California Department of Education

Measures for a College and Career Indicator: Research Brief on Innovative Measures

June 17, 2014

Educational Policy Improvement Center (EPIC)

Introduction

In September of 2012, Governor Jerry Brown signed into law Senate Bill 1458, which calls for California's school accountability system to shift from a near exclusive reliance on state test scores to a broader range of measures demonstrating student achievement. At the high school level, starting in the 2015–2016 school year, the Academic Performance Index (API) will include an indicator composed of measures reflecting students' college and career preparedness.

To determine exactly what measures will be included in this new indicator, the State Superintendent of Public Instruction and the State Board of Education will consider input from regional public meetings, a statewide survey, and recommendations from the Public Schools Accountability Act (PSAA) Advisory Committee. To further support this decision-making process, the California Department of Education has contracted with the Educational Policy Improvement Center (EPIC) to conduct analyses of six different types or clusters of potential measures of college and career preparedness, summarized in a series of six white papers and a final summary report.

This white paper considers innovative measures—specifically metacognitive assessments, performance assessments, and the California State Seal of Biliteracy—as potential measures to be included in California's College and Career Indicator (CCI). The white paper begins by describing the criteria that each innovative measure is evaluated against. Next, each innovative measure is introduced and then evaluated separately against the analytical framework to determine the technical quality, stakeholder relevance, and system utility when used as a component measure of accountability. The white paper concludes with a summary of the analysis of innovative measures as a cluster of indicators, identifying major strengths, weaknesses, and trade-offs.

Evaluation Against an Analytical Framework

Working in collaboration with the PSAA Advisory Committee, EPIC developed an analytical framework to provide a set of criteria by which each measure can be evaluated for its potential to contribute to a revised Academic Performance Index (API). This framework was adapted from the Advisory Committee's API Guiding Principles and was supplemented with additional criteria specific to the charge of designing a CCI. The 10 criteria are grouped into three dimensions: technical quality, stakeholder relevance, and system utility.

Evaluating innovative measures, the topic of this white paper, requires a refinement and adaptation of these 10 criteria. By definition, innovative measures do not conform well to current practice. Therefore, they often violate the assumptions upon which traditional accountability systems operate. The operational definitions of seven of the ten criteria used in previous white papers (EPIC, 2014a, 2014b) have been adapted for the purpose of judging the suitability of innovative measures as contributors to a revised API; the seven adapted criteria are: A2. Fair Comparisons; A3. Stability; B1. Value to Students (redefined from Student Currency); B2. Public Understanding; B3. Content, Skills, and Competencies; C1. Minimal Burden; and C2. Student Coverage. Innovative measures will be rated on a three-point scale (strong, moderate, or weak) for all ten criteria.

A. Technical Quality: For the purposes of this white paper, technical quality is defined as the degree to which a measure has predictive validity for forecasting how students will perform in postsecondary pathways, how well it allows for fair comparisons among different subpopulations of

students, and the degree to which it is sufficiently stable, which is necessary to track trends longitudinally.

A1. Relationship to Postsecondary Success: For the purposes of this white paper, postsecondary success consists of a wide array of outcome variables including college matriculation, persistence, course grades, grade point average (GPA), and degree completion. Career success outcomes are more challenging to measure because most of them occur after the end of formal education. Examples include rate of employment, starting salary, advancement in a career pathway, or self-reported job satisfaction. In-school indicators could include participation in internships and other forms of career exploration, declaration of a major and then completion of the declared major, decrease in undeclared majors, completion rates in certificate programs, or number of certificate programs in which a student enrolls before completing one program.

A2. Fair Comparisons: This analytic criterion is based on the requirement that the API must give all students a fair chance to show what they know and have learned. For the purposes of this white paper, the extent to which a measure provides fair comparisons across students and schools is determined by ascertaining the degree of systematic bias the criterion evidences. Definitions of fairness vary across sectors and domains, but most incorporate an understanding of equality and equity as overlapping but nonidentical components. Equality may refer to level opportunities or level outcomes, while equity focuses on redressing unequal opportunities or outcomes through differential inputs. What constitutes fairness must be clearly specified for innovative measures, particularly those that rely on judgment-based methods such as student self-evaluation or teacher observations. These methods are viewed by many psychometricians as more vulnerable to certain types of bias than data from standardized test scores, although it is worth noting that such tests are likewise criticized as reflecting other forms of bias (Conley, 2013).

A3. Stability: This evaluative criterion is chiefly concerned with the ability to make comparisons over time, both within one school and between schools. In order to measure school performance and improvement comparably over time, the measurement system should be based on definitions that remain relatively constant from year to year. This may be more challenging for innovative measures that are still being developed, refined, and field-tested. For example, California Office to Reform Education or "CORE" districts plan to implement metacognitive assessments into their accountability system, but they have not accumulated longitudinal data sufficient to ascertain how stable such measures will be. The multidimensional nature of college and career preparedness and the fact that definitions of this phenomenon are still evolving may mean that no single measure will be stable enough over time to address all needs to establish trends. It may be necessary to triangulate innovative measures against more traditionally stable measures for a period of time, both to establish the stability of the innovative measures and to transition beyond the traditional measure or redefine its utility.

B. Stakeholder Relevance: Accountability measures provide greater value to education systems when they are relevant to a variety of stakeholder groups. To the extent that measures can serve multiple purposes, they may help increase stakeholder acceptance of an accountability system. Many innovative measures have the characteristic of being important, particularly to students and teachers. They can reflect many of the aspects of teaching and learning that motivate teachers to enter teaching in the first place. They can also represent behaviors that students understand, and they are more clearly and directly associated with college and career preparedness than are many of the tests used currently to judge preparedness.

B1. Value to Students: This evaluative criterion is chiefly concerned with the extent to which component measures of the CCI are likely to be actionable, accepted, and valued by students. A CCI that incorporates value to students reflects and creates incentives for behaviors and performances that directly affect or improve individual prospects for preparedness to succeed after high school.

B2. Public Understanding: The API is intended to give educational stakeholders—educators, parents, students, and the public at large—a clear picture of a school's status and growth. Therefore, the CCI should communicate clearly how it supports college and career preparedness in order to be easily understood by laypersons as well as educators. In the instance of innovative measures, some of them are easily understood, while others will be new to educators and laypersons alike. Increasing public understanding will likely occur as measures are included, explained, and repeated over time.

B3. Content, Skills, and Competencies: In order for a revised API that consists of more than test scores to provide a valid description of school quality, its multiple measures must seek to gauge more fully the range of skills and competencies that are taught and learned in schools. Innovative measures presented in this white paper are not necessarily equally useful for rating content, skills, and competencies. Innovative measures are generally better suited to assess skills that transcend content knowledge and academic competencies. Some measures might assess all three in some instances, but not universally. Metacognitive assessments, such as inventories of student learning skills, are not measures of subject-matter knowledge; analyzing their value on the basis of their ability to evaluate content mastery would not add value to the API. Such a measure would be useful, however, when gauging skills. In other words, some innovative measures will provide insight into skills without revealing much about content knowledge.

B4. Emphasis on Student Performance: The legislative charge to California's school accountability system is to focus on educational outcomes rather than inputs. As important as it is to account for different features of quality schooling (e.g., teachers, instructional resources, curriculum, and school organization), this evaluative criterion looks at the extent to which potential component measures of the CCI emphasize student performance. Innovative measures use a more expansive conception of student performance, one that extends beyond tests of content knowledge to include other aspects of performance.

C. System Utility: Measures to be included in an accountability system have greater utility if they add minimal burden to the education system yet reflect the performance of as many students as possible. For the purposes of this white paper, a measure's system utility is also a function of the degree to which it provides information on students who will pursue a variety of postsecondary pathways.

C1. Minimal Burden: Minimizing the burden of an indicator means constraining the time and cost of implementation and data collection based on what schools can manage. For example, this criterion considers the overall amount of time necessary to prepare for and take a test, and the test's direct and indirect effects on students, teachers, administrators, and the system. As innovative measures have not been implemented at scale for accountability purposes, this white paper relies on research conducted in a variety of school settings and on evidence from classrooms, schools, and systems.

C2. Student Coverage: The API Guiding Principles state that the API should include as many students as possible in each school and district. This inclusion principle was cornerstone to an accountability system based entirely off universal measures (e.g., all students must take state assessments including populations requiring testing accommodations). A CCI by definition includes measures that are not universal because not all students can or should be compelled to go to college immediately upon completion of high school. Students and their parents retain the right to choose which path makes the most sense for them, and college is one option among many. In addition, students can demonstrate preparedness through an array of measures that address different knowledge, skills, and aspirations associated with postsecondary success. This analytic criterion gives preference to scaled or scalable measures over local and unique ones. Consequently, this criterion considers innovative measures differentially than more universal measures. Therefore, when applicable, student coverage will be discussed in terms of the potential for student participation.

C3. Postsecondary Pathways: The last criterion is less a measure than a global determination of the overall validity of the CCI on the basis of whether the indicator encompasses the full range of postsecondary pathways.

Determining the appropriateness of any potential indicator is not simply a matter of rating it on each of these criteria and then summing up the score. While each alternative measure will be analyzed based on the research and practices that support it and then rated on the three-point scale described earlier, overall judgments will be holistic and interpretive. The analyses may reach conflicting conclusions (e.g., a measure may be outstanding in one area but have a potentially fatal weakness in another). The purpose of this work is not to reach a summary judgment on each alternative measure, but to equip decision makers with sufficient information to consider and deliberate strengths, weaknesses, and trade-offs associated with a range of measures that have the potential to be included in the CCI.

Next, this white paper turns to a systematic analysis of a number of alternative measures, including metacognitive measures, performance assessments, and the California State Seal of Biliteracy, against the described analytical framework and rating criteria. The white paper concludes with a summary of these innovative measures in relation to these rating criteria.

Innovative Measure 1: Metacognitive Assessments

Metacognitive skills, also known as "noncognitive learning skills" or "21st-century skills," refer to the learning strategies, attitudes, and behaviors students employ and improve upon during the learning process. Although the term "noncognitive" is more familiar as a result of work by economist James Heckman, in this white paper EPIC uses the term "metacognitive" to reflect that these activities require significant cognitive processing to complete. Specifically, when students reflect on their own thinking they are engaged in more and deeper cognitive processing than when they are retrieving information from long-term memory and processing it in working memory, the type of cognition that content tests typically elicit.

Researchers in sectors other than education have noted the importance of metacognitive learning skills (Almlund, Duckworth, Heckman, & Kautz, 2011; Barrick, Mount, & Judge, 2001; Goldberg, 1992; Lindqvist & Vestman, 2011; Peterson et al., 1997). Research in the field of education has now begun to focus more on the importance of metacognitive learning skills such as conscientiousness (Poropat, 2009), self-efficacy (Multon, Brown, & Lent, 1991), grit (Duckworth, Peterson, Matthews, & Kelly, 2007), and others (Robbins et al., 2004). These skills positively predict college grades and retention. Furthermore, some research suggests that metacognitive ability can be more important than cognitive ability in explaining success in the labor market (Lindqvist & Vestman, 2011).

The educational standards movement, which began in earnest in the early 1990s, created a need for measures of student performance on the standards. Many of the standards were sufficiently expansive to require measures beyond criterion-based content tests. States such as New York, Connecticut, Maryland, Ohio, Vermont, Kentucky, and California all developed assessment systems that included more complex measures of student learning that in many cases required learners to reflect upon the ways in which they solved problems or developed solutions in addition to answering questions correctly (Darling-Hammond & Adamson, 2010). However, since the passage of the No Child Left Behind Act (NCLB) of 2001, which required all states to develop standards in mathematics and reading and to test students in Grades 3–8 and once again in high school, states have relied exclusively on standardized tests of content knowledge to evaluate learner progress and school quality. NCLB focused on content knowledge assessment to the exclusion of measures of thinking skills or other metacognitive skills associated with college and career preparedness (Conley & Darling-Hammond, 2013). While standardized test scores are useful as one source of information on school quality, such scores reveal only a partial picture of student learning and school quality. Comprehensive literature reviews have shown the importance of metacognitive factors (Farrington et al., 2012; Pellegrino & Hilton, 2012) in an overall picture of school and instructional quality.

Metacognitive assessments can complement standardized content knowledge tests when administered in low-stakes environments. While high-stakes content assessments do provide useful information to educators about what students know, they are not particularly useful in helping educators understand why students are or are not learning effectively (Conley & Darling-Hammond, 2013). Measures that provide insight into student mastery and use of specific learning skills can be much more informative about why students performed well or failed to perform well on tests. So can information on student attitudes toward learning, such as whether learners attribute their success to aptitude or effort (Pecheone, Kahl, Hamma, & Jaquith, 2010).

Gathering information on student metacognitive skills presents its own unique challenges (Soland, Hamilton, & Stecher, 2013). Recall that standardized achievement tests were first introduced and adopted on a large scale by U. S. schools in the 1950s and have been in the processes of being improved and refined ever since. Legions of technicians, social scientists, statisticians, and educators have devoted time and energy to developing new versions of these tests. Numerous companies have made it their business to generate revenue from these tests. As a result, the tests tend to improve over time in terms of their technical adequacy, and they certainly are by now much more familiar to and accepted by educators and parents. These tests have been woven into the fabric of schools and schooling, perhaps grudgingly at times, and nearly everyone has grown up taking them and therefore have firsthand experience with them, whether positive, negative, or indifferent. This ability to meet technical standards, coupled with widespread familiarity, has led to an acceptance of these types of tests as the best, most useful, and most accurate portrayal of student knowledge that exist. All other measures tend to fall somewhat short in comparison, not necessarily because they are not potentially as valuable or more valuable, but because they lack the same technical rigor and are far less familiar.

Measures of metacognitive skills in particular have not received anywhere near the same attention from psychometricians nor had the resources from testing companies devoted to their development and refinement as have content knowledge tests (Pellegrino & Hilton, 2012). As a result, far less progress has been made evolving and adapting these types of instruments to meet higher technical standards and to increase familiarity with them. Not surprisingly, policymakers have demonstrated considerable reluctance to include metacognitive measures in accountability systems or to encourage their use broadly as performance measures in schools. This sends a signal to educators that the information gained from a metacognitive assessment might be less valid or valuable (Conley, 2013). The result is a cycle in which such measures are not seen as technically rigorous, are not used, and therefore the technical rigor is never improved. Use remains limited to boutique schools and esoteric settings, and the general public's familiarity with these instruments never increases.

Recently, however, researchers have begun to paint a clearer and more compelling picture of the potential contributions that metacognitive measures might be able to make to an overall picture of student achievement and preparedness. Policymakers and educators seeking to measure metacognitive skills have at their disposal, as the Asia Society and RAND Corporation put it, "a dizzying array of options" from which to choose (Soland et al., 2013, p. 9). The Educational Policy Improvement Center (EPIC) identified 143 assessments that were claimed to be measures of metacognition, "soft skills," interpersonal skills, intrapersonal skills, 21st-century skills, or some other type of skill not based on content, and reported the results in an unpublished paper (Conley, Gilkey, Seburn, Bryck, & Shanley, 2012). Of these assessments, EPIC reviewed 33 against 12 evaluative criteria.[1] These formative or summative assessments came in a variety of formats including multiple-choice tests, self-report questionnaires, closed-ended computer-based items, video games, and performance tasks. Table 1 provides some representative examples of the 33 metacognitive assessments EPIC reviewed. Each identifies the skills it assesses. An expanded list of metacognitive assessments can be found in Soland et al. (2013).

[1] Predictive validity, reliability, fairness, resistance to faking, administrative feasibility, operational costs, population and subpopulation, item and response types, delivery mode, scoring method(s), exemplary components, and strengths and weaknesses.

Table 1. Examples of Metacognitive Assessments

Assessment	Format	Metacognitive skills assessed
ACT®: ENGAGE	Student self-report	Motivation, social engagement, self-regulation
EPIC: CampusReady™	Student, teacher, counselor, and administrator self-reports	Key learning skills and techniques
H&H: Learning and Study Strategies Inventory (LASSI)	Student self-report	Attitude, motivation, time management, anxiety, concentration
ETS: Standardized Letters of Recommendation (SLR)	Teacher-generated ratings	Creativity, communication, motivation, self-organization, teamwork

All of the assessments in Table 1 have research bases to support them. For example, ACT's ENGAGE (formerly known as the Student Readiness Inventory) was developed with a constructed validation approach using meta-analytic research on motivation and academic-related and social engagement skills (Le, Casillas, Robbins, & Langley, 2005). The original item pool for the Learning and Study Strategies Inventory (LASSI) was developed from an analysis of existing instruments and inventories that measured study skills and learning strategies (Weinstein & Palmer, 1990). The SLR was created after studying the limitations of letters of recommendations, researching the relationship between metacognitive skills and graduate admission, and surveying graduate faculty and administrators (Walters, Kyllonen, & Plante, 2006).[2]

EPIC's CampusReady school diagnostic is derived from research conducted in the College Readiness Evaluation for Schools and Teachers (CREST) study, which analyzed programs and practices at 38 carefully selected high schools that consistently graduated college-ready students from underrepresented groups. CampusReady also drew from multiple analyses of the content of college courses and other source material. Since Fall 2013, more than 43,000 students, 3,700 teachers, 300 administrators, and 270 counselors at 148 schools in 20 states have used CampusReady to launch, plan, and prioritize their college and career preparedness goals. Lombardi, Seburn, and Conley (2011) found that CampusReady is a reliable measure of goal-driven behaviors, persistence, study skills, and self-monitoring.

Metacognitive skills have yet to be included in any statewide accountability system. However, seven California school districts that serve more than one million students, known as the California Office to Reform Education or "CORE" districts, have designed an accountability system to meet federal requirements that plans to measure student metacognitive skills. The CORE districts were granted a waiver from NCLB requirements based on a plan that includes a School Quality Improvement Index comprising two domains: Academic and Social-Emotional & Climate/Culture. When the School Quality Improvement Index is fully implemented in the 2014–2015 academic year, the Social-Emotional & Climate/Culture domain will consist of absentee rates; suspension/expulsion rates; English learner redesignation rates; special education identification rates; student, staff, and

[2] The SLR is included in this list for illustrative purposes to give an example of a teacher-generated assessment that could be triangulated against other information, such as academic measures, but not used on its own.

parent climate/culture surveys; and metacognitive assessments. The Social-Emotional & Climate/Culture domain will account for 40% of a school's School Quality Improvement Index score.

Currently, the CORE districts are piloting four initial metacognitive assessments across 20 schools, using two versions of each metacognitive assessment. For each metacognitive assessment, one version has been selected from existing measures; the other version has been developed in partnership with methodological experts in an effort to improve upon existing measures. Table 2 shows the existing assessments being piloted by the CORE districts. These measures consist of teacher reports for students in grades K–12, plus student self-reports for students in grades 5–12.

Table 2. CORE District Metacognitive Assessments Currently Being Piloted

Developer: Assessment	Format	Metacognitive skills assessed
Chicago Consortium on Chicago School Research (CCRS): Becoming Effective Learners Project	Student self-report	Growth mindset
Chicago Consortium on Chicago School Research (CCRS): Becoming Effective Learners Project	Student self-report	Self-efficacy
Angela Duckworth: Character Growth Card	Student self-report and teacher report	Self-management
Collaborative for Academic, Social, and Emotional Learning (CASEL) and American Institutes for Research (AIR): Collaborating Districts Initiative	Student self-report and teacher report	Social awareness

A. Technical Quality: A1. Relationship to Postsecondary Success

This section describes the correlational and theoretical research bases showing the relationship between metacognitive factors and college and career postsecondary success and offers an overview of the validity research around the assessments presented in Table 1.

Researchers from industrial-organizational psychology, developmental psychology, human resource development, and economics have analyzed the relationship between metacognitive skills and postsecondary success (Pellegrino & Hilton, 2012). Across these academic disciplines, many metacognitive skills have been shown to predict college and career success. The most widely researched and validated set of personality traits associated with academic and workplace success are known as the "Big Five": conscientiousness, openness, agreeableness, emotional stability, and extroversion (Goldberg, 1993; McCrae & Costa, 1987). Of these five traits, conscientiousness, which is defined as being well organized and taking responsibility for one's learning, has emerged as the best predictor of overall attainment and achievement in a variety of settings, including job performance across a broad range of occupational categories (Almlund et al., 2011; Poropat, 2009; Barrick et al., `2001).

In addition to conscientiousness, the ability to persevere when confronted with challenges, also known as *grit* or *persistence*, has shown a strong positive relationship with academic outcomes in a wide range of settings such as retention at West Point or success in the National Spelling Bee. Duckworth coined the term grit to describe the quality displayed by students who overcome obstacles to achieve success (Duckworth et al., 2007; Duckworth & Quinn, 2009). Grit has been found to be a better overall predictor of academic achievement than cognitive ability (Duckworth & Quinn, 2009).

A meta-analysis conducted by Robbins et al. (2004) synthesized 109 studies from educational persistence and motivational theory, analyzing the relationship between two college outcomes (cumulative GPA and retention) and nine psychosocial and study skills factors (PSFs): achievement motivation, academic goals, institutional commitment, perceived social support, social involvement, academic self-efficacy, general self-concept, academic-related skills, and contextual influences. The study found that academic self-efficacy significantly predicted both college outcomes. Academic goals and academic-related skills significantly predicted college retention, while achievement motivation significantly predicted cumulative GPA.

Assessing metacognitive skills for purposes other than formative feedback presents its own set of unique challenges, and more research will be necessary to understand the effects of using data from metacognitive assessments in accountability systems. However, several assessments have been shown to be a significant predictor of college grades and retention. ACT's ENGAGE is one of these (Peterson, Casillas, & Robbins, 2006; Robbins, Allen, Casillas, Peterson, & Le, 2006). The CampusReady instrument was found to be a reliable measure of goal-driven behaviors, persistence, study skills, and self-monitoring, and also a significant predictor of college success (Lombardi et al., 2011). Research on the LASSI found that eight of its ten subscales significantly predicted college GPA, the exceptions being anxiety and selecting main ideas (Griffin, MacKewn, Moser, & VanVuren, 2012).

This research suggests that measuring student metacognitive skills provides a better overall insight into potential success than do cognitive measures on their own. At the very least, it seems that combining information from measures of metacognitive skills with information from content tests would lead to a more complete picture of student readiness.

Rating: Moderate

A. Technical Quality: A2. Fair Comparisons

Fairness is the degree to which a metacognitive assessment is unbiased to various subgroups (e.g. gender, race and ethnicity, socioeconomic status, English learners, and students with disabilities). Le et al. (2005) analyzed whether ACT's ENGAGE allowed for fair comparisons across different subgroups (men-women, high school-community college-university students, and majority-minority). The results showed statistically significant differences between subgroups, but effect sizes of these differences were so small the authors concluded that the differences were of "little *practical* significance" (Le et al., 2005, p. 503, emphasis added).

Lombardi et al. (2011) conducted separate multivariate analyses of variance tests to determine if the characteristics of race, gender, and first-generation status predicted goal-driven behaviors, persistence, study skills, and self-monitoring in the high school grade constructs in CampusReady.

Significant differences were found in Grade 9 between genders for Hispanic/Latino students. However, these differences did not persist throughout high school, suggesting that all students, regardless of characteristics, may benefit from being taught the importance of academic behaviors.

The only relevant research addressing whether LASSI allows for fair comparisons across different subgroups found females significantly outscoring males across eight LASSI subscales: attitude, concentration, information-processing skill, motivation, self-testing and review techniques, use of study-support techniques, time management, and effective test-taking strategies. The research found that the significant difference between female and male academic performance disappeared after controlling for the variance explained by LASSI scores. This suggests that learning and study strategies may explain previous research findings (Leonard & Jiang, 1999) showing females outperforming males academically (Griffin et al., 2012).

The limited research available to date suggests that metacognitive assessments can be administered in a way that ensures fairness across subgroups of students. Results from the CORE district field-testing and implementation may provide greater insight into large-scale incorporation of different types of metacognitive assessments.

Rating: Moderate, with promising but insufficient evidence

A. Technical Quality: A3. Stability

Instruments used to assess metacognitive skills need to yield reliable results over time to be considered stable measures of college and career preparedness. To date, no metacognitive assessment has been employed in a statewide accountability setting. However, the California CORE districts' incorporation of metacognitive assessments into the School Quality Improvement Index will soon provide some initial longitudinal data to help analyze the reliability and stability of metacognitive assessment as contributors to the API.

One of the main concerns about most metacognitive measures is the potential for social-desirability bias, or faking, which is the potential for students to report what they think are desirable answers on self-report questionnaires. Research shows that students have a clear sense of the personality attributes of an "ideal" student (Huws, Reddy, & Talcott, 2009).

Faking becomes more of an issue as stakes increase. One way to confirm to some degree the responses on a self-report instrument is to triangulate data with results from other, complementary measures. For example, student self-reports can be compared to teacher reports of student characteristics such as persistence and goal focus. Additionally, self-reports can be compared to scores on content tests to help spot serious inconsistencies. When too many inconsistencies exist for a school, the overall results could be called into question. A variation on this approach is used statewide in Victoria, Australia, to compare the results of teacher-marked collections of student work against a low-stakes statewide reference exam that tests cognitive skills and strategies (Darling-Hammond & Adamson, 2010). In cases where serious anomalies are found, the state can either adjust scores or retrain teachers in scoring techniques. In the U. S., results from SBAC and PARCC tests could serve the same purpose in the future, as an external benchmark to judge the overall validity of self-reports or teacher ratings of student metacognitive skills.

Another potential strategy to make inferences about student metacognitive skills is from assignments requiring such skills. For example, a research paper that takes multiple drafts to complete could be used as an indicator of both time management and persistence. Did the student complete all drafts on time and with high quality? Did the student give up and fail to complete the final draft, or was the final draft of such low quality that it was evident that the student did not spend sufficient time on it? Similarly, students could be asked to set short-term and medium-term goals in class, and teachers (and students) could observe if they achieved or worked toward those goals. These types of ratings could not be as easily faked, although they would still be dependent on the integrity of educators to rate students accurately and honestly.

Given the current lack of longitudinal data available to substantiate stability and the potential vulnerability to faking without safeguards, metacognitive assessments are rated currently as weak. This rating does not imply that metacognitive assessments cannot reach the same level of stability as multiple-choice standardized assessments. More research, new techniques and strategies, large-scale experimentation, triangulation, and refined instrument development will likely improve the stability of metacognitive assessments as potential statewide measures of college and career preparedness.

Rating: Weak

B. Stakeholder Relevance: B1. Value to Students

Although the vast majority of postsecondary institutions do not currently award students credit or scholarships for metacognitive assessment scores,[3] institutions such as Boston College, DePaul University, Tufts University, and Oregon State University (OSU) use information gleaned from metacognitive assessments for admission purposes (Tomsho, 2009). In its 2009 admissions cohort, DePaul used four personality assessment questions (e.g., How would you compare your educational interests and goals with other students in your high school?) to admit about 150 students whose tests scores made them marginal applicants and to screen out about 50 applicants whose responses were judged to be "lackadaisical." In 2004, OSU began asking applicants to complete its six-item Insight Resume, which is designed to measure capacity to deal with adversity. Two admissions counselors score each applicant's response to prompts such as one's experience facing/witnessing discrimination and one's response to it. Unlike DePaul's approach, OSU does not disqualify students on the basis of its Insight Resume, instead using the measure to attract and keep minority and low-income applicants or those who do not meet typical grade or test-score thresholds for admissions.

Developing metacognitive skills also has employment value for students. For instance, employers are increasingly using assessments such as ACT's WorkKeys to measure both the foundational and "soft" skills necessary to be successful in the workplace (ACT, 2014). Furthermore, 225 recently surveyed U.S. employers placed high value on communication skills, positive attitudes, solid teamwork skills, and the ability to think critically to solve problems (Millennial Branding, 2012).

Rating: Moderate

[3] One exception: the Jack Kent Cooke Foundation awards college scholarships based on metacognitive attributes including persistence.

B. Stakeholder Relevance: B2. Public Understanding

Metacognitive assessments to date have not been implemented in any state accountability system, so a low baseline level of public understanding can be expected. However, the concept of metacognitive skills is intuitive, and students, parents, educators, and policymakers observe these phenomena daily. Many of these skills have long been rated on primary grade report cards, but then disappear by middle school in favor of letter grades on purely academic indicators. Most educators and parents would take notice if a student showed a marked increase or decrease in motivation, self-efficacy, conscientiousness, or grit. They understand that these are important. The value of feedback to students and teachers alike on metacognitive skills is that such feedback raises awareness of the importance of these skills and prompts educators to improve these skills among students. Although public understanding of metacognitive assessments will be initially low, it may be reasonably easy to increase understanding relatively rapidly.

Rating: Moderate currently, strong post-implementation

B. Stakeholder Relevance: B3. Content, Skills, or Competencies

Metacognitive assessments do not measure content knowledge or competencies. Instead, they provide insight into the strategies, dispositions, and behaviors that students apply when learning academic content. Understanding how students employ metacognitive skills can be an important source of information to inform development of effective strategies for teaching content knowledge. Incorporating metacognitive assessments into the API will send a signal that developing these learning skills is a valuable educational endeavor. Such measures may also serve as a "canary in the coal mine," signaling both the overall health of a school and potential problems in schools that see rapid decreases in scores on metacognitive measures even as their academic scores may remain relatively constant.

Rating: Moderate

B. Stakeholder Relevance: B4. Emphasis on Student Performance

The formative nature of metacognitive assessments provides educators with information on how to improve student learning, which subsequently leads to improvements in student performance on content knowledge assessments while also developing the 21st-century skills necessary for career success. Metacognitive assessments provide unique insights into why students are or are not learning key content knowledge. While a content knowledge test tells which questions a student answered correctly or incorrectly, it does not provide insight into why the student learned some material but failed to learn other material. Skillful teachers develop insights into the causes of academic deficiencies, but measures of metacognitive skills hold the promise of helping educators understand more systematically which learning skills are contributing to success and the lack of which skills are hindering achievement. This is particularly important in schools with large concentrations of students from groups historically underrepresented in college. These students currently tend not to get much training in metacognitive skills because their schools may be concentrating on content knowledge transmission geared to state tests. These students and their teachers could conceivably benefit from more information about the reasons students are doing well or struggling.

Rating: Moderate

C. System Utility: C1. Minimal Burden

Other than test-taking time, the burden of metacognitive assessments is minimal. Most such assessments do not take much time to complete.

Assessment format determines the direct costs to districts. These costs vary greatly across types of metacognitive assessments. More expensive assessments, generally those using complex computer-based systems, surpass administration costs of traditional standardized tests. Indirect costs include staff time required to administer these assessments, any scoring that needs to be completed by school or district staff, and the opportunity cost associated with reallocating time from direct instruction (Soland et al., 2013). EPIC's review of metacognitive assessments found that most are student self-report using closed-ended response options that can be scored by computer, greatly reducing direct costs to districts (Conley et al., 2012).

Direct costs to the CDE depend on the level of stakes attached to metacognitive assessments. At a minimum, direct costs to the CDE include collecting and reporting results. Again, these costs vary by format. Furthermore, if districts are free to choose from a menu of metacognitive assessments, the CDE will likely need to provide additional staff time and resources to equate results across assessments and then provide materials so that various stakeholders might interpret results correctly.

Subsuming metacognitive measures into extant large-scale cognitive assessments could eliminate many direct costs to districts and the CDE. One such large-scale example, ACT's WorkKeys, measures both cognitive skills (such as applied mathematics and business writing) and "soft skills" (including motivation, integrity, and interpersonal interaction) to complement the results of the cognitive portion of the assessment. Both the SAT and ACT include a range of optional items in which students report attitudes and behaviors. This addition of metacognitive items to content knowledge tests could result in some sufficiently valid and useful information on a set of basic metacognitive skills being gathered in a cost-effective manner that could serve as a baseline or starting point for other universal instruments or measures at the school level.

Rating: Moderate

C. System Utility: C2. Student Coverage

Statewide student participation numbers have not been collected for metacognitive assessments due to the absence of large-scale implementation. However, metacognitive assessments do have the potential to become universal measures of college and career preparedness. Because metacognitive skills are associated with both college and career success, requiring all students to take a metacognitive assessment would not force students into postsecondary pathways they do not want to pursue. Furthermore, including one universal measure within the CCI would provide policymakers with information for both college and career preparedness across schools.

Rating: Strong, with potential for universal coverage

C. System Utility: C3. Postsecondary Pathways

Including metacognitive assessments in the CCI could be valuable for students pursuing both college and career-going postsecondary pathways by leading teachers and students alike to focus more time on developing and mastering key metacognitive skills. Essentially all the metacognitive

skills discussed in this white paper are applicable to both college and careers. For example, conscientiousness best predicted college *and* career success in a comprehensive meta-analysis (Almlund et al., 2011). Metacognitive assessments are the only measures considered for inclusion in the CCI that provide insight directly into preparedness for both college and career-going postsecondary pathways.

Rating: Strong

Summary

Overall, the research suggests that metacognitive assessments hold great potential as a means to drive improvement in student achievement. Table 3 presents a summary of the evaluative criteria ratings. Metacognitive assessments were rated strongest on being understandable, recognizing both postsecondary pathways, and the potential for universal student coverage. Metacognitive skills are strongly related to college and career success, but less evidence is available for metacognitive assessments. A relative unknown is the stability of large-scale metacognitive assessments.

Table 3. Metacognitive Assessment Evaluative Criteria Ratings

A. Technical quality			B. Stakeholder relevance				C. System utility		
A1	**A2**	**A3**	**B1**	**B2**	**B3**	**B4**	**C1**	**C2**	**C3**
Moderate	Moderate	Weak	Moderate	Moderate	Moderate	Moderate	Moderate	Strong	Strong

The standards and accountability movement has tended to exclude measures of metacognitive skills even though many standards imply that such skills are necessary to master the standard. As a result of the NCLB flexibility waiver, the California CORE districts are set to become the first large-scale test of the use of metacognitive assessments for accountability purposes. Many valid and reliable metacognitive assessments are currently available to educators and policymakers, but more research and experimentation is necessary to understand fully the value and limitations of administering metacognitive assessments and using results for accountability purposes.

Metacognitive assessments have the advantage of producing actionable information that students and educators can put into practice immediately to improve achievement. Metacognitive skills generally are connected with success in college, careers, and life. These skills and their attendant measures are important and often overlooked components of effective learning systems.

Metacognitive assessments provide three distinct advantages over standardized assessments by (a) producing actionable information that students and educators can put into practice immediately to improve achievement; (b) connecting to success in college, careers, and life; and (c) being understood by all stakeholders as having utility both within and outside school settings. When a state operationalizes metacognitive skills as part of its accountability system, these skills will realize increased importance resulting in increased technical quality for their attendant measures. Until then, metacognitive skills may remain as overlooked components of effective learning systems.

Innovative Measure 2: Performance Assessments

The Common Core State Standards (CCSS) specify the concepts and skills needed for success in the 21st century. By creating a system of fewer, clearer, and higher standards geared to college and career readiness, the CCSS seek to encourage deeper learning within schools. In fact, research demonstrates the ways in which deeper learning skills are required to master the CCSS (Conley, 2011). Two consortia of states, the Smarter Balanced Assessment Consortium (SBAC) and the Partnership for Assessment of Readiness for College and Careers (PARCC), are creating new assessments designed to assess many of the CCSS. However, Conley and Darling-Hammond (2013) critique both assessments as being unlikely to cover some standards (e.g., communication, collaboration, and problem solving) because they cannot be validly measured using the SBAC and PARCC assessments designs. More cognitively demanding standards can be assessed only in the context of more authentic and complex student work products. A system of assessments that gauges the development of deeper learning skills and provides useful diagnostic information for instructors will require as one component information from student work that can only be completed over a more extended period of time (Conley & Darling-Hammond, 2013).

Performance assessments (also referred to as performance-based assessments or performance tasks) are designed to cause students to construct original responses to authentic problems. They can consist of simple tasks completed in a single class period, semester-long research projects, or many options in between. The more extended period of time that students devote to performance tasks permits a great deal more cognitive processing to occur. As a result, these tasks are much better suited to measuring the thinking and reasoning skills critical for college and career preparedness. Performance assessments are not the same as local teacher-generated assignments. The content focus and technical quality of externally designed performance assessments is more tightly controlled, the conditions for administration more highly specified, and the scoring methods more systematic and consistent. The result is scores that are more valid and reliable than those from teacher-designed assignments (Conley, 2013).

The 1990s witnessed the high point of performance assessment use in schools and in state testing and accountability systems. The National Science Foundation's Systemic Science Initiative funded states' development of "hands-on" science and mathematics assessments. This funding helped Connecticut, Maine, Vermont, New Hampshire, Rhode Island, Missouri, Kentucky, New York, and Ohio develop and use performance-based assessments (Darling-Hammond & Adamson, 2010). The short-lived California Learning Assessment System (CLAS) was recognized for the design quality and challenge level of its performance tasks. CLAS was designed to be an improvement over the California Assessment Program (CAP) by providing a more accurate measure of student content knowledge though the use of performance assessment (Kirst & Mazzeo, 1996). Oregon's Certificate of Initial Mastery incorporated a series of performance tasks in mathematics and English, and combined the scores with the results from multiple-choice tests. A number of states instituted culminating performance-based projects as graduation requirements, most prominently Massachusetts, Oregon, and Washington (Conley, 2013). States that adopted performance assessments in the late 1980s and early 1990s had largely abandoned them by the end of the decade for a variety of reasons including technical adequacy, cost, training and scoring demands, reporting issues, and, ultimately, the looming requirements imposed by NCLB. These issues will be discussed

in depth in A3. One of the longer-lasting large-scale performance assessments, the Maryland State Performance Assessment Program (MSPAP), ended in 2002.

Pecheone et al. (2010) discussed eight examples of promising practices from states that use performance assessments for high-stakes purposes. In New York, a network of 27 schools formed the New York Performance Standards Consortium. Instead of taking the traditional New York Regents exams required to graduate, students in these schools complete and defend a graduation portfolio that contains performance assessments including scientific investigation, a mathematical model, a history or social science research paper, and a literary essay. New Hampshire has developed a competency-based system for graduation that no longer relies on Carnegie units (also called credit hours). These students now earn course credits by taking course-based performance assessments both in and out of school.

Performance tasks can be found in a number of California school districts. For example, many use performance assessments from the Mathematics Assessment Resource Services (MARS) program, which includes the Mathematical Assessment Project (MAP) Summative Assessment Tasks. Created collaboratively by researchers at the University of California, Berkeley, and the University of Nottingham Shell Centre, the MAP Summative Assessment Tasks require students to apply complex knowledge and skills to solve performance-based problems (Darling-Hammond & Adamson, 2010). More recently, the Envision Schools Charter Management Organization (ESCMO) in San Francisco collaborated with the Stanford Center for Assessment, Learning, and Equity (SCALE) to develop the College Success Portfolio, a graduation requirement for students in four charter schools. Summit Public Schools is a network of charters with seven schools in California that emphasizes the use of performance-based projects designed to assess deeper learning (Summit Public Schools, 2014). The College Success Portfolio includes performance assessment outcomes in six core content areas: ELA, mathematics, science inquiry and science literacy, history-social science, foreign language, and the arts.

Many high-achieving educational systems, including Finland, Singapore, Hong Kong, Australia, and England, use performance tasks to assess higher-order thinking skills (Darling-Hammond & Wentworth, 2010). England's General Certificate of Secondary Education (GCSE) functioned as a touchstone for the creation of similar assessments in Australia, Hong Kong, and Singapore, as well as for the International Baccalaureate (IB) and the New York State Regents examinations. Both the GCSE and the IB Diploma Programme are two-year courses of study assessing students within and at the end of courses using open-ended items and extended classroom-based performance tasks. GCSE's assessments are either created by educators and scored by an awarding body, or designed by the awarding body and scored by educators. IB develops its own assessments but solicits critique from teachers after the culmination of each assessment. New York involves educators in the development and scoring of items and tasks of the Regent examinations (Darling-Hammond & Adamson, 2010).

In addition to these state and international examples, other organizations have developed performance assessments and task banks. For instance, EPIC's ThinkReady (formerly known as C-PAS) uses performance tasks to measure a student's mastery of five key cognitive strategies (problem formulation, research, interpretation, communication, and precision and accuracy) measured on a novice-to-expert continuum. Other organizations with expertise in performance assessment, such as SCALE, the Asia Society, the Literacy Design Collaborative, and the Center for Collaborative Education, partner with schools, districts, and states to provide resources that enable

teachers to create and manage performance assessments. SCALE partners with SBAC and CTB/McGraw-Hill to develop the performance tasks that will assess student knowledge of the CCSS in 25 states. EPIC and SCALE have partnered with the Council of Chief State School Officers (CCSSO) to work with a group of nine states including California that collaborate to identify and implement student-centered education reforms. These Innovation Lab Network states agree to create innovation zones in their states where schools can experiment with implementation of performance tasks as assessments that can provide information useful to the state as well as to the schools that administer them.

A. Technical Quality: A1. Relationship to Postsecondary Success

Performance assessments have the potential to measure deeper learning, something standardized tests cannot do as well (Conley, 2013; Darling-Hammond & Adamson, 2010; Darling-Hammond & Pecheone, 2009; Lane, 2010). Pellegrino and Hilton's (2012) comprehensive literature review on deeper learning defines it as including cognitive, interpersonal, and intrapersonal skills, and finds that mastery of these skills correlates positively with educational, workplace, and life outcomes. However, the authors suggest that the limited evidence is uneven and call for foundations and federal agencies to support more research establishing the connection between 21st-century skills and educational, workplace, and life outcomes.

The greatest strength of performance assessments is their ability to be highly valid representations of the cognitive processes they are designed to measure (Lane, 2010). The most prevalent criticism of them is the reliability of scoring (Darling-Hammond & Adamson, 2010). Studies of the predictive validity of performance assessments have yielded positive results. Goldschmidt, Martinez, Niemi, and Baker (2007) found that 9th grade performance on an English language arts performance assessment predicted 10th grade scores on the California High School Exit Exam, after controlling for demographic characteristics and past performance. Kobrin, Patterson, Barbuti, Mattern, & Shaw (2008) determined that the writing section of the SAT was a more effective and consistent predictor of first-year college GPA than the standardized SAT mathematics or verbal sections. Likewise, Hojay et al. (2000) found that the writing section of the Medical College Admission Test (MCAT) predicted clinical science evaluations and ratings of clinical competence better than achievement on the reasoning sections of MCAT tests in biological sciences, physical sciences, and verbal ability.

One strength of performance assessments is the ability to create tasks that are valid representations of the type of tasks students typically encounter in college and careers. College assignments are quite often more like performance assessments than are the assignments students encounter in high school (McGaughy, 2014). Professional fields such as medicine and law have long used performance assessment for high-stakes decisions (Tung, 2010). Of 38 clusters of majors listed on the College Board website (http://collegeboard.org), half depend almost exclusively on performance as the primary means of assessment. Those clusters range from communication and visual/performing arts to engineering and natural science. Another 13 clusters of majors in areas such as English language and literature, history, or some social sciences may employ essay-based responses that generally can be considered performance tasks.

Career and technical education has a rich tradition of performance assessment. For instance, the National Occupational Competency Testing Institute (NOCTI) offers approximately 100 "job-ready" performance assessments across 15 occupational areas at both the secondary and postsecondary levels. These performance assessments are meant to replicate the tasks students will

encounter in the workplace. Federal requirements have spurred states to develop performance assessments designed to measure career and technical skill attainment. For example, students in Utah must pass both the multiple-choice and performance task sections of Career Technical Education (CTE) Skill Certification Tests to receive a certificate in one of eight CTE program areas.[4] Wyoming's Career Technical Assessment is solely performance-based and requires students to demonstrate generic skills across six content areas[5] (Klein, 2006).

Performance assessments have the potential to measure cognitive and metacognitive skills in a way that leads to greater insights into deeper learning than traditional standardized tests. They are common measures of knowledge and skill acquisition in many college disciplines and most career and technical fields, making their use at the secondary level important as authentic demonstrations of the knowledge and skills needed for college and careers. The predictive value of some types of performance assessments to later academic success suggests that performance tasks can be contributing elements to high school accountability systems that reference college and career preparedness.

Rating: Strong

A. Technical Quality: A2. Fair Comparisons

Fairness is the degree to which performance assessments are unbiased to various subgroups (e.g., gender, race and ethnicity, socioeconomic status, English learners, and students with disabilities). Bias concerns for performance assessments generally revolve around the degree to which the content of the prompt or wording of the task is potentially unfamiliar to certain subgroups of students. Raters that demonstrate systematic bias toward particular groups or subgroups are also a potential issue (Darling-Hammond & Adamson, 2010). The issue of scoring reliability will be explored in greater detail in the following section.

Well-designed performance assessments can improve accessibility for English learners and students with disabilities when compared with multiple-choice assessments (Darling-Hammond & Adamson, 2010). This is because performance assessments allow respondents to demonstrate knowledge in many ways, whether through a graphical display or a hands-on science activity. Multiple-choice assessments sometimes require selecting the "best" option among more than one plausibly correct answer. This introduces issues with the comprehension of complex linguistic features, such as passive voice and relative clauses. Such obscure and complex language increases difficulty for English learners and students with learning disabilities (Abedi, 2010). Performance tasks may help level the playing field by providing learners with multiple ways to comprehend the prompt. Goldschmidt et al. (2007) found that results on the ELA performance assessment were not sensitive to students' socioeconomic status or ethnicity. By contrast, ELA performance assessment scores were sensitive to student variables associated with English language proficiency, home language, immigrant status, and special education status. The authors do not speculate about why these subgroups of students performed below other subgroups of students, but research by Aguirre-

[4] Agricultural education, business education, family and consumer sciences, health science and technology education, marketing education, technology education, trade and technical, and information technology.
[5] Communication, applied math, affective and thinking, technology, pre-employment, and employability

Munoz et al. (2006) suggests that English learners or students with disabilities may struggle with the linguistic demands of the ELA performance assessment.

Scores on performance assessments have been shown to correlate less highly with student demographics than do standardized achievement test scores. Goldschmidt et al. (2007) found that the gap between white students, English-only students, and traditionally disadvantaged students was larger on the standardized Stanford Achievement Test, 9th edition, than on the language arts performance task portion of the California High School Exit Examination.

The potential to minimize systematic differences in performance among subgroups is a promising feature that warrants careful attention to the overall utility and value of performance assessments as one component in a larger system of assessments. The development of high-stakes performance assessments needs to include thorough field-testing to ensure that the linguistic demand embedded in the assessment is equal for all subgroups of students. Research shows that failing to do so will likely create unfair comparisons.

Rating: Moderate, promising but insufficient evidence

A. Technical Quality: A3. Stability

Many challenges would have to be overcome before a state would be able to implement a reliable performance assessment system. Designing and implementing performance assessments on a large scale poses a series of vexing challenges to states that wish to attempt them. They must assemble development teams capable of creating tasks that elicit the precisely desired responses in the content area and cognitive skill to be tested. They must field-test them, which is much more difficult because each task takes students much more time to complete, thereby taking away from classroom time, which means that far fewer schools are willing to field-test them. They must be replaced on a yearly basis if used for high-stakes purposes. Scoring them is challenging and requires significant resources for training of scorers and the actual scoring activity itself (Conley, 2013). All of this must be done to ensure a level of reliability and validity that will permit comparisons of scores from year to year. Oregon abandoned its mathematics performance tasks precisely because scores could not be made sufficiently comparable across years.

Research over the past two decades and examples from successful states do demonstrate that the challenges to implementing reliable performance assessments can be overcome. For example, common scoring guides, rubrics, and training can be created, and teachers can be trained to use them to generate consistent, reliable scores (Lane, 2010). Examples from states such as Kentucky show that achieving high rates of inter-rater reliability is possible by instituting a statewide audit system and investing in teacher training (Pecheone et al., 2010). Agreement on the content to be assessed and the conditions of administration also enhances reliability. Retaining high levels of reliability and consistency becomes increasingly challenging as the number of students assessed increases. However, advances in technology, such as computer-based training, calibration, and scoring have led to better methods of ensuring that performance assessments are valid and reliable measures of student achievement and growth (Darling-Hammond & Adamson, 2010).

The reliability of student scores is related to the number of performance tasks within an assessment. Stecher (2010) presents research showing that 2 to 20 performance tasks are required for reliable student scores and argues that there is no simple answer to how many tasks are needed for reliability

because tasks differ dramatically in content and format. Combining performance tasks with multiple-choice questions into one assessment may reduce the number of tasks needed for reliability (Stecher, 2010). Results from the CCSS assessments will provide additional information on the relationship between the number of tasks and student score reliability.

Although high-stakes performance assessments do exist (e.g., the New York Regents Examinations), none are available for immediate use in California. This is partly due to a lack of field-testing necessary to make valid and reliable generalizations in a statewide setting and also to the fact that local and state investments in assessment have flowed almost exclusively to the development of standardized multiple-choice tests. However, research and technological advances over the last two decades and successful examples from states in the late 1980s and early 1990s show that creating a system of stable performance assessments is conceivable and feasible.

Rating: Moderate, with promising but insufficient evidence

B. Stakeholder Relevance: B1. Value to Students

Postsecondary institutions and employers have long used versions of performance assessments to qualify applicants. For instance, many private and selective colleges expect students to submit a portfolio of work, including performance-based tasks such as research projects, along with traditional application materials (Ehley, 2006). Similarly, four out of five employers in a recent survey indicated that an electronic portfolio of student accomplishments would be useful to help ensure applicants possessed the knowledge and skills necessary to be successful employees (Hart Research Associates, 2013). Postsecondary institutions and employers see the importance in using performance assessments to screen potential applicants because these assessments provide additional information that cannot be gleaned from grades, references, or traditional standardized test scores.

Rating: Strong

B. Stakeholder Relevance: B2. Public Understanding

The public understanding, acceptance, and perception of performance assessment scores are largely dependent on the type of skill being measured. For instance, a performance assessment measuring the ability to solve a mathematics equation and apply the findings in a certain context is much more easily understood than a performance assessment measuring the ability to communicate or collaborate because the definitions of effective communication and collaboration are subjective. However, the concept of performance assessment is well understood by anyone who has taken a driving test, undergone CPR certification, auditioned for a school play, or competed in a tryout for a sports team. The public's general understanding of performance assessment results may be low initially, but is expected to increase as more educators integrate these assessments into their curriculum and policymakers disseminate information about the rationale, design, and intended use of performance assessments. The New York Regents Examinations are one of the few long-lasting statewide high-stakes performance assessment systems. They survive for a variety of reasons, not the least of which is their institutionalization and familiarity. All native New Yorkers who completed high school went through the Regents system. Although not universally loved, they are universally understood. They serve as a concept proof that more complex examination systems can work at a state level.

Rating: Moderate

B. Stakeholder Relevance: B3. Content, Skills, and Competencies

Performance assessments have the potential to measure both content knowledge and metacognitive skills in tandem. Performance assessments can provide educators with important information on how students apply learning strategies and skills to formulate responses to tasks. Performance assessments can offer formative value to educators by generating information necessary to improve student learning, in addition to providing summative information on content knowledge and skills. The degree to which performance assessments measure what is taught and learned in the classroom depends on the alignment between the curriculum and the performance tasks.

Rating: Strong

B. Stakeholder Relevance: B4. Emphasis on Student Performance

Well-designed performance assessments measure both content knowledge and metacognitive skills. They provide teachers with potentially useful information about student academic strengths and areas in need of improvement. Properly designed and scored performance assessments can also provide insight into student metacognitive skill development and how such skills contribute to the successful completion of the performance task. High quality performance assessments have the potential to contribute information for use in accountability systems as well as for classroom-level formative feedback to improve student learning.

Rating: Strong

C. System Utility: C1. Minimal Burden

Student test time varies substantially based on the type of performance assessment. For example, short-answer or essay exams generally take one or two class periods. On the other end of the spectrum, extended performance assessments may take several days, weeks, or even months, with students completing components over time or working on multiple drafts.

Estimating the burden on states and districts to incorporate large-scale performance assessments into an accountability system is more complex. In general, performance assessments place greater administrative burden on educators, are more costly to develop, and require more resources to score than multiple-choice assessments (Stecher, 2010). The benefits of performance tasks, however, may outweigh their costs. For instance, educators and administrators in Vermont and Kentucky perceived their portfolio assessment (created in the 1990s) as burdensome, but thought that the instructional benefits resulting from the program were worthwhile despite the burdens (Koretz, Barron, Mitchell, & Stecher, 1996; Koretz, Stecher, Klein, & McCaffrey, 1994). Advances in research and technology are reducing the costs of developing and administering performance assessments, making these assessments more feasible to implement on a large scale (Darling-Hammond & Adamson, 2010; Stecher, 2010).

Another way to look at the burden created by performance assessments is to consider the combined cost of current state and local ELA and mathematics assessments relative to their value. The combined cost, which on average is approximately $50 per pupil, includes test preparation, administration, scoring, and any professional or curriculum development associated with the

assessments (Darling-Hammond & Adamson, 2013). These tests do not gauge higher-order thinking skills very well. A RAND study concluded that only 2% of mathematics and 20% of ELA items on current assessments measure higher-order thinking (Yuan & Le, 2012). Darling-Hammond and Adamson (2013) argue that, at this level of spending, states could support the development of performance assessments that measured deeper learning more accurately and comprehensively.

Incorporating performance assessments into the CCI would add little additional burden for students, while it would create some initial burden for educators. The burden is counterbalanced to some degree by the ability of performance assessments to measure deeper learning and inform teaching. States may also want to take into account advances in technology and methodology that have steadily reduced the costs of developing, administering, and scoring performance assessments.

Rating: Moderate

C. System Utility: C2. Student Coverage

Copious evidence exists on the feasibility and challenges of using performance tasks in state accountability systems. It has been done, and it can be done universally. Doing so does require a different type of organization and commitment by the state to work with schools on the conditions of administration and the time necessary for task completion. Under such circumstances, universal student coverage is highly feasible, although accommodations and modifications of performance tasks for special needs student populations pose particular challenges. The language requirements of some tasks can also be challenging for English language learners. These challenges, though, are not outside the range of issues encountered when administering traditional content knowledge tests for accountability purposes, which also require time to be completed and adaptations for special populations.

Rating: Strong, with potential for universal coverage

C. System Utility: C3. Postsecondary Pathways

Performance assessments have the potential to provide useful information on student preparedness for college and career postsecondary pathways. Performance assessments can provide insight into student mastery of the cognitive and metacognitive skills that are essential for college success. Performance assessments can capture more complex constellations of skills, of the type required for workplace success. They can also gauge the degree of mastery of cognitive strategies such as problem formulation and interpretation that are necessary to complete assignments in many entry-level college courses. Familiarity with performance assessments will benefit students entering both college and career postsecondary pathways. The information generated from them when used for accountability purposes provides a unique insight into college and career preparedness, one that no content test can provide.

Rating: Strong

Summary

Table 4 presents the evaluative criteria ratings in relation to incorporating performance assessments in the CCI. Performance assessments contain trade-offs when balancing concerns about technical

quality, stakeholder relevance, and system utility. Research suggests that performance assessments can measure many of the more complex skills required in college and careers better than can multiple-choice standardized tests. Performance assessments have value for formative as well as summative assessment purposes because they can be used to diagnose student strengths and weaknesses and provide information used to improve instruction.

Table 4. Performance Assessment Evaluative Criteria Ratings

A. Technical quality			B. Stakeholder relevance				C. System utility		
A1	**A2**	**A3**	**B1**	**B2**	**B3**	**B4**	**C1**	**C2**	**C3**
Strong	Moderate	Moderate	Strong	Moderate	Strong	Strong	Moderate	Strong	Strong

Using performance tasks on a large scale for accountability purposes has proven very challenging in the past. Scoring issues, in particular, have deterred many states from incorporating performance tasks even though such tasks can better reflect what teachers actually teach in their classrooms. Other nations have found ways around the scoring dilemma that allow them to use the results from performance tasks for a range of high-stakes purposes. In these nations, the somewhat lower reliabilities that have been associated with performance assessment scoring when compared to standardized content knowledge tests are more than compensated for by the validity of what they test and their ability to signal to teachers and students what is important to learn. In this country, states such as Kentucky have achieved very high reliabilities with performance task scoring by investing in training and by leaving the system in place long enough for everyone to become familiar with it. This allows teachers and scorers alike to develop common mental models needed to operationalize the different levels of performance for different pieces of work. Common mental models provide a foundation for consistent judgments about work quality.

Performance task scoring can be made more manageable by a) specifying more clearly the content and skills to be tested and then writing tasks that more directly measure those elements, b) devoting sufficient time and resources to field-testing, c) training scorers to high levels of reliability initially (and adjusting scoring guides when needed to facilitate reliable scoring better), d) using techniques such as back reads and anchor papers to enhance scoring consistency, and then e) employing technology to spot scorer drift or outliers and retrain them in real time.

Teachers can be enlisted as scorers to a greater degree if performance assessments are incorporated into the classroom grading system, although safeguards against score inflation would be needed. As Darling-Hammond and Adamson (2013) note, current investments in state mathematics and ELA standardized tests could be redirected to performance assessment development. The economies of scale created by SBAC and PARCC could free up some resources, although states would still need to be willing to make investments of their own.

The greatest potential benefit of incorporating performance tasks into state accountability systems is to help counteract educator perceptions that what is measured in such systems does not reflect well what is happening in their classrooms. While performance tasks do not solve this problem entirely, they do indicate a willingness by the state to seek more valid information about student achievement in ways more directly connected to classroom learning. These assessments are more complicated to develop, administer, and score, which will always be a deterrent to their large-scale use. Targeted experiments on a local scale that demonstrate the best ways to use performance assessments for accountability purposes may be a logical next step in exploring their potential utility and value.

Innovative Measure 3: California State Seal of Biliteracy

Schools or school districts award the California State Seal of Biliteracy (CSB), a gold seal that appears on the transcripts or diplomas of students who have attained proficiency in two or more languages by high school graduation. In addition to the high school CSB, some school districts in California award pathway awards to graduating preschool, elementary, and middle school students who progress to biliteracy. To demonstrate proficiency and earn the CSB, students whose first language is English must do the following:

1. Complete all English language arts (ELA) requirements for graduation with an overall grade point average (GPA) of 2.0 or above,
2. Pass the Grade 11 California Standards Test [6] in ELA at or above the "proficient" level, and
3. Demonstrate proficiency in one or more languages other than English through one of the following:
    a) Score 3 (out of 5) or higher on an Advanced Placement (AP) exam with content in a language other than English
    b) Score 4 (out of 7) or higher on an International Baccalaureate (IB) exam with content in a language other than English
    c) Successfully complete a four-year high school course of study in a language other than English with a GPA of 3.0 or above in those courses
    d) Pass an approved school district language examination
    e) Score 600 or higher on a SAT II foreign language exam

Students whose first language is not English must achieve the "Early Advanced Proficiency" level on the California English Language Development Test (CELDT) and meet the requirements in steps 1, 2, and 3 above.

There are numerous reasons for providing incentives to schools that encourage student biliteracy. Biliteracy has been demonstrated to strengthen brain functioning (Adesope, Lavin, Thompson, & Ungerleider, 2010; Rodriguez, Carrasquillo, & Lee, 2014; Soveri, Laine, Hamalainen, & Hugdahl, 2011) and is associated with higher student performance on achievement tests (Armstrong & Rogers, 1997; Dumas, 1999). Beyond cognitive and academic benefits, biliteracy is increasingly important in a global economy, creating and enhancing career opportunities not available to those who know only one language. In 1980, 11% of the U.S. population spoke a language other than English at home. By 2009, that statistic increased to 20% (Ortman & Shin, 2011). This growth has been particularly acute in California, where nearly 44% of residents over the age of five speak a language other than English at home (Ryan, 2013).

In 2012, California became the first state to award a state Seal of Biliteracy, after passing legislation in 2011 (California AB 815, 2011). In 2012, California awarded more than 10,000 seals in 29 languages, including American Sign Language. The number of seals awarded in 2013 doubled to 21,655. More than 197 school districts and 19 charter schools awarded seals. Florida, Maryland, and Massachusetts have since added state Seals of Biliteracy, and pending legislation in New York, Texas, New Mexico, Illinois, and Washington indicates that other states are following California's lead. In 2009, Utah began a Kindergarten–Grade 3 dual-language immersion (DLI) program for

[6] The CSB criteria will need to be revised to reflect that the California Standards Test has been replaced by the Smarter Balanced Assessment System (SBAC).

1,400 students, implementing a 50/50 instructional split between English and Chinese, French, Portuguese, or Spanish (Hales, Dickson, & Roberts, 2013). In 2013, Utah's DLI program served more than 20,000 students. Oregon, Minnesota, and Delaware are working toward implementing language programs or providing incentives aimed at increasing biliteracy among students.

A. Technical Quality: A1. Relationship to Postsecondary Success

Due to the recent implementation of the CSB, there is no research directly measuring the long-term effects of the program on student college and career outcomes. However, biliteracy has been shown to improve cognitive skills, student achievement, and wage premiums. Acquiring a second language alters the density of the brain tissue responsible for information processes (Rodriguez et al., 2014). A recent meta-analysis of 63 studies involving 6,022 participants found that bilingualism associates reliably with increased attention control, working memory, metalinguistic awareness, and abstract and symbolic representation skills (Adesope et al., 2010). A study of Finnish-Swedish bilinguals confirmed that they can better direct attention and inhibit irrelevant stimuli (Soveri et al., 2011), an asset for college-bound students.

Learning a second language can contribute to academic progress in other subjects, including outperforming control groups on standardized tests (Armstrong & Rogers, 1997; Dumas, 1999). However, the vast majority of this research pertains to elementary and middle school students; little research has explored the effects on high school student achievement.

Mastering a second language may also produce career benefits. College graduates with fluency in a second language earn wages 2–3% higher than graduates knowing one language only. The returns differed by language. For instance, the return for speaking German is 4%; French, 2.7%; and Spanish, 1.7%. Individuals in the personal services, business support, management positions, and those who speak a language known by a smaller number of people have the highest returns (Saiz & Zoido, 2005). Furthermore, the Bureau of Labor Statistics (2014) estimates that the employment of interpreters and translators will grow by 46% from 2012–2022.

Despite these advantages, little to no research has directly linked the study of additional languages to improved college outcomes. However, EPIC examined admissions policies at higher education systems and found that the flagship universities in 36 of 50 states require a minimum of two years of courses in languages other than English for admission. More selective public schools publish increased requirements for foreign language coursework (e.g., three years or more for admission to the University of Wisconsin–Madison) or recommended coursework (e.g., the University of Michigan, the University of North Carolina at Chapel Hill, the University of Texas–Austin, and several campuses of the State University of New York recommend at least three years) to demonstrate the level of high-school academic rigor in foreign language necessary for admission.

Horn, Kojaku, and Carroll (2001) show that students who completed rigorous programs of study in high school, which included three years of languages other than English, were more likely to earn higher college GPAs and showed higher retention rates. However, the effect of language coursework was not differentiated, making it impossible to know the effect size or whether excluding languages other than English from a rigorous program would influence college outcomes. More research is needed to confirm the strength of relationships between taking a second language course and indicators of college and career success.

Although effects of bilingualism seem far reaching, the paucity of research exploring relationships between demonstrating additional language proficiencies in high school and future college success somewhat limits the value of the CSB as an indicator in the CCI. In terms of career-going pathways, Saiz and Zoido (2005) show that proficiency in a second language leads to higher wage premiums; however, this research does not say whether bilingualism relates positively to job performance. Subsequently, this white paper finds a weak relationship between the CSB and college and career success pending further empirical findings.

Rating: Weak

A. Technical Quality: A2. Fair Comparisons

One aspect of fairness is whether schools offer students similar opportunities to gain proficiency in a language other than English. Sung, Padilla, and Silva (2006) examined the language offerings at 220 public high schools in California in relation to the schools' API, socioeconomic status, percentage of students eligible for free and reduced-price lunch, and percentage of English learners. Schools with high percentages of economically disadvantaged students had a smaller percentage of students enrolled in classes in languages other than English, as well as fewer foreign language instructors, feeder middle-school programs, and opportunities and resources for those courses. California students in disadvantaged schools may not have the same opportunity to learn languages other than English as do counterparts in schools with high percentages of economically advantaged students.

Another aspect is whether the standardized pathways to the CSB (e.g., AP/SAT/IB foreign language exam scores) allow for fair comparisons among subgroups of students. Previous EPIC white papers identified the AP and SAT exams as allowing moderately fair comparisons; insufficient evidence exists for IB (EPIC, 2014a, 2014b). The California Standards Test, one of the assessments currently used in the API, has been shown as a fair measure of student performance. Approved school district language examinations, the SAT II, and coursework GPAs provide the most uncertainty for fair comparisons between schools and districts; without more detailed information regarding bias and the consistency of course/test quality and results between subgroups, it is an open question whether or not they are fair measures.

Students in schools with high percentages of economically disadvantaged students do not have the same access to coursework in languages other than English as do students in more economically advantaged schools. Because access to language instruction is not equal, the standardized pathways to earn the CSB are relatively weak in terms of fair comparisons.

Rating: Weak

A. Technical Quality: A3. Stability

Most of the approved pathways for students to earn the CSB are stable measures. For instance, previous EPIC white papers showed that the AP and IB exams are stable measures of student performance (EPIC, 2014b). SAT II exams in languages other than English were not studied directly in either white paper.[7] The California Standards Test has been shown to be a stable measure of

[7] Although the SAT II subject tests were not studied specifically in the EPIC research brief, the College Board ensures year-to-year comparability of test forms through ongoing equating studies.

student performance. Approved school district language examinations and coursework GPAs present many variables that threaten their stability as measures of the CSB.

The stability of the CSB is rated as moderate due to differences between pathways employing stable measures (i.e., AP and IB) rather than measures for which stability has not been validated. Approved school district language examinations and coursework GPAs have uncertain validity and may not generalize to the state level.

Rating: Moderate

B. Stakeholder Relevance: B1. Value to Students

The CSB, in and of itself, does not provide direct educational value to students. However, students in AP or IB pathways earning the CSB could receive college credits. Furthermore, taking the SAT II subject test is recommended for applicants to some UC campuses (University of California, 2010). AP/IB exams, SAT II subject tests, or the CSB itself could also serve as college application resume builders. Additionally, flagship universities in 42 states include demonstration of capacity in a language other than English as a requirement for admission. States such as Indiana, North Carolina, and Oregon have increased admissions requirements and/or recommendations for language experience since 1997, either for their flagship public universities or systemically. The CSB demonstrates moderate value for its indirect ability to promote students' skills and/or course-taking behaviors that align with perceptions of college preparedness.

Earning the CSB may improve a student's chances of getting a job by creating and enhancing career opportunities not available to those who know one language only. This is especially true for students seeking careers in the service industry, business, or other industries that require translators or interpreters. As the demographic research above shows, over 40% of Californians over the age of five speak a language other than English at home. The demand for employees who speak multiple languages will only increase in coming years.

Rating: Strong

B. Stakeholder Relevance: B2. Public Understanding

Because the CSB was introduced very recently, few Californians are likely to know its program specifics. Public awareness and understanding of the program will grow as the number of districts awarding the CSB increases. The concept of biliteracy is not complex, and the CDE and CSB websites provide a program overview. CaliforniansTogether, a statewide coalition of parents, teachers, and other stakeholders, provides additional resources and up-to-date news on the implementation of the CSB. Additionally, as state legislators such as Sen. Ricardo Lara (Los Angeles) and education officials including San Francisco superintendent Richard Carranza call for revisiting Proposition 227, a 1998 voter-approved law requiring non-English speaking students to be taught in English (Miranda, 2014), the topic of language instruction in California schools should intensify. While public understanding of the CSB may be weak currently, that is likely to change over time with more participation and potential revisits to language instruction policies in the state.

Rating: Moderate

B. Stakeholder Relevance: B3. Content, Skills, and Competencies

To earn the CSB, students must complete all ELA requirements for graduation with an overall GPA of 2.0 and pass the Grade 11 California Standards Test in ELA at the "proficient" level. Furthermore, whether it is an AP or IB course, or four years of instruction in a language other than English, students will be assessed on the content, skills, and competencies taught within their school. This holds true both for native English speakers and English learners.

Rating: Strong

B. Stakeholder Relevance: B4. Emphasis on Student Performance

All requirements necessary for earning the CSB, whether by course grade or exam score on the California Standards Test, AP, IB, or SAT II exams, are completed by individual students. These measures assess students directly, resulting in a strong emphasis on student performance. However, because of differential access to language programs based on school demographics, the CSB inherently includes an indirect measurement of a school's inputs and processes, which will influence student performance.

Rating: Moderate

C. System Utility: C1. Minimal Burden

There are no direct burdens to students for completing the first two requirements necessary to earn the CSB. Students completing a four-year course of study in a language other than English and earning a GPA of 3.0 or better to satisfy the third requirement also have no direct burdens beyond normal coursework. Students electing to take an AP, IB, or SAT II exam to satisfy the third requirement face a minimal test time burden, with all exams taking between 2–5 hours to complete. However, students taking the AP ($89), IB ($157 registration fee and $108 per exam), or SAT II ($48.50) will be required to pay a test fee. The College Board and the CDE provide test fee waivers for students who are eligible to receive free and reduced-price lunch.

The burden will be greater in districts that need to create new courses or programs to ensure that students have an equal opportunity to earn the CSB. All districts are required to submit an Insignia Request Form to the CDE listing the school name and number of students eligible to receive the CSB. Other than the time it takes to track students and submit the Insignia Request Form to the CDE, there are no other direct costs to districts. Indirect costs include time and cost required to create and manage a district plan for implementing a CSB program.

The direct costs to the CDE include managing the requests for Seals of Biliteracy and purchasing and sending to districts the CSB insignia that is affixed to a student's diploma or transcript.

As a conditional measure for schools that choose to offer the CSB, the inclusion of the measure would include minimum burden for students, educators, and the system. To the extent that schools were required to offer the CSB, the burden to schools and districts could rise depending on whether they currently offer universal access to language programs.

Rating: Strong

C. System Utility: C2. Student Coverage

In 2012, the first year of the CSB, more than 10,000 insignias were awarded to students in approximately 100 districts and 17 charter schools. More than 70% of students earned CSBs in Spanish, 10% in French, 7% in Mandarin, and 2% each in Cantonese, Japanese, and German (California Department of Education, 2013, 2014). In 2013, the number of CSBs awarded nearly doubled to 21,655 (CaliforniansTogether, 2014).

With approximately half a million high school graduates in California each year, the percentage of students earning the CSB is low; however, the rate of increase from 2012 to 2013 shows that the program has the potential to expand rapidly. The program will be able to expand only as fast as language offerings within schools grow. As a result, maximizing student coverage and minimizing the system burden are at odds. As more students demand languages other than English, schools will have to hire more qualified teachers to create opportunities for more students to become proficient.

Rating: Weak, potential for moderate coverage

C. System Utility: C3. Postsecondary Pathways

The CSB supports both the college and career-going postsecondary pathways. All students who earn the CSB will have a potential college application resume builder. Students who elect to take an AP, IB, or SAT II exam to demonstrate proficiency in a language other than English may earn college credit. Furthermore, students who earn the CSB will have satisfied admissions requirements for languages other than English at most colleges in the United States.

The importance of knowing more than one language is growing in an increasingly global economy and especially in California. Bilingual individuals have higher average lifetime earnings (Saiz & Zoido, 2005), although there is no evidence whether proficiency in multiple languages leads to better job performance.

Relative to some of the other potential measures that have been reviewed, the CSB has utility for both college and career pathways. The college and career benefits may be modest, in terms of college credits earned, or distant, in terms of future career earnings, but do exist for a variety of postsecondary pathways. The availability of more direct evidence between biliteracy and a variety of college and career outcomes would strengthen the ratings on this measure.

Rating: Moderate

Summary

The CSB is an attractive policy option for many reasons. The CSB is relevant to various stakeholders within and outside the educational system. Students have tangible educational and employment value by demonstrating proficiency in a language other than English. Additionally, the CSB will create few burdens for educators and the system as a whole.

Table 5. California State Seal of Biliteracy Evaluative Criteria Ratings

A. Technical quality			B. Stakeholder relevance				C. System utility		
A1	**A2**	**A3**	**B1**	**B2**	**B3**	**B4**	**C1**	**C2**	**C3**
Weak	Weak	Moderate	Strong	Moderate	Strong	Moderate	Strong	Weak	Moderate

Disadvantages relate to limited research available regarding the relationships between biliteracy and college and career success in general, and between CSB attainment and college and career success in particular. Also, schools with lower proportions of economically advantaged students tend to be less likely to offer a variety of language courses, creating an inequality of access issue that may be an important consideration in the development of the CCI.

The CSB is new and, therefore, there are many unknowns about its value as an indicator of school quality. Differential offerings across schools with differing demographic characteristics confound understanding of whether high CSB participation is a measure of school quality versus a measure of school resource adequacy. In addition, more time is needed to determine the relationships between CSB attainment and college and career success. As a conditional measure in the CCI, the CSB can serve the purpose of rewarding schools that are doing a good job of instructing students to proficiency in multiple languages.

Innovative Measures Conclusion

The classes of innovative measures reviewed in this white paper present questions about scalability. But innovations, by definition, must be discussed as trade-offs between feasibility and the power to drive improvement. Compared to current standardized tests, metacognitive and performance assessments both provide educators with actionable, immediate feedback, allowing classrooms to become more responsive to addressing student learning needs. As a result, these innovative assessments can foster paradigmatic shifts toward the powerful use of assessment data and a focus on deeper learning. The rapid growth in the number of students attaining the CSB suggests an enthusiasm in the state for recognizing the importance of biliteracy as a key achievement for a high school graduate. At least eight other states are implementing or holding public debates about similar programs. Innovative measures would fill a noticeable void in the state's accountability system by measuring 21st-century skills, including conscientiousness and self-efficacy, known to be determinants of college and career preparedness or future career success. Typically, standardized tests may punish students who lack metacognitive skills, but do not measure those skills directly. The lack of intentional focus on metacognitive skills limits a school's ability to comprehensively prepare students to pursue a variety of postsecondary pathways. Simply stated, these measures capture career and college readiness in a way that content-specific, multiple-choice examinations cannot.

These benefits should not mask the challenges of implementing innovative measures statewide. Limited empirical research, a common experience during periods of innovation, creates uncertainty about the selection of appropriate measures. For example, a robust menu of metacognitive assessments exist, but performance assessment is still maturing in its development. The CSB focuses on demonstrating proficiency in a language other than English, which is crucial, but only one of many 21st-century competencies. Scalability processes are needed that equalize access, a particular concern with the CSB, and do not create long-term, undue burdens. However, statewide decisions to include innovative measures within the CCI would likely lead to crowdsourcing, replication, and research that would drive down costs of time and resources while dramatically increasing public understanding and knowledge of best practices in the field. In particular, metacognitive skills already resonate with most stakeholders, but focusing assessment on the acquisition and development of those skills departs from the type of education most parents, educators, and policymakers received. Including innovative measures, particularly metacognitive assessment, in accountability systems would create bold, systemic change.

References

Abedi, J. (2010). *Performance assessments for English language learners*. Stanford, CA: Stanford Center for Opportunity Policy in Education.

ACT. (2014). ACT WorkKeys: Overview. Retrieved from http://www.act.org/products/workforce-act-workkeys/

Adesope, O. O., Lavin, T., Thompson, T., & Ungerleider, C. (2010). A systematic review and meta-analysis of the cognitive correlates of bilingualism. *Review of Educational Research, 80*(2), 207–245.

Aguirre-Munoz, Z., Boscardin, C. K., Jones, B., Park, J. E., Chinen, M., Shin, H. S., et al. (2006). *Consequences and validity of performance assessment for English language learners: Integrating academic language and ELL instructional needs into opportunity to learn measures* (CSE Report No. 678). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Almlund, M., Duckworth, A. L., Heckman, J. J., & Kautz, T. D. (2011). *Personality psychology and economics* (No. 16822). Cambridge, MA: National Bureau of Economic Research.

Armstrong, P. W., & Rogers, J. D. (1997). Basic skills revisited: The effects of foreign language instruction on reading, math, and language arts. *Learning Languages, 2*(3), 20–31.

Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). Personality and performance at the beginning of the new millennium: What do we know and where do we go next? *International Journal of Selection & Assessment, 9*(1–2), pp. 9–30.

Bureau of Labor Statistics. (2014). Occupational outlook handbook: Interpreters and translators. Retrieved from http://www.bls.gov/ooh/media-and-communication/interpreters-and-translators.htm

California Assembly Bill 815 (2011-2012), Chapter 618 (Cal. Stat. 2011).

California Department of Education. (2013). California Department of Education news release: State schools chief Tom Torlakson announces more than 10,000 students earn new state seal of biliteracy. Retrieved from http://www.cde.ca.gov/nr/ne/yr12/yr12rel68.asp

California Department of Education. (2014). Informational letter to the field. Retrieved from http://www.cde.ca.gov/sp/el/er/ssb14anlltrtofield.asp

CaliforniansTogether. (2014). The California campaign for biliteracy. Retrieved from http://www.californianstogether.org/

Conley, D. T. (2011). *Crosswalk analysis of deeper learning skills to Common Core State Standards*. Eugene, OR: Educational Policy Improvement Center.

Conley, D. T. (2013). *Getting ready for college, careers, and the Common Core: What every educator needs to know*. New York, NY: John Wiley & Sons.

Conley, D. T., & Darling-Hammond, L. (2013). *Creating systems of assessment for deeper learning*. Stanford, CA: Stanford Center for Opportunity Policy in Education.

Conley, D. T., Gilkey, L., Seburn, M., Bryck, R., & Shanley, C. (2012). *Non-cognitive assessments*. Eugene, OR: Educational Policy Improvement Center.

Darling-Hammond, L., & Adamson, F. (2010). *Beyond basic skills: The role of performance assessment in achieving 21st century standards of learning.* Stanford University, CA: Stanford Center for Opportunity Policy in Education.

Darling-Hammond, L., & Adamson, F. (2013). *Developing assessments of deeper learning: The costs and benefits of using tests that help students learn.* Stanford University, CA: Stanford Center for Opportunity Policy in Education.

Darling-Hammond, L., & Pecheone, R. (2009). Reframing accountability: Using performance assessments to focus learning on higher-order skills. In L. M. Pinkus (Ed.), *Meaningful measurement: The role of assessments in improving high school education in the twenty-first century* (pp. 25–53). Washington, DC: Alliance for Excellent Education.

Darling-Hammond, L., & Wentworth, L. (2010). *Benchmarking learning systems: Student performance assessment in international context*. Stanford University, CA: Stanford Center for Opportunity Policy in Education.

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology, 92*(6), 1087–1101.

Duckworth, A. L., & Quinn, P. D. (2009). Development and validation of the Short Grit Scale (GRIT–S). *Journal of Personality Assessment, 91*(2), 166–74.

Dumas, L. S. (1999). Learning a second language: Exposing your child to a new world of words boosts her brainpower, vocabulary, and self-esteem. *Child, 72*(74), 76–77.

Educational Policy Improvement Center. (2014a). *Measures for a college and career indicator: Research brief on SAT & ACT*. Eugene, OR: Author.

Educational Policy Improvement Center. (2014b). *Measures for a college and career indicator: Research brief on Advanced Placement and International Baccalaureate.* Eugene, OR: Author.

Ehley, L. (2006). *Digital portfolios: A study of undergraduate student and faculty use and perceptions of Alverno College's diagnostic digital portfolio* (Doctoral dissertation). Cardinal Stritch University, Milwaukee, WI.

Farrington, C. A., Roderick, M., Allensworth, E., Nagaoka, J., Keyes, T. S., Johnson, D. W., & Beechum, N. O. (2012). *Teaching adolescents to become learners: The role of noncognitive factors in shaping school performance.* Chicago, IL: University of Chicago Consortium on Chicago School Research.

Goldberg, L. R. (1992). The development of markers for the big-five factor structure. *Psychological Assessment, 4,* 26–42.

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist, 48*(1), p. 26–34.

Goldschmidt, P., Martinez, J. F., Niemi, D., & Baker, E. L. (2007). Relationships among measures as empirical evidence of validity: Incorporating multiple indicators of achievement and school context. *Educational Assessment, 12*(3–4), 239–266.

Griffin, R., MacKewn, A., Moser, E., & VanVuren, K. W. (2012). Do learning and study skills affect academic performance? An empirical investigation. *Contemporary Issues in Education Research, 5*(2), 109–116.

Hales, B., Dickson, S., & Roberts, G. (2013). *Critical languages: Dual language immersion education appropriations report.* Salt Lake City, UT: Utah State Office of Education.

Hart Research Associates. (2013). *It takes more than a major: Employer priorities for college learning and student success.* Washington, DC: Author.

Hojay, M., Erdmann, J. B., Veloski, J. J., Nasca, T. J., Callahan, C. A., Julian, E., & Peck, J. (2000). A validity study of the writing sample section of the Medical College Admission Test. *Academic Medicine, 75*(10), S25–S27.

Horn, L., Kojaku, L. K., & Carroll, C. D. (2001). *High school academic curriculum and the persistence path through college* (NCES 2001-163). Retrieved from http://nces.ed.gov/pubs2001/2001163.pdf

Huws, N., Reddy, P. A., & Talcott, J. B. (2009). The effects of faking on non-cognitive predictors of academic performance in University students. *Learning and Individual Differences, 19*(4), 476–480.

Kirst, M. W., & Mazzeo, C. (1996). *The rise, fall, and rise of state assessment in California, 1993-1996*. Stanford, CA: Policy Analysis for California Education.

Klein, S. (2006). *Assessing technical achievement in secondary career technical education.* St. Paul, MN: National Research Center for Career Technical Education.

Kobrin, J. L., Patterson, B. F., Barbuti, S. M., Mattern, K. D., & Shaw, E. J. (2008). *Validity of the SAT for predicting first-year college grade point average* (Research Report No. 2008-5). New York, NY: The College Board.

Koretz, D., Barron, S., Mitchell, M., & Stecher, B. (1996). *Perceived effects of the Kentucky Instructional Results Information System (KIRIS).* Santa Monica, CA: RAND Corporation.

Koretz, D., Stecher, B., Klein, S., & McCaffrey, D. (1994). The Vermont portfolio assessment program: Findings and implications. *Educational Measurement: Issues and Practice, 13*(3), 5–16.

Lane, S. (2010). *Performance assessment: The state of the art.* Stanford, CA: Stanford Center for Assessment, Learning and Equity.

Le, H., Casillas, A., Robbins, S., & Langley, R. (2005). Motivational and skills, social, and self-management predictors of college outcomes: Constructing the Student Readiness Inventory. *Educational and Psychological Measurement, 65*, 482–508.

Leonard, D. K., & Jiang, J. (1999). Gender bias and the college predictions of the SATs: A cry of despair. *Research in Higher Education, 40*(4), 375-407.

Lindqvist, E., & Vestman, R. (2011). The labor market returns to cognitive and noncognitive ability: Evidence from the Swedish enlistment. *American Economic Journal: Applied Economics, 3*(1), 101–128.

Lombardi, A, Seburn, M., & Conley, D. (2011). Development and initial validation of a measure of academic behaviors associated with college and career readiness. *Journal of Career Assessment 19*(4): 375–91.

McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology, 52*(1), 81–90.

McGaughy, C. (2014). *Understanding entry-level courses in American institutions of higher education* (Working Paper). Eugene, OR: Educational Policy Improvement Center.

Millennial Branding. (2012). *Millennial Branding student employment gap study*. Retrieved from http://millennialbranding.com/2012/05/millennial-branding-student-employment-gap-study/

Miranda, N. (2014, April 16). In California, a push to restore bilingual education. *NBC Bay Area KNTV*. Retrieved from http://www.nbcbayarea.com/news/local/In-California-A-Push-To-Restore-Bilingual-Education-255583271.html

Multon, K. D., Brown, S. D., & Lent, R. W. (1991). Relation of self-efficacy beliefs to academic outcomes: A meta-analytic investigation. *Journal of Counseling Psychology, 38*(1), 30–38.

Ortman, J. M., & Shin, H. B. (2011, August). *Language projections: 2010 to 2020*. Paper presented at the meeting of the American Sociological Association, Las Vegas, NV.

Pecheone, R., Kahl, S., Hamma, J., & Jaquith, A. (2010). *Through a looking glass: Lessons learned and future directions for performance assessment*. Stanford, CA: Stanford Center for Assessment, Learning and Equity.

Pellegrino, J., & Hilton, M. (Eds.). (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. Washington, DC: National Academy Press.

Peterson, C. H., Casillas, A., & Robbins, S. B. (2006). The Student Readiness Inventory and the Big Five: Examining social desirability and college academic performance. *Personality and Individual Differences, 41*, 663–673.

Peterson, N. G., Mumford, M. D., Borman, W. C., Jeanneret, P. R., Fleishman, E. A., & Levin, K. Y. (Eds.). (1997). *O*NET final technical report (Vols. 1–3)*. Salt Lake City: Utah Department of Employment Security, on behalf of the U.S. Department of Labor Employment and Training Administration.

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin, 135*(2), 322–338.

Robbins, S. B., 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. B., 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.

Rodriguez, D., Carrasquillo, A., & Lee, K. S. (2014). *The bilingual advantage: Promoting academic development, biliteracy, and native language in the classroom*. New York, NY: Teachers College Press.

Ryan, C. (2013). *Language use in the United States: 2011* (ACS-22). Retrieved from http://www.census.gov/prod/2013pubs/acs-22.pdf

Saiz, A., & Zoido, E. (2005). Listening to what the world says: Bilingualism and earnings in the United States. *Review of Economics and Statistics, 87*(3), 523–538.

Soland, J., Hamilton, L. S., & Stecher, B. M. (2013). *Measuring 21st-century competencies*. New York, NY: The Asia Society.

Soveri, A., Laine, M., Hamalainen, H., & Hugdahl, K. (2011). Bilingual advantage in attentional control: Evidence from the forced-attention dichotic listening paradigm. *Bilingualism: Language and Cognition, 14*(3), 371–378.

Stecher, B. (2010). *Performance assessment in an era of standards-based educational accountability*. Stanford, CA: Stanford Center for Opportunity Policy in Education.

Summit Public Schools. (2014). *Our approach: High impact teaching*. Retrieved from http://www.summitps.org/approach/high-impact-teaching

Sung, H., Padilla, A. M., & Silva, D. M. (2006). Foreign language education, academic performance, and socioeconomic status: A study of California schools. *Foreign Language Annals, 39*(1), 115–130.

Tomsho, R. (2009, August 19). Adding personality to the college admissions mix. *The Wall Street Journal*. Retrieved from http://online.wsj.com/news/articles/SB10001424052970203612504574342732853413584.

Tung, R. (2010). *Including performance assessments in accountability systems: A review of scale up efforts*. Boston, MA: Center for Collaborative Education.

University of California. (2010). SAT subject tests. Retrieved from http://admission.universityofcalifornia.edu/freshman/requirements/examination-requirement/SAT-subject-tests/

Walters, A. M., Kyllonen, P. C., & Plante, J. W. (2006). Developing a standardized letter of recommendation. *Journal of College Admission, 191*, 8–17.

Weinstein, C. E., & Palmer, D. (1990). *LASSI-HS user's manual*. Clearwater, FL: H&H Publishing.

Yuan, K., & Le, V. (2012). *Estimating the percentage of students who were tested on cognitively demanding items through the state achievement tests*. (RAND Working Paper No. 967). Santa Monica, CA: RAND Corporation.