of consistency in the knowledge and

[12] Selectivity can be measured by traits of enrolled or admitted students (e.g., mean HSGPA, percent ranked at in the top 10% of their class) but is most often related to standardized test scores.

skills required for success in entry level courses which have been confirmed by several different faculty surveys. Still, this will be a continuing issue of debate in higher education and empirical studies will need to address this issue to verify a common benchmark is adequate preparation for success across different types of institutions and majors. If this is not found, then establishing separate benchmarks may eventually need to be reexamined.

College ready and Career ready: The same or different?

The types of empirical benchmark studies described for college readiness are relatively easy to do because we are simply matching performance on the predictor (test) with outcome data (college grades) and establishing some probability level (e.g., 67%, 50%) of a specific outcome (e.g., grade in Biology, FGPA). This doesn't work quite as well for career readiness for several reasons:

- What is career readiness – at this time, career-readiness appears to be defined as possessing the academic skills and knowledge required to be placed and succeed in a post-secondary vocational or career training program.

- What is the outcome – it would appear to be grades in such career-training programs, although no multi-institutional validity studies employing cognitive tests have been cited. In addition, many such programs may not maintain such data, and acquiring such data when it exists may be difficult and unsystematic. At this time we can not estimate how much data are available (do these programs give grades of just pass/fail, can data be matched to K-12 systems? Such programs are likely to be either certification programs within a 2-yr college or stand alone private institutions).

As noted above, the lack of criterion data from post secondary training institutions complicates efforts to define and measure career readiness. Such a limitation has not stopped many organizations and policy makers from pronouncing that college and career-readiness is the same thing. Studies by the

American Diploma Partnership (2004) and ACT (2006) are frequently cited when arguing that the same KSAs and level of performance on tests equates to both college and career readiness. Both studies focused on occupations which offer a sufficient wage to support a small family and are projected to increase in the future. Such jobs generally require some combination of vocational training and/or on-the- job experience, or some college, but less than a four-year degree. Results from these studies indicate that:

- Job incumbents had most often taken a core college preparatory curriculum and obtained good grades in these courses.
- Employers who were surveyed "reiterated the value of the knowledge and skills typically taught in Algebra I, Geometry and Algebra II" (American Diploma Partnership, 2004, p.106).
- The level of knowledge and skills required for these occupations (as measured by Reading for Information and Applied Math on Workeys) corresponded to the ACT College Readiness Benchmark (ACT, 2006).

The ADP study (2004) first used data from the National Educational Longitudinal Study (NELS) to determine that highly paid workers required 4 years of high school English and skills taught in Algebra II. Next, they worked with content experts to extract the knowledge and skills taught in such courses. Finally, they front-line managers from diverse industries reviewed preliminary workplace expectations and confirmed the importance of this content and skills emphasized in these courses during interviews. The ACT study (2006) provided minimal description of the methodology, sample, and analyses. They did report that a concordance was conducted between students in a single state who took Workeys and ACT. The ACT College Readiness benchmarks for Math and Reading were linked to ratings of the average KSAs required of job profiles. However, alignment studies between the CCSS and Workeys and ACT tests reported significant differences (NAEP, 2011a) so it is difficult to determine whether these tests actually

measure the same constructs and meet the statistical requirements to produce robust concordances (Dorans, Lyu, Pommerich & Houston, 1997).

While it would be enormously convenient to conclude that competencies and performance levels required for of the typical student entering all types of postsecondary institutions are identical, there is simply not sufficient evidence to make such claims or assumptions at this time. It may be difficult to conduct such research, but it appears there are substantial differences in the ability levels of students entering career training programs and two- or four-year colleges.

Performance Level Descriptors and Assessment Design

Performance standards provide high-level expectations regarding what students should know and be able to do. Performance levels also provide convenient categories for classification of students based on their test performance (Haertel, 1999). Cut scores are simply numeric points on a scale which are used to separate students into each of these categories. Historically, performance level descriptors (PLDs) have been developed quite late in the assessment design and development process – typically developed to provide a cognitive framework to aid panelists in standard setting, and more recently, in reporting (Egan, Schneider and Ferrara, In press). However, there is increased recognition that PLDs should be developed at the beginning of the assessment design process, especially when **evidence centered design (ECD)** approaches are used. When PLDs are established early in the design and development process they are employed to develop assessment frameworks, to specify the level of rigor of the performance, and to ensure that test items are directly related to the claims and evidence (Hendrickson, Huff, and Luecht, 2010; Perie, 2008).

PLDs should also be developed early in PARCC's assessment design process to ensure coherence of PLDs across grades. PLDs should be:

- Consistent with the CCSS content and skills

- Support the validity argument and intended uses of any definition of CCR with empirical data which distinguishes between the performance levels and the criterion measures selected for college and career success.

- Used in developing assessment blueprints, item specifications, test items, and efforts to maintain comparability across forms

- Complement the ECD approach[13] and help us to unpack and drill down to a deeper level of the standards to first build"claims and evidence statements" and corresponding tasks. The CCSS outline the domain of content knowledge and skills for ELA and math at each grade level. The CCSS will become more concrete when actual tasks are linked to them, and the PLDs and ECD process should provide specific observable evidence about what students must know and be able to do.

Empirical data for descriptions

There are at least three requirements when grounding PLD descriptors to empirical outcome data: (a) determine the appropriate criterion (e.g., placement in college courses, FGPA), (b) determine the appropriate performance level on the criterion (e.g., grades of B or higher, placement in college credit math courses corresponding to college level algebra or higher, completion of a career training program within 150 percent time), and (c) deciding on probabilities (e.g., 50 percent, 65 percent, 75 percent). For example, the ACT and the SAT benchmarks include all three of these elements in their definitions (e.g., 65 percent probability of FGPA of 2.65 or higher).

[13] This paper will not review ECD and readers are referred to Hendrickson et al., (2010) and Huff and Plake, (2010) for greater detail on how PLDs evolve from the ECD framework and can be used to guide item writing.

Policy capturing approaches with a diverse range of highly informed policymakers and experts is a promising method to derive initial PLDs and interpretative statements which would be subject to validation efforts. The College Board employed such a process with policymakers and leaders in higher education to set the initial definitions for their college readiness benchmarks (Wyatt et al., 2011). One risk in developing PLDs is that panelists may enter the process with strong opinions about which outcomes should be used. Several panelists who participated in the College Board process insisted that six-year graduation rates were the ultimate criterion for any effort to establish empirical evidence of college readiness despite the limitations of cognitive tests in prediction and the difficulty in obtaining such data. Extremely high cut scores on tests will often be required to arrive at an acceptable probability with outcomes like graduation which have weaker statistical relationships. Contingency tables were helpful in illustrating the impact of various combinations of different performance levels (grades of B vs C) and probability levels (e.g., 50 percent, 65 percent, and 80 percent) and served as an important tool in grounding panelists in the realities of college success. If such an approach is adopted by PARCC the following processes should be considered:

- Select panelists based on their background and expertise in college and career readiness and their knowledge of college and career training outcomes, as opposed to their political influence. Potential panelists could be selected from admissions and enrollment officers in two-year, four-year colleges and career training programs, researchers, business leaders with direct experience in certification or hiring, and administrators.

- Ensure panelists are trained on issues such as percentiles vs. percentages and differences in aggregate outcomes (percent of students who would be CCR) vs. individual probabilities. Additional panelists could be selected from groups which had some involvement in the

development of the CCSS, as long as they have a broader perspective and expertise beyond a specific discipline (e.g., validation panel, review panel).

- Employ more than one panel independently; it will reduce the impact of dominant panelists and provide cross validation of draft descriptors.

- Provide the context on what the assessments are designed to do and the types of interpretative statements that will be made based on test scores and PLDs for individual students and groups of students.

- Provide data on the relationship between various predictors and potential outcomes; provide some context on the meaning of such relationships.

- Consider introducing constraints on the definition before beginning the process – do not empower the panel to discuss any and all potential criteria but limit the process to the criteria which are most credible, defensible, measureable, and obtainable across post secondary institutions. Provide multiple sources of data and associations with other educational data if available (e.g., correspondence between scores on PARCC assessments and other assessments, correspondence between probabilistic outcomes and ACT/SAT/AP/NAEP outcomes). Assist panelists to understand the context for each type of data provided and do not include data which could be misleading.

Requirements for PLDs for PARCC Assessments

Development of PLDs should proceed in an iterative approach. Initial PLDs should be established as soon as possible to guide assessment design and development. These PLDs would then be refined during item development, following standard setting and again as several streams of empirical

validation evidence are obtained. Incorporating empirical evidence in the PLD process is much more critical with CCR assessments than has been the case with existing state assessments. Existing evidence and impact data from studies conducted by NAEP, test publishers, and even states should be included in setting initial PLDs. PARCC should also consider whether linking studies could be conducted between tests used in CCR benchmarking and both existing state assessments and field trial forms of PARCC assessments.

First, generic policy definitions should be developed which are consistent with the intended use and interpretative arguments for the assessments (Perie, 2008). At this point, PLDs are similar to what Kane (2001) labels as the policy assumption. These PLDs are often considered "draft PLDs" and are written at a high level.

In this specific instance, the policy assumptions and interpretative arguments all reference college and career readiness and success. Therefore, it seems reasonable to establish PLDs which provide meaningful distinctions in reference to CCR and college and career outcomes. There should be at least three PLDs.

Example 1

- Advanced or Exceeds –Students are above the standard for CCR (HS: Students who exceeded the standard for CCR; Grades 3-8: Students who exceeded the standard to be on track for CCR at grade X).
- Proficient or Meets (the Standard) – Students are CCR (HS: Students who have met the standard for CCR; Grades 3-8: Students who have met the standard to be on track for CCR at grade X)

- Basic or Emerging/Approaching – Students are below the standard for CCR (HS: Students who have not met the standard for CCR; Grades 3-8: Students who have not met the standard required to be on track for CCR at grade X)

PLDs can also be set which reference specific probability levels for CCR. That is, if empirical data are relied on to determine cut scored across performance levels then probability levels may be the best means for differentiation. Below are two examples of draft PLDs based on probabilities. The first set establishes the same probability for different outcome, while the second set uses different probabilities for similar outcomes. These example PLDs are appropriate for the high school assessments.

Example 2

- Advanced or Exceeds –Students have obtained the preparation in mathematics such that they have at least a 70% chance of succeeding in entry level college credit or and career training courses in mathematics (or general education).

- Proficient or Meets (the Standard) – Students have obtained the preparation in mathematics such that they have at least a 70% chance of enrolling, without remediation, in entry level college credit or and career training courses in mathematics (or general education).

Example 3

- Advanced or Exceeds –Students have obtained the preparation in mathematics such that they have at least a 70% chance of succeeding in entry level college credit or and career training courses in mathematics (or general education).

- Proficient or Meets (the Standard) – Students have obtained the preparation in mathematics such that they have at least a 50% chance of succeeding in entry level college credit or and career training courses in mathematics (or general education).

In these most recent examples, 'succeeding' should be defined and could be set at a grade (e.g., C, 2.0) or grades (FGPA of 2.0 or above). An additional performance level may be required to represent readiness in specific postsecondary settings, if empirical evidence does not support assumptions that college and career readiness are similar across different types of postsecondary institutions. It is possible that different levels of performance could be required for readiness and success in at least four types of postsecondary environments: (a) career training programs, (b) two-year colleges, (c) four-year colleges, and (d) STEM majors or programs. Higher expectations in mathematics are expected for students entering STEM areas. Certainly, students entering STEM majors are generally required to have completed more advanced courses in high school, such as precalculus and calculus, than students entering other postsecondary fields who may have not taken courses beyond advanced algebra. Similar differences in the knowledge and skills required in literacy may also exist in different postsecondary institutions. Proposing such distinctions in readiness would likely be met with controversy among policymakers who seek simple and declarative standards for all students, and strong empirical evidence would be required if PARCC were to proceed in this manner. PARCC states may wish to have a fourth performance level between Proficient and Basic which describes students who have met expectations for high school but are not CCR or a more popular level of below basic.

PLDs[14] for high school assessments should be derived from both the CCSS and empirical evidence related to CCR. Best practice for developing descriptors would also include using the same number of

[14] Egan et al., (in press) differentiates between different types of PLDs, for example, target, range, and reporting PLDs. I would suggest that we focus on one set of PLDs which become more specific and refined overtime, as

levels, labels, and generic descriptions for PLDs across grades (Egan, et al., in press; Perie, 2008). The draft PLDs or generic policy descriptions should be informed by both content experts who are able to begin to relate the knowledge and skill required for each level of performance to specific CCSS, and policymakers who can develop initial probabilistic statements related to college and career success (outcomes). It also seems appropriate to begin PLD development at the high school level where the relationship between predictor (PARCC assessments) and criteria (outcomes in college and career training programs) are temporally closest. Once draft PLDs are established for high school tests the task becomes one of sequencing or back mapping PLDs to sequential grades.

Hambleton and Plake (In press) identify a several concerns in basing PLDs on empirical data from postsecondary success and find more traditional standard setting approaches more appealing. Results from PARCC assessment will almost certainly be used as a major component in evaluating teachers and they note that much of the success associated with performance during freshmen year in postsecondary educational programs can be associated with the quality of postsecondary instruction and other factors that are not directly related to previous instruction (e.g., motivation, persistence, study skills, attitudes). This is a very compelling argument, but can also apply to success on high school outcomes (e.g., grades, scores on PARCC assessments) to some degree. Overall, using PARCC assessments as a principle means of evaluating teachers raises a number of concerns with the design, the specification of PLDs and the interpretative argument required to assemble validation evidence to support such uses. As discussed above, empirical data from postsecondary outcomes can make provide a source of strong external evidence and objectivity to classification of students as CCR, but can result in negative consequences if used as a significant factor for teacher evaluations. Developing PLDs and cut scores in the more

---

opposed to supporting different versions of the same PLDs for different purposes or audiences since that can lead to confusion and misuse.

traditional ways, based on judgmental processes and impact data in high school, may reduce the negative consequences of using scores in teacher evaluations, but this method will not eliminate them. Additionally, if there are discrepancies between the judgments and postsecondary outcomes the assessments may been views as less legitimate or credible as a measure of CCR. At the end of the day, PARCC leadership will need to balance the advantages and disadvantages of approaches in light of the primary purposes of the assessments (to indicate CCR or to evaluate teachers).

Content descriptors

Subject-matter experts (SMEs) and other educators should apply their knowledge about grade-level content standards to supplement the initial generic policy based descriptions. Educators from adjacent grades or levels should also be included within the grade specific panels (e.g., 6th and 8th grade ELA teachers serving on the 7th grade panels), as well as grade level SMEs who have experience with students with limited English proficiency and students with disabilities. Faculty teaching entry level postsecondary courses in the disciplines should also be included in panels developing PLDs for high school assessments. Care should be given to include faculty from two-year and four-year colleges, and several different career training programs. One possible method would be to use the five occupations selected by NAEP as the focus of initial efforts to establish PLDs which are responsive to career training programs. Given these requirements, high school panels will likely be larger than panels for other grades.

There are a number of other issues which must be considered in developing and refining the PLDs, such as the impact of accommodations, administration conditions, and accessibility of the assessment design. State testing programs have struggled with defining accommodations for their reading and ELA tests, particularly at the high school level, even before considering college and career readiness, grounding their work in the content and achievement standards for reading at the high school level

(Johnstone, Thurlow, Thompson, & Clapper, 2008; Thompson, Johnstone, Thurlow, & Clapper, 2004; Johnstone & Thurlow, 2010). The National Accessible Reading Assessment Projects (NARAP – DARA, PARA, TARA) are just completing a series of projects that resulted in a set of principles that define how to assess students with atypical profiles in *how* they read. In the Common Core State Standards, on p. 6 of the section on including students with disabilities it says "for students with disabilities *reading* should allow for the use of Braille, screen reader technology, or assistive devices…"

PARCC partners need to define their positions on the appropriate assessment of students who use technology or other tools to read, and then reflect these decisions in how the PLDs are worded and interpreted. Defining "reading" broadly to include Braille would embrace current practices in most states, where students who are blind are permitted to use Braille on all assessments. In most states, one can assume that the PLDs are meant to include students who use Braille to respond to the reading assessment, regardless of how they are currently worded. PARCC partners should think broadly beyond Braille given improved awareness and understanding of how adults with disabilities succeed in college and career, and choose words in PLDs carefully. The use of the words like "consider" rather than "read" and "produce" rather than "write" may be appropriate. For example, writing to a prompt could be incorporated into the PLDs - - "when asked to consider a technical passage about a specific issue, proficient students are able to produce a coherent analysis of the pros and cons of that issue and provide specific examples from the passage." That reflects the necessary skills for college and career of reading and writing, while still allowing for multiple ways to "read" and "write."

Wording and formatting of the descriptors must also be coherent across grades (Perie, 2008). PLDs can be related to the performance of the"typical" student or performance of the"borderline" student. While there are good reasons for each approach, a focus on students at the border seems most appropriate for CCR. As noted above, once this initial process is completed, the PLDs can inform the

development of the assessment framework, test specifications, and item development. This process will likely raise some questions or concerns that will result in refinements to statements in the PLDs, as well as greater specification of the knowledge and skills students are expected to demonstrate. The final review and refinements of PLDs will come as a result of the standard setting process and as validation evidence becomes available. In summary, great care is needed in developing the PLDs early enough so they are a central driver of the assessment framework, test specifications, and item development. Descriptions must be challenged and revisited as assessment development progresses, standard setting is conducted, and empirical data become available. Ambiguous PLDs, as well as PLDs which are not strongly related to statements (or predictions) concerning CCR and future performance, pose a significant threat to the validity of the standard setting and interpretative statements associated with the assessment (Huff and Plake, 2010).

PLD progression and coherence across grades

For ELA, the design of the standards already allows for a natural back mapping of the high school PLDs to earlier grades. The College and Career Readiness Anchor Standards (Common Core State Standards, 2010, p. 35, 41, 48, and 51) provide general standards which span across all grades and work in tandem with the K-12 grade-specific standards. The coding of the standards allows one to reference each specific K-8 standard to the corresponding Anchor Standard. For math, the high school standards are organized by conceptual categories (e.g., Functions, Modeling, etc., located on pp 58-79). K-8 mathematics standards could be categorized under each conceptual category giving us a sense of the precursor knowledge required for each category. The math standards also address key "mathematical fluencies," or areas of content mastery required at certain grades (mainly K-8), and the lower grade PLDs would need to consider these fluencies.

Once PLDs have been developed for high school assessments, SMEs should be convened again to establish the achievement targets at lower grade levels, keeping the actual grade-specific standards in mind. If trained in the principles of Evidence Centered Design the SMEs can articulate the on-track PLDs at the lower grade levels, and they would need to consider both content and skills (or practices). As part of the ECD process in AP, detailed "skills" or verb charts and matrices were developed for different subject areas. A resource like this for each discipline could be helpful in back mapping the PLDs.

As noted earlier, empirical methods should drive the development of PLDs for high school assessments unless PARCC leadership determines the use of assessment scores for teacher evaluation is of more importance than determination of CCR of students. Statistical methods are still relevant but may have less utility in establishing interim PLDs and cut scores for assessments in grades 3-8 for several reasons. One practical limitation is the difficulty in back mapping the high school PLDs to grade 8 for students who complete different courses in different sequences and at different times. PLDs from grades 3-8 can be sequenced, but then there will be a huge gap until one reaches the PLD for the cumulative scores on the high school assessments. The absence of annual measures (e.g., grade 9, grade 10) and the differences in curriculum complicate and may undermine reliance on empirical methods alone in the earlier grades. PARCC leadership may consider two different processes in establishing PLDs for grades 3-8 and high school. Empirical data can still play a role in setting cut scores and establishing PLDs for grades 3-8, but such data could serve to validate and supplement other methods until longitudinal data are collected on students who complete the assessments and go on to postsecondary experiences.

Empirical data can supplement or validate a PLD and standard setting process that relies on judgmental methods for assessments in grades 3-8. For example, statistical projection methods could determine the range of scores on an 8th grade test that is associated with the CCR benchmark on the

high school assessments, as well as benchmarks on NAEP and pre-admission tests. Similarly, the range of scores on a 7th grade test could be linked to the 8th grade benchmark or directly to benchmarks on other tests (if a common persons design can be implemented). ACT and the College Board employed statistical projection techniques to establish CCR benchmarks from their admissions tests to lower grades (e.g., 8th-11th grades). ACT used a database of 150,000 students who took Explore (8th graders), PLAN (10th graders), and ACT. Success on Explore and PLAN was defined as meeting the ACT benchmark value with a 50 percent probability (Sconing, 2010). Proctor, Wyatt and Wiley (2010) used a similar procedure establishing a benchmark for PSAT/NMSQT with both 10th and 11th graders based on a 65 percent probability of attaining the SAT benchmark. Currently, the College Board is undertaking a national field test to scale ReadiStep, an 8th grade test modeled after the PSAT/NMSQT and SAT. The study should result in statistical linkages between all three assessments which will permit interpretations related to student progress and create new benchmarks that will determine if students are on track toward reaching their CCR benchmarks. Statistical projection, including logistic and linear regression, are the most common means of back mapping empirical benchmarks, but there are other methods available. For example, the College Board also calculated benchmarks using a borderline groups method (Livingston and Zieky, 1982), where students who scored at the border of CCR on the PSAT/NMSQT were identified and their mean score on ReadiStep was used as an interim benchmark. Contrasting groups (Livingston and Zieky, 1982) was attempted but abandoned. This method involved using the PSAT/NMSQT score at which the distributions of students who just met and just missed reaching the SAT CCR benchmark overlapped. Finally, IRT offers another approach where you would concurrently calibrate matched groups and use the test characteristic curve to identify the PSAT/NMSQT score that map to the corresponding SAT benchmark score (*K. Sweeney personal correspondence, August 30 2011*).

Validation and Concluding Thoughts

The goal of each section in this paper was to discuss assumptions, issues, and options which impact important decisions about the overall validation of inferences and interpretations based on test scores. As noted at the outset of the paper, the primary purpose of the PARCC assessments is to determine if students completing high school are CCR and if students in earlier grades are on track to reach that milestone. The term CCR needs to be operationally defined, and then specific criteria and metrics are required to measure whether students in fact have attained this level of success. If the definition of CCR or the way we measure student success in college or career training programs is vague or ambiguous, the assessment results and interpretations are more likely to be confused and misused.

PLDs should be established early in the process to support the initial work in designing the assessment framework, specifications, and tasks. PLDs should also include empirical benchmarks in the form of expected outcomes (e.g., grades, placement in college credit courses), levels (e.g., grades of C or better, FGPA) and probabilities. PLDs must "define college and career readiness, in early grades and at the end of high school, predict future outcomes such as continuing achievement across grades, following a trajectory toward readiness for college and career, and subsequent successful performance in college or in the workplace" (Egan et al., in press, p.33). While PLDs need not include the specific probabilities in their descriptions, the probabilities will be implied and will constitute part of the validation argument. For these reasons it seems appropriate to define college and career readiness in terms of specific outcomes and probabilities. The most appropriate and defensible college outcomes include placement into college credit bearing courses and achievement in freshman courses (e.g., specific course grades or GPA).

Defining and measuring readiness for success in career training programs is more nebulous and elusive. Despite the efficiency offered by a single standard, at this point there is insufficient evidence to

conclude that the same standard or benchmark will serve college and career readiness equally. Many advocates of a single standard will argue that even among colleges a single standard is ineffective, but here we have data that can be brought to this issue. We can conduct research to determine the classification accuracy a cut score has across types of colleges (e.g., selectivity, two-year vs. four-year), courses, and even programs (e.g., STEM vs non-STEM). Such data is largely absent from any large and representative sample of career training programs and entry level jobs. That is, we can estimate error over time and across institutions with the type of college outcome data that is generally used in higher educational validity studies, but that data are not available to inform decisions about career readiness. Content validation approaches, including surveys, reviews of job requirements, and perhaps some local concurrent and predictive validation studies will be required to provide a base of evidence related to how a CCR benchmark relates to career readiness and success.

The general validation approach and assumptions of the NAGB research agenda can form a blueprint for both the types of research and questions PARCC needs to consider establishing a CCR benchmark and accumulating supporting evidence. Empirical studies conducted over time and across different postsecondary institutions will provide the most compelling evidence of CCR benchmarks and standards. A variety of studies can be designed to provide this evidence, some of this while the first cohort of students is still enrolled in high school. For example, concurrent validation studies can be conducted by comparing student performance on PARCC assessments with performance in college level work (AP examinations, dual enrollment courses, IB examinations). A second strand of validation evidence may come from statistical linking studies between PARCC assessments and other tests such as EXPLORE and ReadiStep (8th grade), PLAN, and PSAT/NMSQT (10th grade), ACCUPLACER, ACT, and SAT (11th and 12th grades), NAEP performance[15] (4th, 8th and 12th grade) and existing state tests where

[15] Using statistical linking studies similar to those described in NAGB (2011) and Loomis (2011).

relationships have been established with external post secondary criteria (this is beginning to occur in more states). A number of states are establishing similar CCR benchmarks and these should also be examined if they are based on empirical evidence. It will be just as important to develop data license agreements with states having K-20 data systems. However, validity studies conducted within state will exclude over a quarter of all students in four-year colleges nationally, and over 40 percent of students in 13 states. Therefore, PARCC needs to begin to plan for cross-state or national validation studies and cooperative agreements which capture this population. Decision accuracy and classification accuracy will likely be used to determine the efficacy and validity of inferences that the consortia and states wish to make on the basis of PARCC assessments. How often do students who meet the CCR proficiency level fail in college and post-secondary training, and how often do students who are on the borderline and fail to meet the CCR proficiency level succeed in postsecondary education?

The requirement for validation evidence is not restricted to empirical relationships but ultimately must support inferences about the knowledge and skills in the CCSS. Is there evidence that the knowledge and skills addressed in the CCSS are required for success in postsecondary education? Do students lacking some of these skills succeed and do students possessing these KSAs fail? The standards were developed using collections of evidence, but it may be a stretch to say they were empirically validated[16]. In the early days of CCSS development, the committee wanted to gather evidence such as student outcome data, curriculum surveys, etc. and link a specific reference to each standard. This occurred with the first draft of the College and Career Readiness Standards, but those standards no longer exist in their original form. The effort to link evidence to each standard did not continue in later drafts. In math, the committee was more diligent about keeping valid references as a centerpiece of the development process. In the introduction section of the math standards, the CCSS (2010) explicitly

[16] In mathematics, Works Consulted can be found on pages 91-93; for ELA, the Bibliography is on pages 35-39.

state, "…the development of these Standards began with research-based learning progressions detailing what is known today about how students' mathematical knowledge, skill, and understanding develop over time" (p.4). ELA does not make this same statement.

Summary Recommendations about PLDs and Validation

In summary, the following recommendations for developing PLDs and a validation argument for CCR should be considered:

1. Determine the primary use(s) and interpretative arguments that will be derived from test scores. The validation argument will rest on the rationale for establishing PLDs. If assessments are primarily used for making determinations about CCR then external data on postsecondary success should drive the process used to set PLDs and cut scores for high school assessments. If scores are primarily used for evaluating teachers then outcomes from postsecondary success are problematic and can undermine the validity argument.

2. CCR needs to be operationally defined, and then specific criteria and metrics that measure whether students in fact have attained this level of success should be identified as part of the PLD development. If the definition of CCR or the way we measure student success in college or career training programs is ambiguous the assessment results and interpretations are more likely to be misused. Definitions should consider the type of criterion data available to validate performance levels and interpretative arguments concerning college- and career-readiness. This issue is particularly problematic in defining and validation inferences about career-readiness.

3. PLDs should be established now. They are needed to support the design and development of the assessment framework, test specifications and exemplar items. This is especially important since ECD will be employed in designing and developing the PARCC assessments.

4. Empirical data from postsecondary outcomes should drive the development of PLDs and cut scores for the high school assessments. Expected outcomes (e.g., grades, GPA), performance levels (e.g., grades of C or higher) and probabilities (e.g., 70%) should be considered in establishing PLDs. If postsecondary empirical data are not used in establishing PLDs and cut scores then great caution should be taken in making any statements about the use of scores to predict or determine CCR. If empirical postsecondary success is not consistent with the outcomes from PARCC assessments (e.g., 50% of students are considered CCR from the PARCC assessments, but half of these students require remediation or fail entry level courses) then use of scores for determining future success in college or career training programs may have to rest on local institutional practices and studies.

5. Different processes can be employed to establish PLDs and cut scores for high school and grade 3-8 assessments. Judgmental processes can be used to supplement or validate the empirical processes to set PLDs at the high school level. A different process can be used to set initial PLDs for PARCC assessments in grades 3-8.

6. Statistical linkages or projections can be useful in setting cut scores and PLDs for PARCC assessments in grades 3-8. These methods might supplement or validate judgmental processes until longitudinal data are collected on students completing assessments across grades and into postsecondary education. Impact data should certainly be considered and articulation procedures should ensure results appear logical and reasonable across grades.

7. PLDs and cut scores established before 2015 should be considered as 'interim' statements and revisited once longitudinal data are available on a large and representative population of students. This process should be revisited every several years (e.g., 5 years).

8. A comprehensive validation argument and approach should be established that can guide research and influence the assessment design and development. The NAEP 12th grade research agenda can form a blueprint for the types of research and questions PARCC needs to address. Empirical studies conducted over time and across different postsecondary institutions will provide the most compelling evidence of CCR benchmarks and standards, but evidence should encompass a variety of approaches.

9. PARCC should begin documenting and evaluating the quality, efficacy and administrative requirements associated with obtaining student level data from states and postsecondary institutions. Specifically, PARCC could conduct a comprehensive survey building on data available from the Data Quality Campaign (DQC) about state K-12 and post secondary databases. The final product would be a document which specifies the availability of data across institutions and the level of consistency used in classifying background and course data. In addition, PARCC should obtain files from each state's K-12 and higher educational system to determine the level of effort requiring in using the data and to identify any administrative issues which could later restrict access to data. Obtaining data from a wide variety of postsecondary institutions within states is not always as efficient or seamless as one would expect. PARCC should request data from the K-12 and postsecondary systems in each PARC C state now and 'test' the assumptions of full cooperation of institutions and quality of data before proceeding with assumptions that could limit or delay the ability to provide validation evidence to support the intended use of scores.

References

ACT (2007). *ACT Technical Manual*. Iowa City, IA: ACT.

ACT (2006). *Ready for college, ready for work: Same or different*. Retrieved from http://www.act.org/research/policymakers/pdf/ReadinessBrief.pdf

Adelman, C. (2006). *The Toolbox Revisited: Paths to Degree Completion From High School Through College.* Washington, D.C.: U.S. Department of Education. Retrieved from http://www2.ed.gov/rschstat/research/pubs/toolboxrevisit/toolbox.pdf

Allen, J., and Sconing, J., (2005). *Using ACT assessment scores to set benchmarks for college readiness*. ACT Research Report 2005-3. Retrieved from http://www.act.org/research/reports/pdf/ACT_RR2005-3.pdf

American Diploma Partnership (2004). *Ready or not: Creating a diploma that counts*. Retrieved from http://www.achieve.org/files/ADPreport_7.pdf

Baum, S. and McPherson, M. (2008). *Fulfilling the committee: Recommendations for reforming federal student aid.* New York, NY: College Board.

Berkner, L., & Chavez, L. (1997). *Access to postsecondary education for the 1992 high school graduates*. (NCES 98-105). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Berry, C. M. & Sackett, P. R. (2008, March). *The validity of the SAT at the individual course level*. Paper presented at the American Educational Research Association Conference, New York, NY.

Bowen, W.G., Chingos, M. M., and McPherson, M. S. (2009). *Crossing the finish line: Completing College at America's Public Universities*. Princeton, NJ: Princeton University Press.

Bowers, B. (2007, December 6). Tracing business acumen to dyslexia. *The New York Times*. Retrieved from http://www.nytimes.com

Burgstahler, S.E. & Cory, R.C. (Eds.). (2008). *Universal Design in Higher Education: From Principles to Practice.* Boston: Harvard Education Press.

Burton, N. W. and Ramist, L. (2001). *Predicting success in college: SAT studies of classes graduating since 1990.* College Board Research Report 2001-2. Retrieved from http://professionals.collegeboard.com/profdownload/pdf/rdreport200_3919.pdf

Camara, A. (2009). *Expenditure analysis: The funding of a college degree*. Unpublished paper.

Camara, W. J. (2005a). Broadening criteria of college success and the impact of cognitive predictors. In W.J. Camara and E.W. Kimmel (Eds.) *Choosing students: Higher education admissions tools for the 21st century* (pp. 53-80). New York, NY: Routledge.

Camara, W. J. (2005b). Broadening predictors of college success. In W.J. Camara and E.W. Kimmel (Eds.) *Choosing students: Higher education admissions tools for the 21st century* (pp. 81-105). New York, NY: Routledge.

Common Core State Standards (2010). Retrieved from http://www.corestandards.org/the-standards

Conley, D. T. (2011). *Redefining college readiness*, Volume 5. Eugene, OR: Educational Policy Improvement Center.

Conley, D. T. (2003). *Understanding University Success*. Center for Educational Policy Research: Eugene, OR. Retrieved from https://www.epiconline.org/files/pdf/UUS_Complete.pdf

Conley, D., Drummon, K.V., DeGonzalez, A. Roosebloom, J., & Stout, O. (2011). *Reaching the goal: The applicability and importance of the common core state standards to college and career readiness.* Eugene, OR: Educational Policy Improvement Center.

Conley, D., McGaughy, C., Brown, D., vander Valk, A., & Young. B. (2009). *Texas career and technical education career pathways analysis study.* Eugene, OR: Educational Policy Improvement Center.

Conley, D., McGaughy, C., Cadigan, K., Forbes, J., & Young. B. (2009). *Texas college and career readiness initiative: Texas career and technical education phase 1 alignment analysis report.* Eugene, OR: Educational Policy Improvement Center.

Dorans, N. J., Lyu, C.F., Pommerich, M., & Houston, W.M. (1997). Concordance between ACT assessment and Recentered SAT I Sum Scores. College and University, 73(2), 24-34.

Egan, K. L., Schneider, C., and Ferrara, S. (In press). *Performance level descriptors: History, practices and a proposed framework.*

Greene, J. P., & Winters, M. A. (2005). *Public high school graduation and college-readiness rates: 1991-2002*. (Education Working Paper No. 8, February 2005). Manhattan Institute. Retrieved April 23, 2008, from http://www.manhattan-institute.org.

Haertel, E. H. (1999). Validity arguments for high-stakes testing: In search of the evidence. *Educational Measurement: Issues and Practice, 25* (4), 5-9.

Hammick, F. A., Schuh, J. H., and Shelley II, M. C. (2004). Predicting higher education graduation rates from institutional characteristics and resource allocation. *Educational policy analysis archives, 12* (19). 1-24.

Harmston, M. T. 2004. *Cross-Validation of Persistence Models for Incoming Freshmen.* AIR Professional File, Number 93. Tallahassee, FL: Association for Institutional Research.

Haertel, E. H. (1999). Validity arguments for high-stakes testing: In search of the evidence. *Educational Measurement: Issues and Practice, 25* (4), 5-9.

Hendrickson, A., Huff, K., and Luecht, R. (2010). Claims, evidence, and achievement-level descriptors as a foundation for item design and test specifications. *Applied Measurement in Education, 23*, 358-377.

Huff, K., and Plake, B. S. (2010). Innovations in setting performance standards for K-12 test-based accountability. *Measurement, 8,* 130-144.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64,* 425-461.

Kane, M. (2001). So much remains the same: Conception and status of validation in setting standards. In G. Cizek (Ed.) *Setting performance standards* (pp.53-88). New York: Routledge.

Kearns, J., Kleinert, H., Harrison, B., Sheppard-Jones, K., Hall, M., & Jones, M. (2011). *What does college and career-ready mean for students with significant cognitive disabilities?* Lexington, KY: National Alternate Assessment Center.

Kobrin, J. L., Patterson, B. F., Shaw, E. J., Mattern, K. D., & Barbuti, S. M. (2008). *Validity of the SAT for Predicting First-Year College Grade Point Average*. (College Board Research Report No. 2008-5). New York: The College Board.

Johnstone, C.J., Thurlow, M.L., Thompson, S.J. & Clapper, A.T. (Spring, 2008). The potential for multi-modal approaches to reading for students with disabilities as found in state reading standards. *Journal of Disability Policy Studies 18*(4), 219-229.

Johnstone, C.J., & Thurlow, M. (2010). Statewide testing of reading and possible implications for students with disabilities? *The Journal of Special Education.* Online preprint available at http://sed.sagepub.com/content/early/2010/05/25/0022466910371984.full.pdf.

Livingston, S. A. & Zieky, M. J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests.* Princeton, NJ: ETS.

Loomis, S. (2011, April). *College readiness, career readiness: Same or different?* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Miller, G. Twing, J. S. and Meyers, J.L. (2008). TAKS Higher Education Readiness Component (HERC TAKS and College Readiness Correlation Study. Unpublished paper.

National Assessment Governing Board (2009). *Making New Links, 12th Grade and Beyond: Technical Panel on 12th Grade Preparedness Research Final Report*. Retrieved from http://www.nagb.org/publications/PreparednessFinalReport.pdf .

National Assessment Governing Board (2011a). COSDAM briefing materials for the August 2011 meeting of the National Assessment Governing Board.

National Assessment Governing Body (2011b). The NAEP Grade 12 preparedness research project judgmental standard setting (JSS) studies summary report. Unpublished paper under contract ED-NAG-10-C-0004. Prepared by Measured Progress.

National Center on Education Statistics (2010). Retrieved from http://nces.ed.gov/programs/digest/d10/tables/dt10_232.asp and http://nces.ed.gov/programs/digest/d10/tables/dt10_231.asp

No Child Left Behind of 2002. Pub.L. 107-110, 115 Stat. 1425 (January 8, 2002).

Perie, M. (2008). A guide for understanding and developing performance-level descriptors. *Educational Measurement: Issues and Practice, 27,* 15-29.

Plake, B., and Hambleton, R. (In press).

Proctor, T., Wyatt, J. and Wiley, A., (2010). *PSAT/NMSQT: Indicators of college readiness*. (College Board Research Report No. 20010-4). New York: The College Board.

Purcell, J. and Clark, A. 2002. *Assessing the Life and Times of First-Semester Academic Failures.* Paper presented at the annual conference of the Southern Association for Institutional Research. Baton Rouge, LA.

Reitz, S. (2011, March 29). *Connecticut governor opens up about dyslexia struggles*. Associated Press. Retrieved from http://www.ap.org/

Robbins, S., Allen, J., Casillas, A., Peterson, C. H., & Le, H. (2006). Unraveling the Differential Effects of Motivational and Skills, Social, and Self-Management Measures From Traditional Predictors of College Outcomes. *Journal of Educational Psychology, 98(3),* 598–616

Robbins, S., Lauver, K., Le, H., Davis, D., Langley, R., & Carlstrom, A. (2004). Do psychosocial and study skill factors predict college outcomes? A meta-analysis. *Psychological Bulletin, 130,* 261–288.

Sackett, P. R., Borneman, M. J., and Connelly, B. S. (2008). High stakes testing in higher education and employment: Appraising the evidence for validity and fairness. *American Psychologist, 63* (4), 215-227.

Schmitt, N., Billington, A. Q., Keeney, J., Oswald, F. L., Pleskac, T., Sinha, R., et al. (2009). Prediction of four-year college student performance using cognitive and noncognitive predictors and the impact on demographic status of admitted students. *Journal of Applied Psychology*, *94*, 1479–1498.

Schultz, P. (2011). *My Dyslexia*. W.W. Norton & Company.

Sconing, J. (2010, June). *Measuring college readiness: Validity, cut scores and looking into the future*. Paper presented at the Large Scale Assessment Conference, Detroit, MI.

Shaw, E.J. , and Patterson, B.F. (2010). What should students be ready for in college? A first look at coursework in four year postsecondary institutions in the U.S. College Board Research Report 2010-1. Retrieved from http://professionals.collegeboard.com/profdownload/pdf/10b_1417_ FirstYrCollCourseRR_WEB_100611.pdf

Thompson, S. J., Johnstone, C. J., Thurlow, M. L., & Clapper, A. T. (2004). *State literacy standards, practice, and testing: Exploring accessibility* (Technical Report 38). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thurlow, M. L., Laitusis, C. C., Dillon, D. R., Cook, L. L., Moen, R. E., Abedi, J., & O'Brien, D. G. (2009). *Accessibility principles for reading assessments.* Minneapolis, MN: National Accessible Reading Assessment Projects. Available on the World Wide Web at www.narap.info.

Tinto, V. (1987). *Leaving college*. Chicago, IL: University of Chicago Press.

Walsh, P. (2011, October 4). Law school test administrator relents to feds, accommodates U grad with ADD. *The Minneapolis Star Tribune*. Retrieved from www.startribune.com. (The actual settlement agreement is posted at http://www.ada.gov/lsac_2011.htm)

Wiley, A., Wyatt, J., and Camara, W. J. (2010). *The development of a multidimensional college readiness index.* College Board Research Report 2010-3. Retrieved from http://professionals.collegeboard.com/profdownload/pdf/10b3110_CollegeReadiness_RR_WEB_110315.pdf

Wyatt, J., Kobrin, J., Wiley, A., Camara, W.J., and Proestler, N. (2011). *SAT benchmarks: Development of a college readiness benchmark and its relationship to secondary and postsecondary school performance.* College Board Research Report 2011-5. Retrieved from http://professionals.collegeboard.com/profdownload/pdf/RR2011-5.pdf

Wyatt, J., Wiley, A., Camara, W.J. and Proestler, N. (In press). The Development of an index of academic rigor for college readiness.