Education Policy Analysis
Archives/Archivos Analíticos de Políticas
Educativas
ISSN: 1068-2341
epaa@alperin.ca
Arizona State University
Estados Unidos

Conley, David
A New Era for Educational Assessment
Education Policy Analysis Archives/Archivos Analíticos de Políticas Educativas, vol. 23, 2015, pp. 1-37
Arizona State University
Arizona, Estados Unidos

SPECIAL SERIES
A New Paradigm for Educational Accountability: Accountability for Meaningful Learning

education policy analysis archives

A peer-reviewed, independent, open access, multilingual journal

epaa aape

Arizona State University

Volume 23 Number 8 February 2nd, 2015 ISSN 1068-2341

A New Era for Educational Assessment[1]

David Conley
Educational Policy Improvement Center
United States

Citation: Conley, D. T. (2015). A new era for educational assessment. *Education Policy Analysis Archives, 23*(8). http://dx.doi.org/10.14507/epaa.v23.1983. This article is part of EPAA/AAPE's Special Series on *A New Paradigm for Educational Accountability: Accountability for Meaningful Learning.* Guest Series Edited by Dr. Linda Darling-Hammond.

Abstract: In this article, David Conley focuses on how to assess meaningful learning in ways that promote student achievement while simultaneously meeting system accountability needs. The article draws upon research that supports the notion that a major shift in educational assessment is needed in order to encourage and evaluate the kind of learning that enables success in college and careers. Over the next several years, almost every state will either implement the Common Core State Standards or develop an alternative version of their own. The question worth posing is whether educational stakeholders should be satisfied with on-demand tests that measure only a subset of the standards, or will they demand something more like a system of assessments in which multiple measures result in deeper insight into student mastery of complex and cognitive challenging standards? This article presents a vision for a new system of assessments, one designed to support the kinds of ambitious teaching and learning that most parents say they want for their children. The article begins with a brief historical overview, describes where educational assessment appears to be headed in the near term, and then discusses some longer-term possibilities, concluding with a series of recommendations for

[1] This article is adapted from: Conley, D. T. (2014). *A new era for educational assessment*. Boston, MA: Jobs for the Future.

Journal website: http://epaa.asu.edu/ojs/
Facebook: /EPAAA
Twitter: @epaa_aape

Manuscript received: 09/01/2014
Revisions received: 11/01/2014
Accepted: 01/05/2015

how policymakers and practitioners can move toward a better model of assessment for teaching and learning.
Keywords: assessment; accountability; meaningful learning; equity; Common Core Standards.

Una nueva era para la evaluación educativa
Resumen: En este artículo, David Conley se centra en la forma de evaluar el aprendizaje significativo de maneras que promuevan el logro de los estudiantes y al mismo tiempo satisfacer las necesidades de responsabilidad educativa del sistema. El artículo se basa en literatrura de investigación que apoya la idea de que es necesario un cambio importante en los modelos de evaluación de la educación con el fin de fomentar y evaluar el tipo de aprendizajes que permitan preparar a los estudiantes para los estudios universitarios y carreras profesionales. En los próximos años, casi todos los estados, implementará las Estándares Estatales Comunes o desarrollararn una versión alternativa por su propia cuenta. Una pregunta que vale la pena considerar es si agentes educativos tienen que estar satisfechos con examenees genéricos que miden sólo un subconjunto de estándares, o van a exigir algo más parecido a un sistema de evaluaciones con múltiples medidas para obtener una perspectiva más profunda de aprendizajes complejso y normas cognitivas desafiantes. Este artículo presenta una visión para un nuevo sistema de evaluación, uno diseñado para apoyar modelos de enseñanza y aprendizaje ambiciosos que la mayoría de los padres dicen que quieren para sus hijos. El artículo comienza con una breve reseña histórica, describe hacia donde la evaluación educativa parece dirigirse en el corto plazo y, a continuación se describen algunas posibilidades a largo plazo, para concluir con una serie de recomendaciones sobre como los que toman decisiones de políticas y los profesionales pueden avanzar hacia un modelo de evaluación de la enseñanza y el aprendizaje de mayor calidads.
Palabras clave: evaluación; responsabilidad educativa; aprendizaje significativ; Estándares Básicos Comunes.

Uma nova era para avaliação educacional
Resumo: Neste artigo, David Conley se concentra em como avaliar aprendizagens significativas de forma de promover o desempenho do aluno e atender as necessidades do sistema de responsabilização educacional. O artigo é baseado em literatrura de pesquisa que apóia a idéia dz necessidade de uma grande mudança nos modelos de avaliação da educação, a fim de promover e avaliar o tipo de aprendizagem que permite preparar os alunos para a universidade e carreiras professionais. Nos próximos anos, quase todos os estados vão a implementar os Estamdares Comuns e Esenciais do Estado ou desarrollararn uma versão alternativa por conta própria. A questão é saber se vale a pena considerar os educadores têm de estar satisfeitos com examenees genéricos que medem apenas um subconjunto de padrões, ou vai exigir algo mais parecido com um sistema de avaliação com medidas várias para obter uma perspectiva mais profunda de aprendizagenes complexoa e padrões cognitivas exigentes. Este artigo apresenta uma visão para um novo sistema de avaliação, destinado a apoiar modelos de ensino e aprendizagem ambiciosa o que a maioria dos pais dizem que querem para os seus filhos. O artigo começa com uma breve história descreve onde avaliação educacional parece caminhar no curto prazo e, em seguida, algumas possibilidades a longo prazo são descritos, concluindo com uma série de recomendações sobre como os tomadores de decisão e à política profissionais podem se mover em direção a um modelo de avaliação do ensino e aprendizagem mais calidads.
Palavras-chave: avaliação; responsabilidade educativa; significativ aprendizagem; Padrões Básicos Comuns.

Introduction

Imagine this scenario: You feel sick, and you're worried that it might be serious, so you go to the nearby health clinic. After looking over your chart, the doctor performs just two tests—measuring your blood pressure and taking your pulse—and then brings you back to the lobby. It turns out that at this clinic the policy is to check patients' vital signs and only their vital signs, prescribing all treatments based on this information alone. It would be prohibitively expensive, the doctor explains, to collect more information.

Most of us would find another health care provider.

Yet this is similar in some ways to how states gauge the knowledge, skills, and capabilities of students attending their public schools. State-mandated reading and math tests are often the only indicators of student achievement that "count." In many cases, these tests consist almost exclusively of multiple-choice items that measure only a subset of what is taught and learned. Nor do they get at students' ability to apply what they know in real-world settings.

Faced with tight budgets, policymakers have demanded that the costs associated with testing be minimized. Policymakers and, increasingly, educators, parents, and others in the broader community then use the quite limited information that these tests provide to make a wide range of decisions and draw major inferences, some appropriate and some not, about overall and individual student academic performance and progress as well as the efficacy of individual schools and the public school system as a whole.

One would have to travel back in time to an agrarian-era mindset of the 1800s to reach the conclusion that the only mission of school should be to get students to master the basics of reading and math. During the industrial age, the mission expanded to include other core subjects such as science, social studies, and foreign languages, along with exploratory electives and vocational education. And in today's postindustrial knowledge economy, an emerging consensus of business leaders, educational reformers, and policymakers has reached the conclusion that *all* young people need opportunities to develop the sorts of advanced content knowledge and problem-solving skills that used to be limited to an elite few (Conley, 2014b; Jobs for the Future, 2005; Secretary's Commission on Achieving Necessary Skills (SCANS), 1991). So why do schools continue to rely on assessments that do not go beyond a subset of information on two of the "Three R's"?[2]

That's a question that is being asked by a growing chorus of educators and, of late, parents, as well. Increasingly, they are voicing their dismay over current testing and accountability practices (Gewertz, 2013b, 2014; Sawchuk, 2014). Indeed, we may now be approaching an important crossroads in American education, as growing numbers of critics call for a fundamental change of course in the way learning is assessed and how schools are held accountable (Tucker, 2014).

The purpose of this article is to contribute to the conversation spurred by the accountability proposal put forth by Darling-Hammond, Wilhoit and Pittenger (2014) in their article, *Accountability for College and Career Readiness: Developing a New Paradigm,* also known as the 51st state model. In that article, Darling-Hammond and her colleagues propose a three-legged framework for strengthening accountability: meaningful learning, resource accountability, and professional capacity. I focus here on the meaningful learning component of the framework and how to assess meaningful learning in ways that promote student achievement while simultaneously meeting system accountability needs.

---

[2] It's always worth noting parenthetically that only one of the "Three R's" actually begins with the letter "r."

The article draws upon a research base developed by me, my colleagues, and others that suggests a major shift in educational assessment will be necessary in order to encourage and evaluate the kind of learning that is associated with success in college and careers. In particular, multiple studies that have analyzed syllabi, assignments, assessments, and student work from entry-level college courses and ascertained the perceptions of instructors of those courses provide a much more detailed picture of what college and career readiness actually entails—the knowledge, skills, and dispositions that can be assessed, taught, and learned that are strongly associated with success beyond high school (Achieve, The Education Trust, & Thomas B. Fordham Foundation, 2004; ACT, 2011; Conley, 2003; Conley, Aspengren, Gallagher, Stout, Veach, & Stutz, 2006; Conley & Brown, R., 2003; Educational Policy Improvement Center, 2014a; Seburn, Frain, & Conley, 2013; Texas Higher Education Coordinating Board & Educational Policy Improvement Center, 2009; The College Board, 2006).

Advances in cognitive science (Bransford, Brown, & Cocking, 2000; Pellegrino & HIlton, 2012), combined with the development and implementation of the Common Core State Standards and their attendant assessments (Conley, 2014a; Council of Chief State School Officers & National Governors Association, 2010a, 2010b), provide states with both incentive and opportunity to move toward the notion of a more comprehensive system of assessments in place of a limited set of often-overlapping and redundant tests of reading and math that measure lower-order, decontextualized content knowledge almost exclusively.

Over the next several years, almost every state will either implement the Common Core State Standards or develop an alternative version of their own. The question worth posing is whether educational stakeholders be satisfied with on-demand tests that measure only a subset of the standards, or will they demand something more like a system of assessments in which multiple measures result in deeper insight into student mastery of complex and cognitive challenging standards? Will schools begin to use measures of student learning that address more than just reading and math? Will policymakers demand evidence that students can apply knowledge in novel and non-routine ways, across multiple subject areas and in real-world contexts? Will educators be willing to collect information in other areas such as metacognitive learning skills such as persistence and information synthesis, which students must develop in order to become true lifelong learners? Will they be willing to integrate assessment more directly into curriculum and instruction? Will a new system of assessments get at *deeper* learning and address the whole constellation of knowledge and skills that young people need in order to be fully prepared for college, careers, and civic life?

This article presents a vision for a new system of assessments, one designed to support the kinds of ambitious teaching and learning that an increasing number of parents say they want for their children. It is important to focus on assessment in this country because testing has an outsized influence on curriculum and instruction here. The decentralized nature of educational governance in the U. S. has meant that assessments are often the only way to gauge educational quality, as opposed to many of the countries with which the U. S. competes that focus on the quality of curriculum and instructional practices, and then use multi-measure systems of assessment with a strong emphasis on formative measures, the results of which are used to improve education first and foremost, not to publicly shame schools and punish educators.

Thankfully, the public schools in the U. S. do not have to create an entirely new system of assessments from scratch—many schools already exhibit effective practices upon which others can build. For that to happen though, Americans must be willing to adopt new ways of thinking about the role of assessment in education, and policymakers (and assessment experts

and test developers) must be ready to move beyond unidimensional models that produce results cheaply and efficiently but that paint an incomplete picture of teaching and learning.

In order to help readers understand how we got to the current model of testing in the nation's schools, I begin with a brief historical overview. I then describe where educational assessment appears to be headed in the near term, and then discuss some longer-term possibilities, concluding with a series of recommendations for how policymakers and practitioners can move toward a better model of assessment for teaching and learning.

Historical Overview

Even though most states tend to favor a narrow set of assessment measures, relying primarily on multiple-choice tests, the decentralized nature of educational governance in the U.S. has led to the creation of a vast array of assessment methods and tools that can be used to gain insight into students' learning at different levels of cognitive complexity. Those methods run the gamut from individual classroom assignments and quizzes to capstone projects to innovative state tests to admissions methods to results from Advanced Placement® and International Baccalaureate® tests. Many measures are homegrown, reflecting the boundless creativity of American educators and researchers. Others are produced professionally and have long histories and a strong commercial commitment to their continuity. Some measures draw upon and incorporate ideas and techniques from other fields—such as business, sociology, psychology, the military, and other countries—where a wider range of methods have solid, long-term track records.

The problem is that not all, or even most, schools or states take advantage of this wealth of resources. By focusing on reading and math scores with the intent of using those scores to judge schools, federal and state policy over the past 15 or so years has, perhaps inadvertently, influenced schools to de-emphasize many of the assessment approaches that could be used to promote and measure more complex student-learning outcomes.

A Historical Tendency to Focus on Bits and Pieces

The current state of educational assessment has much to do with a longstanding preoccupation in the U.S. with *reliability* (the ability to measure the same thing consistently) over and above a concern with *validity* (the ability to measure the right things). To be sure, psychometricians—the designers of educational tests—have always considered validity to be critical, at least in theory (American Educational Research Association [AERA], American Psychological Association [APA], & National Council for Measurement in Education [NCME], 2014). In practice, though, they have had far more success in assuring the reliability of individual test forms than in dealing with messier and more complex questions about what *should* be tested, for what purposes, and with what consequences for the people involved.[3]

Over the past hundred years, this emphasis on reliability has led to the creation of tests made up of lots of discrete questions, each one pegged to a very particular skill or bit of knowledge. The more specific the skill, the easier it was to create test items that get at the same skill at the same level of difficulty from test to test, which translates to individual students and groups of students scoring remarkably consistently from test form to test form.

---

[3] The just-released version of the Standards for Educational and Psychological Testing take up the issue of validity in greater depth, but test-development practices for the most part have not yet changed dramatically to reflect a greater sensitivity to validity in the sense of what should be tested.

These types of items do not distinguish between a correct answer and the depth of student understanding or mastery of the content. For example, some content needs only be understood at a novice level, well enough to know what it is or what it is named. Other content needs to be understood more deeply, well enough to use and apply it in non-routine ways. Still other knowledge may be conceptually straightforward but may have to be applied with automaticity, in other words, without thinking twice about it, and with near-100% accuracy. And in many cases, items end up measuring how savvy students are as test takers, and whether they can discern the correct answer from among several choices without necessarily even knowing the content at all. Tests may be highly reliable and yet not get at how well the content is known and mastered, and at what level, let alone how this knowledge would be applied in context or in more holistic ways in new situations.

This focus on the parts and pieces has had a clear impact on instruction. In order to prepare students to do well on such tests, schools have treated literacy and numeracy as a collection of distinct, discrete components to be mastered independently, with little attention to students' ability to assemble those components into an integrated whole or to *apply* them in the context of the discipline, or, where appropriate, to other subject areas.

Further, if the fundamental premise of educational testing in the U.S. is that any type of knowledge can be disassembled into discrete pieces, measured as pieces, and that the results of the pieces can be summed to a whole that represents mastery of a knowledge domain, in other words, that testing students on just a somewhat random sample of a subject yields an adequate representation of the student's overall knowledge of the given subject.[4]

It's a bit like the old connect-the-dots puzzles, with each item on a test representing a dot. Connect enough items and you get the outline of a picture or, in this case, an outline of a student's knowledge that, via inference, can be generalized to untested areas of the domain to reveal the "whole picture."

This certainly makes sense in principle, and it lends itself to the creation of very efficient tests that purport to generate accurate data on student comprehension of the given subject. But what if these assumptions aren't true in a larger sense? What if understanding the parts and pieces is not the same as getting the big picture that tells whether students truly grasp concepts, can apply knowledge, and, perhaps most important, can transfer knowledge and skills from one context to an entirely new one? If it's not possible to do these critical things, then current tests will judge students to be well educated when in practice they may have gaping holes in their knowledge base or may not be able to use what they do know to solve problems in the subject area (what is known as "near transfer") or in other subject areas or situations outside the context where they learned the content (known as "far transfer"[5]).

Assessment Built on Intelligence Tests and Social Sorting Models

Another reason for this focus on measuring literacy and numeracy in a particularistic fashion has to do with the unique evolution of assessment in this country. Interestingly, a very

[4] This assumes high alignment between the test items and the content or standards being tested, an assumption that has not been met in many cases, as evidenced by alignment studies of state tests over the past decade.

[5] The concept of far transfer can be interpreted to mean that the learner is taught a skill in the context of a subject area and then expected to apply that skill in an entirely new context or setting. The term as used here is more restricted, referring to learning a skill in one subject area and then being able to apply it to other subject areas with fidelity, such as using algebra to solve science problems after being taught algebra in a math class, or applying principles of English grammar to writing outside of an English class.

different approach, what would now be called "performance assessment" (referring to activities that allow students to show what they can *do* with what they've learned) was common in schools throughout the early 1900s, although not in a form readily recognizable to today's educator. Recitations and written examinations (which were typically developed, administered, and scored locally) were the primary means for gauging student learning. In fact, the College Board (originally the College Entrance Examination Board) was formed in 1900 by colleges and high schools in partnership to standardize the multitude of written essay entrance examinations that had proliferated among the colleges of the day.

These essay exams were not considered sufficiently "scientific" at the time, an important criticism in an era when principles of science were being applied to the management of human organizations. Events in the field of psychological measurement from the 1900s to the 1920s exerted an outsized influence on educational assessment. The nascent research on intelligence testing gained favor rapidly in education at a time when the techniques of scientific management had near-universal acceptance as the best means to improve organizational functioning (Tyack, 1974; Tyack & Cuban, 1995). Further, tests administered to all World War I conscripts seemed to validate the notion that intelligence was distributed in the form of a normal curve (hence "*norm*-referenced testing") among the population: immigrants and people of color scored poorly, whites scored better, and upper income individuals scored the best, which seemed to confirm the social order of the day (Cherry, 2014).

At the same time, public education in the U.S. was experiencing a meteoric increase in student enrollment, along with rising expectations that students would stay in school beyond elementary and middle school. Confronted with the need to manage such rapid growth, schools applied the thinking of the day, which led them to use tests to categorize, group, and distribute students according to their presumed abilities (Tyack, 1974). Children of differing ability should surely be prepared for differing futures, the thinking went, and "scientific" tests could determine abilities and likely futures cheaply and accurately. All of this would be done in the best interest of children to help them avoid frustration and failure (Oakes, 1985).

Unfortunately, the available testing technologies were not then nor never have been sufficiently complex or nuanced enough to make these types of predictions very successfully, and so assessments have been used (or misused, really) throughout much of the past century to categorize students and assign them to different educational opportunities, or tracks, designed to lead to different economic and social futures ("Structural inequality in education," 2014).

Moreover, additional problems with such norm-referenced testing—designed to see how students stack up against each other—are readily apparent. In the first place, it's not clear how to interpret the results. By definition, some students will always come out on top and others will always rank at the bottom. But there's no reason to assume that the top-scorers have mastered the given material (since the tests always measure a degree of test savvy along with individual educational opportunity and may have ceiling effects if content is not sufficiently challenging). Nor can it be assumed that the low-scorers are in fact less capable (since, depending on where they happen to go to school or how they were taught, they may never have had a chance to study the given material at all). Even if they could be trusted to sort students into winners and losers, such tests would still fail to take into account the fact that if all students improved dramatically, those at the bottom would still be deemed to be failures. Neither do norm-referenced tests generally provide much actionable information to help teachers and students know what need to be done to change the teaching and learning process to improve scores.

Assessment to Guide Improvement

Since the late 20th century, the use of intelligence tests and academic exams to sort students into tracks has been largely discredited (Goodlad & Oakes, 1988; Oakes, 1985). In today's rapidly changing economy and society, when everyone needs to be capable of learning throughout their careers and lives, it would be especially counterproductive to keep sorting students in this way—far better to try to educate all children to a high level than to label some as losers and anoint others as winners as early as possible.

The first limited manifestation of an alternative approach to sorting based on norm-referenced scores was the mastery learning movement of the late 1970s (Block, 1971; Bloom, 1971; Guskey, 1980a, 1980b, 1980c). Consistent with the prevailing tendency for assessments to measure parts and pieces, mastery learning focused entirely on basic skills in reading and math and reduced those skills down to the smallest testable units possible. Mastery learning did, however, represent a real departure from the status quo, since it argued that essentially all students could master a specified body of knowledge if they received high quality instruction and opportunities to practice the relevant content. The purpose of assessment was not to put students into categories but, instead, to generate information about their performance, in order to help them master specified content.

One of the problems with mastery learning, though, was that it was limited to content that could be broken up into dozens of distinct subcomponents that could be tested in detail (Horton, 1979). As a result, educators were quickly overwhelmed trying to keep track of student progress on all the elements. Equally vexing was the fact that mastering those elements didn't necessarily lead to proficiency in the larger subject area, or the ability to transfer what had been learned to new contexts (Horton, 1979). Students could pass the reading tests only to run into trouble when they encountered new and different kinds of written material, and they could ace the math tests only to be stumped by unfamiliar applications of the same content. In other words, they had very limited ability to transfer what they had learned. To critics of mastery learning, the approach highlighted the limitations of shallow-learning models (Slavin, 1987), a problem that "criterion-referenced" testing was designed to address.

Whereas norm-referenced tests aim to show how students stack up against each other, criterion-based assessments are meant to determine where students stand in relation to a specific standard ("Criterion-referenced test," 2014). While mastery learning uses tests to help students master discrete bits of content, criterion-based assessments measure student performance in relation to specific learning targets and standards of performance. Like mastery learning, the goal isn't to identify winners and losers but, rather, to show where students stand in relation to a learning target. "Criterion" as used in this context refers to the specific criteria the learner needs to meet to be deemed to know the desired content and skills. This approach to testing supports an instructional model designed to get as many students as possible to learn the given knowledge and skills.

Early Statewide Performance Assessment Systems

The first wave of more complex standards necessary to support criterion-based assessment emerged in the late 1980s and early 1990s (Brandt, 1992/1993). Initially referred to as outcomes-based education, they borrowed from mastery learning in the sense that students were supposed to master them. However, these standards were more expansive and multidimensional, designed to produce a well-educated, well-rounded student, not just one who could demonstrate discrete literacy and numeracy skills. Thus, for example, they included not

just academic content knowledge but also outcomes that related to thinking, creativity, problem solving, and the interpretation of information (Spady, 1992).

These more complex standards created a demand for assessments that went well beyond measuring bits and pieces of information. Thus, the early 1990s saw the bloom of statewide performance assessment systems that sought to gauge student learning in a much more ambitious and integrated fashion. In those years, states such as Vermont and Kentucky required students to collect their best work in "portfolios," which they could use to demonstrate their full range of knowledge and skills. Maryland introduced performance assessments (Hambleton, Impara, Mehrens, & Plake, 2000), California implemented its California Learning Assessment System (CLAS), and Oregon created an elaborate system that included classroom-based performance tasks, along with certificates of mastery at the ends of grades 10 and 12, requiring what amounted to portfolio evidence that students had mastered a set of content standards (Rothman, 1995).[6]

These assessments represented a radical departure from previous norm-referenced achievement tests and criterion-referenced mastery learning models. They were also quite difficult to manage and score initially—requiring more classroom time to administer, more training for teachers, and more support by state education agencies—and they quickly encountered a range of technical, operational, and political obstacles.

Vermont, for example, ran into problems establishing reliability quickly (Koretz, Stecher, & Deibert, 1993), the holy grail of U.S. psychometrics, as teachers were slow to reach a high level of consistency in their ratings of student portfolios (although their reliability did improve as the design of the portfolios was refined and teachers became more familiar with the scoring process). In California, parents raised concerns that students were being asked inappropriately personal essay questions (Dudley, 1997; Kirst & Mazzeo, 1996). Attempts to conduct authentic assessments ran into logistical problems, such as the time the fruit flies shipped to schools for a science experiment died en route, jeopardizing a statewide science assessment. In Oregon, some assessment tasks turned out to be too hard, and others too easy. And everywhere, students who had excelled at taking the old tests struggled with the new assessments, leading to a backlash among angry parents of high achievers.

In the process, a great deal was learned about the do's and don'ts of large-scale performance assessment (Pecheone and Kahl, 2014). These initial challenges took a political toll on the new assessments, and support for them weakened in many states. At the same time, standards in a number of states were revised to be more specific and detailed, resulting in an increased emphasis on testing students on individual bits and pieces of academic content, particularly in reading and mathematics. And while several states continued their performance assessments systems throughout the decade, most of these systems came under increasing scrutiny due to their costs, the challenges involved in scoring them, the amount of time it took to administer them, and the difficulties involved in learning to teach to them.

The federal No Child Left Behind (NCLB) legislation passed in 2001, which mandated testing in reading and mathematics in grades 3-8 and once in high school, created pressures on states to use assessments that passed federal muster. The technical requirements of NCLB (as interpreted in 2002 by Department of Education staff) could conceivably be met with a variety of types of assessments (and states such as Massachusetts and the New England Common

---

[6] For a review of these state efforts, see Pecheone, R. and S. Kahl (2014). Where we are now: Lessons learned and emerging directions, in Darling-Hammond and Adamson, *Beyond the bubble test: How performance assessments support 21st century learning*. San Francisco, CA: Jossey-Bass Wiley.

Assessment Program did continue to include in their state tests a range of open-ended response item types). However, in practice, it turned out to be more economical and easier for states to adopt standardized tests consisting exclusively of selected-response (i.e., multiple-choice) items (Linn, Baker, & Betenbenner, 2002; U. S. Department of Education, 2001).[7]

The designers of NCLB were not necessarily philosophically opposed to performance assessment. First and foremost, though, they were intent on using achievement tests to hold educators accountable for how well they educated all student populations (Linn, 2005; Mintrop & Sunderman, 2009). Thus, although the law wasn't specifically designed to eliminate or restrict performance assessment, this was one of its consequences. A few states (most notably Maryland, Kentucky, Connecticut, and New York) were able to hold onto the performance elements of their testing systems throughout the decade. Most states, though, retreated from almost all forms of assessment other than multiple-choice items and short essays, often as much due to the cost of performance assessment as to federal requirements.

Fast forward to 2014, however, and things may be poised to change once more. As I will discuss in the next section, this trend may now be on the verge of changing direction for a variety of reasons, not the least of which is an expanding interpretation of how NCLB assessment requirements can be met.

Why It's Time for Assessment to Change

An important force to consider when viewing the current landscape of assessment in U.S. schools is the rising weariness with test-based accountability systems of the type that NCLB has mandated in every state. Although the expectations contained in NCLB were both laudable and crystal clear—that all students become competent readers and capable quantitative thinkers—the means by which these qualities were to be judged led to an over-emphasis on test scores derived from assessments that inadvertently devalued conceptual understanding and deeper learning. Even though student test scores improved in some areas, educators were not convinced they were associated with real improvements in learning (Jennings & Rentner, 2006). A desire to increase test scores led many schools to a race to the bottom in terms of the instructional strategies employed, which included an outsized emphasis on test-preparation techniques and a narrowing of the curriculum to focus, sometimes exclusively, on those standards that were tested on state assessments (Cawelti, 2006). One side effect of these strategies has been a decrease in U. S. scores on the Programme for International Student Assessment (PISA), a test that emphasizes knowledge application in and transfer to new and novel settings (Darling-Hammond, Wilhoi,t and Pittenger,2014).

But in addition to the public and educators tiring of NCLB-style tests (as well as the U.S. Department of Education's apparent willingness to allow states to experiment with new models), at least two other important reasons help explain why the time may be ripe for a major shift in educational assessment:

First, the results from recent research clarifies what it means to be college and career ready. These findings make it increasingly difficult to defend the argument that multiple-choice tests are valid measures of skills student need to be prepared for postsecondary success.

---

[7] To be entirely clear on this point, the law did not require multiple-choice tests, but states often found it difficult to get approval for other types of assessments from the U. S. Department of Education, which would not allow federal funds designated to support NCLB tests to be used to pay for performance assessments.

Second, recent advances in cognitive science have yielded new insights into how humans organize and use information, which makes it equally difficult to defend tests that treat knowledge and skill as nothing more than a collection of discrete bits and pieces.

What Does It Mean To Be College And Career Ready?

The term college and career ready itself is relatively recent. Up until the mid-2000s and even now in many places, a secondary school education was geared toward making at least some students *eligible* to attend college, but not necessarily to make them ready to succeed.

For students hoping to attend a selective college, eligibility was achieved by taking required courses, getting sufficient grades and admission test scores, and perhaps garnering a positive letter of recommendation and participating in community activities. And for most open-enrollment institutions, it was sufficient simply for applicants to have earned a high school diploma, then apply, enroll, and pay tuition. Whether students could succeed once admitted was largely beside the point. Access was paramount.

The new economy has changed all of that. A little college, while better than none, is nowhere near as useful as is a certificate or degree. Being admitted to college doesn't mean much if the student is not prepared to complete a program of study. Further enhancing the value of readiness and the need for students to succeed is the crushing debt load ever more students are incurring to attend college now. A college education essentially has to improve a student's future economic prospects, if for no other reason than to enable debt repayment.

Why have high school educators been focused on students' eligibility for college and not on their readiness to succeed there? A key reason is that they weren't entirely sure what college readiness entailed. Until the 2000s, essentially all the research in this area used statistical techniques that involved collecting data on factors such as high school grade-point average, admission tests, and the titles of high school course taken, and then trying to determine how those factors related to first-year college course grades or retention in college beyond the first term.[8] These results were useful in some ways, identifying certain high school experiences and achievements that correlated to some measures of college success. Leaving aside the methodological limitations of these methods, this line of research wasn't able to zero in on what, specifically, enabled some students to succeed while others struggled nor, for the most part, the actions students should take to become more likely to be ready for college, other than take challenging courses.

In recent years, however, researchers have been able to identify a series of very specific factors that, in combination, maximize the likelihood that students will make a successful transition to college and perform well in entry-level courses at any of a wide range of postsecondary institutions. In comparison to what was known just fifteen years ago, we now have a much more comprehensive, multi-faceted, and rich portrait of what constitutes a college-ready student.

This includes numerous studies, including many conducted by me and my colleagues, identify the demands, expectations, and requirements that students tend to encounter in entry-level college courses (Brown, 2007; Conley, 2003; Conley, 2011, 2014b; Conley, Aspengren, & Stout, 2006; Conley, Aspengren, Gallagher, Nies, 2006a, 2006b; Conley, Drummond, DeGonzalez, Rooseboom, & Stout, 2011; Conley, McGaughy, Brown, van der Valk, & Young, 2009; Conley, McGaughy, Brown, van der valk, & Young, 2009; Conley et al., 2008; Conley, McGaughy, Cadigan, Forbes, & Young, 2009; Educational Policy Improvement Center, 2014a; Seburn et al., 2013; Texas Higher Education Coordinating Board & Educational Policy

[8] These methods are still widely used, particularly by colleges themselves.

Improvement Center, 2009). These studies have analyzed course content including syllabi, texts, assignments, and instructional methods and have also gathered information from instructors of entry-level courses to determine the knowledge and skills students need to succeed in their courses.

This body of research has reached remarkably consistent conclusions about what it means to be ready to succeed in a wide range of postsecondary environments. And the key finding is one that has far-reaching implications for assessment at the high school level: In order to be prepared to succeed in college, students need much more than content knowledge and foundational skills in reading and mathematics.

On its face, this may not seem all that surprising. Yet, the prevailing methods of college admission in this country, and much research on college success, largely ignore just how critical it is for aspiring college students to develop a wide range of cognitive strategies, learning skills, knowledge about the transition to higher education, and other aspects of readiness.

For clarity's sake, I have organized these factors into a set of four "keys" to college and career readiness. Before introducing this model, though, it's worth noting that other researchers have offered conceptual models of their own, choosing to arrange these factors into other categories, using different terminology than I present here.[9] Ultimately, though, the specific model is not the critical issue. On the most important points—having to do with the range of factors that contribute to college readiness—researchers have reached a strong consensus. Different models represent different ways of carving up the pie, but the substance is the same.

That said, the Four Keys model derives from research on literally tens of thousands of college courses at a wide range of postsecondary institutions. It highlights four main factors that contribute to college readiness:

- **Key Cognitive Strategies.** The thinking skills students need to learn material at a deeper level and to make connections among subjects.
- **Key Content Knowledge**. The big ideas and organizing concepts of the academic disciplines that help organize all the detailed information and nomenclature that constitute the subject area along with the attitudes students have toward learning content in each subject area.
- **Key Learning Skills and Techniques**. The student ownership of learning that connects motivation, goal setting, self-regulation, metacognition, and persistence combined with specific techniques such as study skills, note taking, and technology capabilities.
- **Key Transition Knowledge and Skills**. The aspiration to attend college, the ability to choose the right college and to apply and secure necessary resources, an understanding of the expectations and norms of postsecondary education, and the capacity to advocate for one's self in a complex institutional context.

In turn, each of these Keys has a number of components, all of which are actionable by students and teachers—in other words, these are things that can be assessed, taught, and learned successfully. (On that score, note that the model does not include certain factors, such as parental income and education level, that are strongly associated statistically with college success but which *are not actionable* by schools, teachers, or students. The point here is to highlight things that can be done to prepare students to succeed, not to list the things that cannot be changed.)

---

[9] See, for example: Farrington, C. A. (2013). Academic mindsets as a critical component of deeper learning. University of Chicago, Consortium on Chicago School Research; Partnership for 21st Century Skills. (2009). "Framework for 21st Century Learning"; Conley, D. T. (2011). Crosswalk analysis of Deeper Learning skills to Common Core State Standards. Palo Alto, CA: Hewlett Foundation.

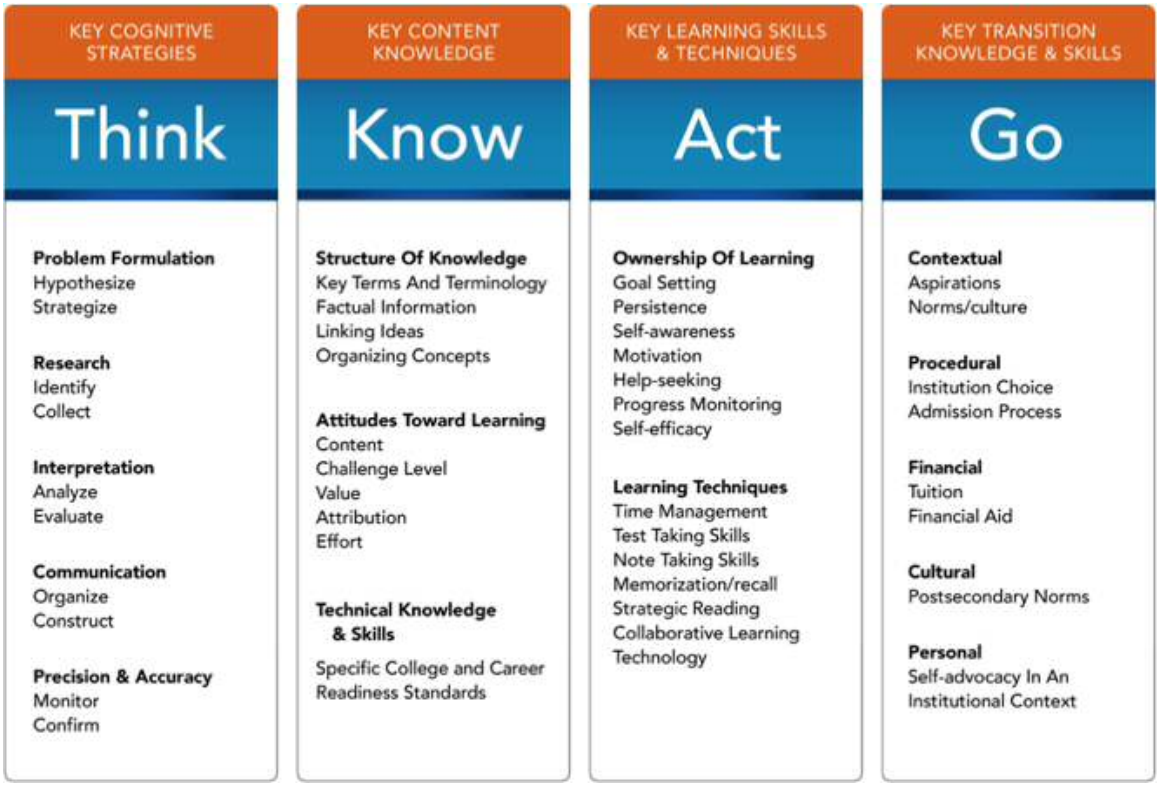

Figure 1. Four Keys Model

Advances In Brain And Cognitive Science

Recent research in brain and cognitive science provides a second major impetus for shifting the nation's schools away from a single-minded focus on current testing models and toward performance assessments that measure and encourage deeper learning.

Of particular importance is recent research into the malleability of the human brain (Hinton, Fischer, & Glennon, 2012), which has provided strong evidence that individuals are capable of improving many skills and capacities that were previously thought to be fixed. Intelligence was long assumed to be a single, unchanging attribute, one that can be measured by a single test. However, that view has come to be replaced by the understanding that intellectual capacities are varied, multi-dimensional, and can be developed over time, if stimulated to do so.

One critical finding is that students' attitudes toward learning academic material turns out to be at least as important as their aptitude (Dweck, Walton, & Cohen, 2011). For generations, test designers have claimed to be able to identify students' "true" ability levels, in order to steer them into academic and career pathways that match their natural talents and capabilities. But the reality appears to be that, far from helping students find their place, such test results serve to discourage many students from making the sorts of sustained, productive efforts to learn that would allow them to succeed in a more challenging course of study.

Recent research also challenges the commonly held belief that the human brain is organized like a library, with discrete bits of information grouped by topic in a neat and orderly fashion, to be recalled on demand (Donovan, Bransford, & Pellegrino, 1999; Pellegrino & HIlton, 2012). In fact, evidence reveals that the brain is quite sensitive to the *importance* of information, and it processes sensory input largely by determining its relevance (Medina, 2008). Thus, the longstanding American preoccupation with breaking subject-area knowledge down

into small bits, testing students' mastery of each one, and then teaching those bits sequentially, may in fact be counter-productive. Rather than ensuring that students learn systematically, piece by piece, this approach could easily deny them critical opportunities to get the big picture and to figure out which information and concepts are most important.

When confronted by a torrent of bits and pieces presented one after the other, without a chance to form strong links among them, the brain tends to forget some, connect others in unintended ways, experience gaps in sequencing, and miss whatever larger purpose and meaning might have been intended. Likewise, when tests are designed to measure students' mastery of discrete bits, they provide few useful insights into students' conceptual understanding or their knowledge of how any particular piece of information relates to the larger whole.

The net result is that students struggle to retain information (National Research Council, 2002). Having received few cues about the relative importance of the given content, and having few opportunities to fit it into a larger framework, it's no wonder that they often forget much of what they have learned, from one year to the next, or that even though they can answer detailed questions about a topic, they struggle to demonstrate understanding of the larger relevance or meaning of the material. Indeed, this is one possible explanation for why scores at the high school level on tests such as the National Assessment of Educational Progress (NAEP)—which gets at students' conceptual understanding, along with their content knowledge—have flat-lined over the past two decades, a period when the emphasis on basic skills increased dramatically.

Ideally, secondary-level instruction guides students through learning progressions that build in complexity over time, moving toward larger and more integrated structures of knowledge. Rather than being taught skills and facts in isolation, high school students should be deepening their mastery of key concepts and skills they were taught in earlier grades, learning to apply and extend that foundational knowledge to new topics, subjects, problems, tasks, and challenges.

In order to provide this sort of instruction, teachers need access to tests and tools that allow them to assess far more than just the ability to recall bits and pieces of content. What's needed, rather, are opportunities for students to demonstrate their conceptual understanding, to relate smaller ideas to bigger ones, and to show that they grasp the overall significance of what they have learned.

Moving Toward a Broader Range of Assessments

Assessments can be described as falling along a continuum, ranging from those that measure bits and pieces of student content-knowledge to those that seek to capture student understanding in more integrated and holistic ways (as shown in Figure 2, drawn from Conley & Darling-Hammond, 2014). But it's not necessary or even desirable to choose just one approach from the continuum and reject the others. As I describe in the following pages, a number of states are now creating school assessment models that combine elements from multiple approaches, which promises to give them a much more detailed and useful picture of student learning than if they insisted on a single approach.

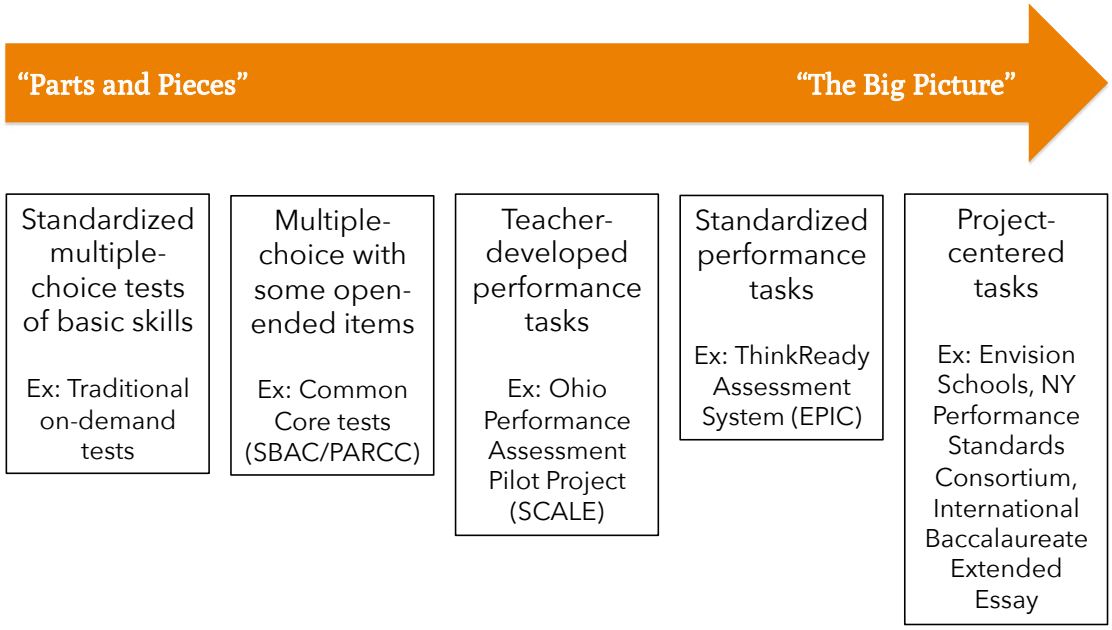

Figure 2. Continuum of Assessments

Traditional Multiple-Choice Tests

Traditional multiple-choice tests have come under a great deal of criticism in recent years, but whatever their flaws, they are a mature technology that offers some distinct advantages. They tend to be reliable, as noted. Also, in comparison to some other forms of assessments, they do not require a lot of time or cost a lot of money to administer, and they generate scores that are familiar to educators. Thus, it's not surprising that a number of states, when given the option of using the more complex tests of the Common Core developed by the two state consortia–Partnership for the Assessment of College and Career Readiness (PARCC) or Smarter Balanced Assessment Consortium (SBAC)–have instead chosen to reinstitute multiple-choice tests with which they are already familiar (Gewertz, 2013a; Ujifusa, 2014). It's likely that multiple-choice tests will continue to be widely used for some time to come as evidenced by the fact that the Common Core assessments continue to include many items of this type in addition to some new item types.

One recent advancement in this area is the design and use of computer-adaptive tests, which add a great deal of efficiency to the testing process. Depending on the student's responses, the software will automatically adjust the level of difficulty of the questions it poses (after a number of correct answers, it will move on to harder items; too many incorrect responses, and it will move back to easier ones), zeroing in more quickly and precisely on student's level of mastery of the given material. Further, the technology makes it a simple matter to include items that test content from previous and subsequent grades, which allows

measurement of a very wide distribution of knowledge and skills (from below grade level to far above it) that might exist in any given class or testing group.[10]

Common Core Tests

The two consortia of states that are developing tests of the Common Core State Standards, and both of them—the Partnership for the Assessment of Readiness for College and Careers (PARCC) and the Smarter Balanced Assessment Consortium (SBAC)—have touted the potential of these tests to overcome many of the shortcomings of NCLB-inspired testing.

These exams will test a range of Common Core standards at grades 3-8 and once in high school, using a mix of methods, including performance tasks that get at more complex learning. However, the tests still rely predominantly on items that gauge student understanding of discrete knowledge and, hence do not address a number of key Common Core standards that require more extensive cognitive processing and deeper learning.

This is a critical point, and it bears repeating: While the PARCC and SBAC assessments have been designed specifically to measure student progress on the Common Core standards, in point of fact they do this only within the student's grade level and assess only *some* of those standards.

Many of the skills that the Common Core defines as necessary preparation for college and careers are ones that can only be tested validly through a wider range of methods than either PARCC or SBAC currently employs. For example, the standards specify that by the time students graduate from high school, they should be able to:

- Conduct research and synthesize information
- Develop and evaluate claims
- Conduct extended investigations
- Use technologies to present information in multiple forms
- Plan, evaluate, and refine solution strategies
- Design and use mathematical models, and
- Collaborate to solve problems

In short, many of the standards contained in the Common Core call upon students to demonstrate quite sophisticated knowledge and skills, requiring more complex forms of assessment than PARCC and SBAC can reasonably be expected to provide from a test that will be administered over several hours on a computer. The tests do include sophisticated performance tasks that ask students to spend time preparing a more in-depth response. However, the tasks are limited in number and in the amount of time that can be devoted to them. This means that core skills such as editing and redrafting can only be assessed in a limited way through the format of an on-demand test.

None of this critique is intended to denigrate those assessments but, rather, to argue that they are not, in and of themselves, sufficient to meet the Common Core's requirements. If states mean to take the standards seriously, then they will want to consider a much broader continuum of options for measuring them, including assessments that are now being developed and used locally, in networks, and, in some cases, by states on a limited basis. Such assessments—including performance tasks, student projects, and collections of evidence of student learning—are both feasible and valid, but they also present challenges of their own.

---

[10] It is worth noting that this technology will not be fully used in the SBAC tests because USDOE is interpreting NCLB to require that all items measure only standards taught at the grade level at which the test is being administered.

Performance Tasks

Performance tasks have been a part of state-level and school-level assessment for decades. They encompass a wide range of formats, including tasks that can take anywhere from 20 minutes to two weeks to complete, and that require students to generate anything from a two-paragraph passage to a whole collection of original work products. Generally speaking, though, most performance tasks consist of activities that can be completed in a few class periods at most, and which do not require students to conduct extensive independent research.

A number of prominent examples of assessments that truly assess deeper learning deserve mention:

In 1997, the New York Performance Standards Consortium, a group of New York schools with a history of using performance tasks as a central element of their school-based assessment programs, sued the State of New York, successfully, to allow the use of performance tasks to meet state testing requirements (Knecht, 2007). Most notable among these schools was Central Park East Secondary School, which had a long and distinguished history of having students present their work to panels consisting of fellow students, teachers, and community members with expertise in the subject matter being presented. Most of these schools were also members of the Coalition of Essential Schools, which also advocated for these types of assessment at its over 600 member schools. The Consortium continues to include several dozen schools that use the same framework, rubrics, and scoring system for these tasks that substitute for the Regents exams. Graduates from Consortium high schools are admitted to and succeed in college at far higher rates than other students in the state, including those who are on average more socioeconomically advantaged ("Educating for the 21st century: Data report on the New York Performance Standards Consortium," n/d).

More recently, my colleagues and I at the Educational Policy Improvement Center (EPIC) developed ThinkReady, an assessment of Key Cognitive Strategies (Baldwin, 2011; Conley, 2007; Conley, McGaughy, O'Shaughnessy, & Rivinus, 2007). Its performance tasks—which take anywhere from a few class periods to several weeks (with out-of-class work) to complete—require students to demonstrate skills in problem formulation, research, interpretation, communication, and the making of precise and accurate claims. Teachers use a common scoring guide that tells them where students stand on a progression from novice to emerging expert on the kind of thinking associated with college readiness. The system spans grades 6-12 and is organized around four benchmark levels that correspond with cognitive skill development rather than grade level.

The Ohio Performance Assessment Pilot Project (OPAPP) was conceived of as a pilot project to identify how performance-based assessment could be used in Ohio (Ohio Department of Education, n.d.). With the support of the Stanford Center for Assessment, Learning, and Equity (SCALE), teachers developed tasks at grades 3-5 and 9-12 in English, mathematics, science, social studies, and career-technical pathways. The tasks were field tested and piloted and then refined. Tasks were scored online and at in-person scoring sessions with high levels of reliability. Teachers described how using the tasks leveraged deeper learning in their classrooms (Wei, Schultz, & Pecheone, 2012).

New Hampshire is in the process of developing common statewide performance tasks that will be included within a comprehensive state assessment system along with Smarter Balanced Assessment Consortium (SBAC) assessments (New Hampshire Department of Education, 2014). Each performance task will be a complex curriculum-embedded assignment involving multiple steps that require students to use metacognitive learning skills. As a result

student performance will reflect the depth of what students have learned and their ability to apply that learning as well.

The tasks will be based on college- and career-ready competencies across major academic disciplines including the Common Core State Standards-Aligned competencies for English Language Arts & Literacy and Mathematics, as well as New Hampshire's K-12 Model Science Competencies recently approved by the New Hampshire Board of Education (New Hampshire Department of Education, 2014). Performance tasks will be developed for elementary, middle, and high school grade spans. They will be used to compare student performance across the state in areas not tested by SBAC, such as the ability to apply learning strategies to complex tasks.

New Hampshire also partnered with the Center for Collaborative Education (CCE) and the National Center for the Improvement of Educational Assessment (NCIEA) to develop the Performance Assessment for Competency Education (PACE), designed to measure student mastery of college and career ready competencies (New Hampshire Department of Education, 2014). PACE includes a web-based bank of common and locally designed performance tasks, to be supplemented with regional scoring sessions and local district peer review audits.

Colorado, Kansas, and Mississippi have partnered with the Center for Education Testing & Evaluation at the University of Kansas to form the Career Pathways Collaborative. The partnership's Career Pathways Assessment System (cPass) is designed to measure high school student readiness for entry into college and/or the workforce (Center for Educational Testing & Evaluation (CETE), 2014). It uses a mix of multiple choice questions and performance tasks both in the classroom and in real-world situations to measure the knowledge and skills necessary for specific career pathways.

It's worth noting parenthetically that the Advanced Placement® (AP) testing program has long included an open-ended component known as a constructed response item and does allow for essays on a number of exams (English, world languages, history) as well as other artifacts of learning on a very small number of exams, such as the Studio Art exams portfolio. In addition, the College and Work Readiness Assessment (CWRA+) combines selected-response items with performance-based assessment to determine student proficiency in complex areas such as analysis and problem solving, scientific and quantitative reasoning, critical reading and evaluation, and critiquing an argument (Council for Aid to Education, 2014). When answering the selected-response items and when writing their own essay or memorandum, students refer to supporting documents such as letters, memos, photographs, charts, or newspaper articles.

Finally, as noted previously, both PARCC and SBAC include a limited number of performance assessments that require students to construct complex written responses to prompts (Partnership for Assessment of Readiness for College and Careers, 2014; Smarter Balanced Assessment Consortium, 2014). The tests will also incorporate some fairly innovative items that elicit a high level of student engagement and reasoning by requiring them to elaborate upon and provide evidence to support the answers they provide as well as more extended reading passages.

Project-Centered Assessment

Project-centered assessments represent a particularly ambitious subset of performance tasks. These assessments engage students in researching or solving open-ended, challenging problems over an extended period of time (Soland, Hamilton, & Stecher, 2013). What distinguishes project-centered assessment is the scope, complexity, and the time and resources these significant tasks require. Projects tend to involve more lengthy, multi-step activities, such as research papers, the extended essay required for the International Baccalaureate Diploma, or

assignments that conclude with a major student presentation of a significant project or piece of research.

For example, Envision Schools, a secondary-level charter school network in the San Francisco area, have made this kind of assessment a central feature of their instructional program, requiring students to conduct semester- or year-long projects that culminate in a series of products and presentations, which undergo formal review by teachers and peers (Stanford Center for Assessment Learning and Equity, 2014). A student or team of students might undertake an investigation of, say, locally sourced food—this might involve researching where the food they eat comes from, what proportion of the price represents transportation, how dependent they are on other parts of the country for their food, what choices they could make if they wished to eat more locally produced food, what the economic implications of doing so would be, whether doing so could cause economic disruption in other parts of the country as an unintended consequence, and so on. The project would then be presented to the class and scored by the teacher using a scoring guide that includes ratings of the students' use of mathematics and economics content knowledge; the quality of argumentation; the appropriateness of sources of information cited and referenced; the quality and logic of the conclusions reached; and overall precision, accuracy, and attention to detail.

Another well-known example is the Summit Charter Network of schools, also located in the Bay Area (Bill & Melinda Gates Foundation, 2014). While Summit requires students to master high-level academic standards and cognitive skills, the specific topics they study and the particular ways in which they are assessed are personalized, planned out according to their needs and interests. The school's schedule provides students ample time to work individually and in groups on projects that address key content in the core subject areas. And in the process, students assemble digital portfolios of their work, providing evidence that they have developed important cognitive skills (including specific "habits of success," the metacognitive learning skills associated with readiness for college and career), acquired essential content knowledge, and learned how to apply that knowledge across a range of academic and real-world contexts. Ultimately, the goal is for students to make public presentations of projects and products that can withstand public critique and are potentially publishable.

Collections of Evidence

Strictly speaking, collections of evidence are not assessments per se. Rather, they offer a way to *organize and review* a broad range of assessment products and results, so that educators can make accurate decisions about student readiness for academic advancement, high school graduation, or postsecondary programs of study (Conley, 2005; Oregon Department of Education, 2006). Portfolios are one manifestation of the collection of evidence model.

For example, New Hampshire recently introduced a technology portfolio for graduation, which allows students to collect evidence to show how they have met standards in this field. And the New York Performance Standards Consortium, which currently consists of more than forty in-state secondary schools and others beyond New York, received a state-approved waiver allowing its students to complete a graduation portfolio in lieu of some of New York's Regents Examination requirements. Students must compile a set of ambitious performance tasks for their portfolios, including a scientific investigation, a mathematical model, a literary analysis, and a history/social science research paper, sometimes augmented with other tasks such as an arts demonstration or analyses of a community service or internship experience. All of these are measured against clear academic standards and are evaluated using common scoring rubrics.

The state of Kentucky adopted a similar approach as a result of its Education Reform Act of 1990, which included KIRIS, the Kentucky Instructional Results Information System (Stecher, Rahn, Ruby, Alt, & Robyn, 1997). Implemented in 1992, KIRIS incorporated information from several assessment sources, including multiple-choice and short-essay questions, performance "events" requiring students to solve applied problems, and collections of students' best work in writing and mathematics (though students were assessed also in reading, social science, science, arts and humanities, and practical living/vocational studies). The writing assessment, which continued until 2012, was especially rigorous: in grades 4, 7, and 12, students submitted 3-4 pieces of written work to be evaluated, and in grades 5, 8, and 12 they completed on-demand writing tasks, with teachers assessing their command of several genres, including reflective essays, expressive or literary work, and writing that uses information to persuade an audience.

In 2009, the Oregon State Board of Education adopted new diploma requirements, specifying that students must demonstrate proficiency in a number of Essential Skills. These include goals in traditional subject areas such as reading, writing, and mathematics, but they also include a number of other complex, cross-cutting outcomes, such as the ability to think critically and analytically, to use technology in a variety of contexts, to demonstrate civic and community engagement, to demonstrate global literacy, and to demonstrate personal management and teamwork skills. Basic academic skills will be tested via the Smarter Balanced exam, while the remaining Essential Skills will be assessed via measures developed locally or selected from a set of approved methods (Oregon Department of Education, 2014).

Such approaches, in which a range of student assessment information is collected over time, permit educators to combine some or all of the elements on the continuum of assessments presented in figure 2. Doing so results in a fuller picture of student capabilities than is possible with any single form of assessment. And because this allows for the ongoing, detailed analysis of student work, it gives schools the option to assess their progress on relatively complex cognitive skills, which is very difficult to measure using occasional achievement tests.

Other Assessment Innovations

Recently, the Asia Society commissioned the RAND Corporation to produce an overview of models and methods for measuring 21st century competencies (Soland et al., 2013). The resulting report describes a number of models that closely map onto the range of assessments described above, in figure 2. However, it also describes "cutting edge measures" such as assessments of higher-order thinking used by the Program for International Student Assessment (PISA) and the Graduation Performance System (GPS) used by Asia Society schools.

Coordinated by the Organization for Economic Cooperation and Development, PISA is a test, first administered in 2000, designed to allow for comparisons of student performance among member countries. Administered every three years to randomly selected 15-year-olds, it assesses knowledge and skills in mathematics, reading, and science, but it is perhaps best known for its emphasis on problem solving skills and other more complex (sometimes referred to as "hard-to-measure") cognitive processes, which it gauges through the use of innovative types of test items.

Beginning in 2015, for example, PISA will introduce an on-line assessment of students' performance on tasks that require collaborative problem solving. Through interactions with a digital avatar (simulating a partner the student has to work with on a project), test-takers will demonstrate their skills in establishing and maintaining a shared understanding of a problem,

taking appropriate action to solve it, and establishing and maintaining team organization. Doing so requires a series of deeper learning skills including analyzing and representing a problem; formulating, planning, and executing a solution; and monitoring and reflecting on progress. During the simulation, students encounter scenarios in which the context of the problem, the information available, the relationships among group members, and the type of problem all vary, and they are scored based on their responses to the computer program's scenarios, prompts, and actions. Early evidence suggests that this method is quite effective in distinguishing different collaborative problem solving skill levels and competencies.

Developed collaboratively by Asia Society and the Stanford Center for Assessment, Learning, and Equity (SCALE), the Graduation Performance System (GPS) measures student progress in a number of areas, with particular emphasis on gauging how "globally competent" they are—i.e., how knowledgeable about international issues and able to recognize cross-cultural differences, weigh competing perspectives, interact with diverse partners, and apply various disciplinary methods and resources to the study of global problems. The GPS assesses critical thinking and communication, and it provides educators flexibility to make choices regarding the specific pieces of student work that are selected to illustrate student skills in these areas.

National testing organizations such as ACT and the College Board, makers of the SAT, are updating their systems of exams to keep them in step with the new demands of measuring college readiness, although these tests will remain in their current formats and not involve student-generated work products beyond an optional on-demand essay. ACT has introduced Aspire, a series of summative, interim, and classroom exams and optional measures of metacognitive skills, designed to determine whether students are on a path to college and career readiness from third grade on (ACT, 2014). The SAT is undergoing a series of changes to require test takers to cite evidence to a greater degree when making claims and to understand what they are reading more deeply than just being able to identify the sequence of events or cite key ideas in a passage. The vocabulary students are expected to know will be less esoteric and more connected to academic learning (The College Board, 2014). An essay option is available on both tests.

A great deal remains to be seen regarding how much these two tests will adapt to reflect new notions of readiness that encompass deeper understanding of a wider range of content and mastery of learning skills and strategies that generalize beyond reading and mathematics. For the foreseeable future, these tests will continue to consist primarily of selected-response items, with all of the attendant limitations of this particular testing method.

Metacognitive Learning Strategies Assessments

Metacognitive learning strategies, skills, and dispositions are the things students do to enable and activate thinking, remembering, understanding, and information processing more generally (Conley, 2014c). Metacognition occurs when learners demonstrate awareness of their own thinking, then monitor and analyze their thinking and decision-making processes or—as competent learners often do—recognize that they are having trouble and adjust their learning strategies.

Indeed, metacognitive capabilities often contribute as much or even more than does subject-specific content knowledge to students' success in college. When faced with challenging new coursework, students with highly developed learning strategies tend to have an important advantage over peers who can only learn procedurally (i.e., by following directions).

Similarly, assessments designed to gauge students' learning and socio-emotional skills offer an important complement to tests that measure content knowledge alone. Ideally, they can

provide teachers with useful insights into why students might be having trouble learning certain material or completing a particular assignment.

However, measures of these skills, strategies, and dispositions are subject to their own set of criticisms. For example, many of them rely on student self-reports—e.g., questionnaires about what was easy or difficult about an assignment—which can lead to what is called social desirability bias, where the student answers based on what they think is the desired response. This limits the use of self-reports for higher stakes purposes without additional triangulation of data to enhance confidence in the honesty of a student's responses. Critics also point out that while these types of measures may not be intended for this purpose, they can easily lead teachers to make character judgments about students, bringing an unnecessary source of bias into the classroom. Finally, the measurement properties of many early instruments in this area have been somewhat suspect, particularly when it comes to reliability. In short, while assessments of metacognition can be useful, educators and policymakers have good reason to take care in their use and in the interpretation of results until the instruments are further refined and validated for higher stakes uses.

Still, though, it is beyond dispute that many educators and, increasingly, policymakers are taking a closer look at such measures, excited by their potential to help have an impact on the achievement gap for underperforming students. For example, public interest has surged, of late, in the role that perseverance, determination, tenacity, and grit can play in learning (Duckworth & Peterson, 2007; MacCann, Duckworth, & Roberts, 2009; Tough, 2012). So, too, has the notion of academic mindset struck a chord with many practitioners who see evidence daily that students who believe that effort matters more than innate aptitude are able to perform better in a subject (Farrington, 2013). And researchers are now pursuing numerous studies of students' use of study skills, their time management strategies, and their goal setting capabilities.

In large part, what makes all of these skills, strategies, and dispositions so appealing is the recognition that such things can be taught and learned, and that the evidence suggests that all are important for success in and beyond school.

One of the best-known assessment tools in this area is Angela Duckworth's Grit Index (Duckworth Lab, 2014), which consists of a dozen questions that can be quickly completed by students and that can predict the likelihood of their completing high school or doing well in situations that require sustained focus and effort. Another, Carol Dweck's Growth Mindset program (mindsetworks, 2014), helps learners understand and change the way they think about how to succeed academically. The program focuses on teaching students that their attitude toward a subject is as important as any native ability they have for the subject.

EPIC's CampusReady instrument is designed to assess students' self-perceptions of college and career readiness in each of the Four Keys described earlier (Educational Policy Improvement Center, 2014b). It touches on many aspects of grit and academic mindset, as well as a number of other attitudes, habits, behaviors, and beliefs necessary to succeed at post-secondary studies.

The California Office to Reform Education (CORE) districts will incorporate measures of social-emotional competencies into their accountability system, starting in the 2014/2015 academic year (California Office to Reform Education (CORE), 2014). Four metacognitive assessments are currently being piloted across twenty CORE schools. These four metacognitive assessments are designed to measure growth mindset, self-efficacy, self-management, and social awareness. For each metacognitive assessment, one version has been selected from existing measures, while the other version has been developed in partnership with methodological experts in an effort to improve upon existing measures.

While a great deal of attention is currently being paid to measures of this type, they still face a range of challenges before they are likely to be used as widely or for as many purposes as traditional multiple-choice tests. Perhaps the greatest obstacle to their use is the fact that most rely to some degree on self-reported information, which is susceptible to social desirability bias, even when no stakes are attached to the assessment.

This issue can be addressed to some extent by triangulating responses and scores against other data sources, such as a test score or attendance record, or even other items in an instrument that are more behaviorally anchored rather than attitudinal. Inconsistencies can indicate the presence of socially desirable responses. Over time, students can be encouraged to provide more honest self-assessments, particularly if they know they will not be punished or rewarded excessively based on their responses. In fact, evidence from the use of CampusReady, EPIC's student self-report on college and career readiness, suggests students find the act of completing the instrument itself to be a valuable form of learning about what it takes to be ready for college and careers, independent of their responses, which enhances their willingness to respond truthfully. This is strengthened when students are provided reports of their responses and links to resources to help them improve their readiness and when they receive scores over multiple administrations, which allows identification of trends for the student.

The use of such instruments to generate longitudinal reports that ascertain overall trends that to determine if students are developing the learning strategies and mindsets necessary to be successful lifelong learners is perhaps one of their strongest attributes. They can help guide teachers and students to develop important strategies and capabilities that enhance learner success and enable deeper learning, but they should not be overemphasized or misused for high stakes purposes, certainly not until more work has been done to understand how best to use these types of instruments.

Toward a System of Assessments[11]

As the implementation of the Common Core proceeds, and as a number of states rethink their existing achievement tests, a golden opportunity may be presenting itself for states to move toward much better models of assessment. It may now be possible to create combinations of measures that not only meet states' accountability needs but that also provide students, teachers, schools, and postsecondary institutions with valid information that empowers them to make wise educational decisions.

Today's resurgent interest in performance tasks, coupled with new attention to the value of metacognitive learning skills, invites progress toward what I like to call a "system of assessments," a comprehensive approach that draws from multiple sources in order to develop a holistic picture of student knowledge and skills in all of the areas that make a real difference for college, career, and life success.

The new PARCC and SBAC assessments have an important contribution to make to this effort, in that they offer well-conceived test items as well as carefully designed performance tasks that require valuable writing skills and problem solving capabilities. These assessments should help signal to students that they are expected to engage deeply in learning and to devote serious time and effort to developing higher-order thinking skills. On their own, however, the Common Core assessments are not a system.

---

[11] Portions of this section are excerpted or adapted from: Conley, D. T. and L. Darling-Hammond (2013). Creating systems of assessment for deeper learning. Stanford, CA: Stanford Center for Opportunity Policy in Education. https://edpolicy.stanford.edu/publications/pubs/1075

A genuine system of assessments would address the varied needs of *all* of the constituents who use assessment data, including public schools; postsecondary institutions; state education departments, state and federal policymaking bodies, education advocacy groups; business and community groups; and others. It would serve purposes that go well beyond the task of rating schools, judging them to be successes or failures. Most importantly, it would avoid placing too much weight on any single source of data. In short, and as I describe below, such a system would produce a nuanced and multi-layered profile of student learners.

A Profile Approach to Readiness and Deeper Learning[12]

A system of assessments yields many more data points than does a single achievement test. Compared to the familiar connect-the-dots sketch of students' knowledge and skills, it offers a much more precise, high-definition picture of where they are, how far they've come, and how far they have to go in order to be ready for college and careers.

Ultimately, this should allow educators to create profiles of individual students that are far more detailed than the familiar high school transcript, which tends to include teacher-generated grades and little more. Rather, it should be possible to generate a more integrative and personalized series of measures, calibrated to individual student goals and aspirations, that highlights much more of what those students know and are able to do.

Such a profile might be thought of metaphorically something like a gears-and-wheels structure, with information from various sources supporting multiple uses and interpretation, rather than a hierarchical model in which information is organized into strict levels. The largest "gear" would include source information on classroom-level performance from a variety of sources more informative than an overall grade. Examples include research papers and capstone projects, students' assessments of their own key learning skills over multiple years, indicators of perseverance and goal focus as evidenced by their completion of complex projects, and teachers' judgments of student characteristics. Other, smaller gears represent more familiar sources of information on readiness, such as, college admission tests, the PARCC and SBAC assessments or state tests, and course-taking patterns such as advanced coursework, dual enrollment, AP, and IB. Another gear might contain information from assessments in other subject areas, and on metacognitive skills, dispositions, and strategies. It is possible to envision other gears in this model that contribute to a dynamic system of data that drives the instructional process. Figure 3 illustrates a model comprising the three gears just described.

[12] For a more detailed discussion of profiles that uses a different metaphor, see: Conley, D. T. (2014). New conceptions of college and career ready: A profile approach to admission. *The Journal of College Admission* (223).

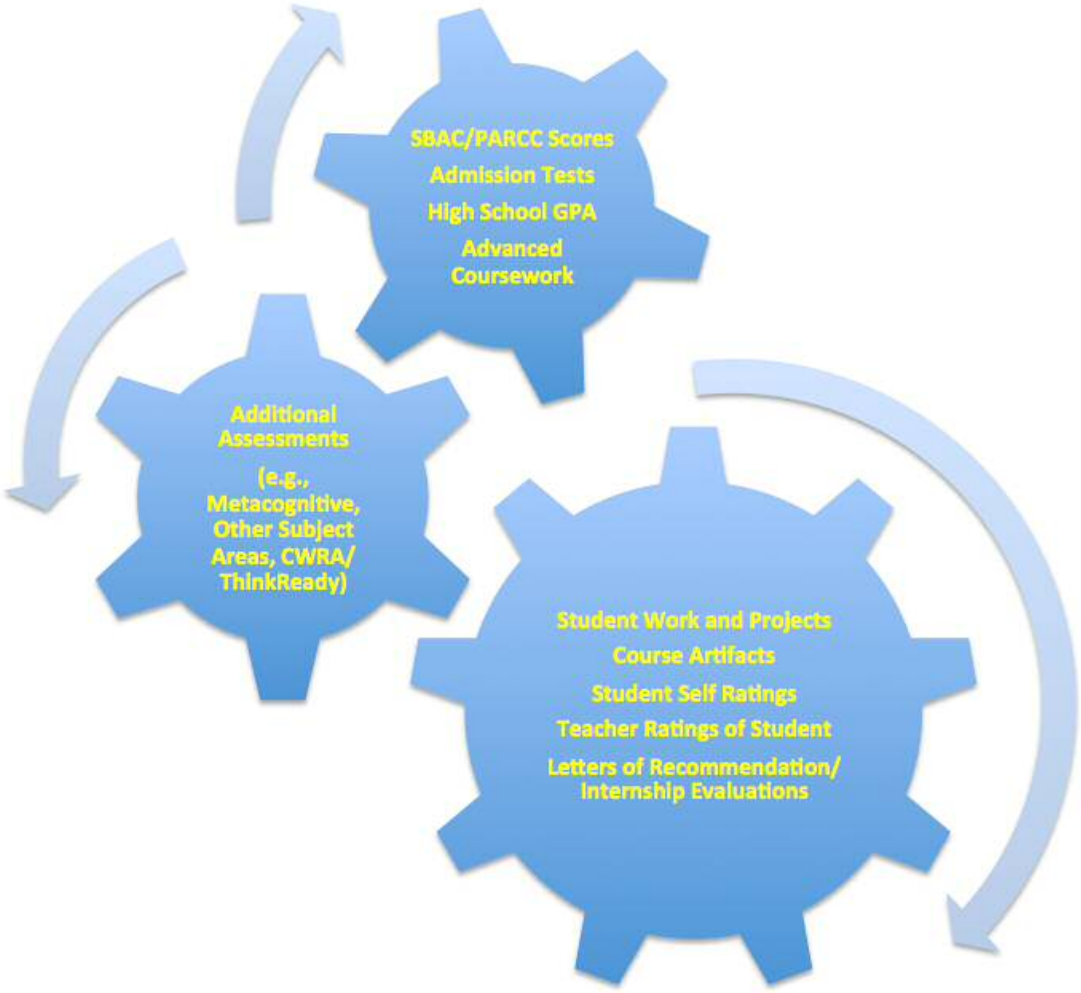

Figure 3. Example of potential data in a multi-dimensional student profile

Subordinate levels of the profile would contain additional information including actual student work with insights into the techniques and strategies they used to generate the work. Student work would be sorted and categorized through the use of metadata tags to array it by characteristic that would make it easy and convenient for a reviewer to pull up samples based on areas of interest, such as interpretive thinking or research or mathematical reasoning.

Note, however, that this would not be the same as the type of portfolio of student work that simply reflects general student growth and achievement. While portfolios of work can be useful within schools for a wide range of purposes, translating them for use outside the school requires systematic organization and additional information that makes them more readily interpretable. A profile approach that offers layers of data (including student work) in a more readily accessible form could serve not just individual students and their teachers and parents but also a range of potential external users, too, such as college admissions officers, advisors, and instructors or potential employers. To be sure, safeguards would have to be in place in order to ensure students' privacy and protect against misuse of their information—just as is true today of student transcripts. But with reasonable safeguards, a profile should offer quite useful

insights into students' progress and valuable diagnostic information that can be used to help students prepare better for college and careers.

Challenges Of Deeper Learning Assessment

Today's data management technologies are sufficiently sophisticated and efficient enough to handle the complex information generated by such a system of assessments. Profiles would, however, still face a series of daunting challenges in order to be implemented successfully and on a large scale.

Although no one has completely cracked the code, some states, researchers, and testing organizations are seeking to develop new methods to assess deeper learning skills on a large scale. For example, the International Baccalaureate has long included a course on theories of knowledge that incorporates in-class assessment and a culminating project. Self-reflective thought is the central focus of Theory of Knowledge courses within the IB Diploma Programme. The primary purpose of the course is to engage student in discussion and self-reflection about how they know what they claim to know. The course culminates with students writing a paper and giving presentation that demonstrates their mastery of these ways of knowing and of thinking. Additionally, the Extended Essay program requires students to analyze, synthesize, and evaluate knowledge by engaging in an in-depth research study that mirrors undergraduate research requirements.

An example of a locally-developed program of deeper learning and assessment can be found in the Danville Independent School District in Kentucky. The Danville Diploma is a set of 11 skills the district commits to develop in its students. Formative assessments help students gauge how well they are using learning skills to achieve lesson outcomes. Additionally, students collect evidence demonstrating progress toward the meeting the Danville Diploma competencies. Several of these skills reflect a commitment to deeper learning, including the following:

- discover how critical thinking skills are used across disciplines;
- adapt and problem solve;
- manage time and create a plan for accomplishing a task or goal;
- know how to find reliable and accurate information; and
- analyze, synthesize, and make inferences from data.

In these examples, work is scored locally by teachers. Scaling up entails attention to large-scale scoring challenges. Coming up with reliable ways to score complex student work may be the holy grail of performance assessment of deeper learning. Until and unless test designers are able to learn to use the most up-to-date methods of automated scoring of open-ended assignments and devise even better ways to score complex student work products, either by teachers or externally, and until states refuse to buy assessments that don't meet these criteria, the Common Core standards that reflect deeper learning will largely be under-assessed, at least by tests used for high-stakes accountability purposes.

As long as the primary purpose of assessments is to reach judgments about students and schools (and, increasingly, teachers), reliability and efficiency will continue to trump validity. Thankfully, though, one important lesson to emerge from No Child Left Behind—and its decade-long rush to judge the quality of individual schools—is that not all assessments are, or should be, summative. The Standards for Educational and Psychological Testing (2014) make it clear that no test should be used for consequential decisions about a student, teacher, or school without additional information—the more consequential the decision, the more information that is necessary.

The majority of the assessment that goes on every day in schools is designed not to hold anybody accountable but to help people make immediate decisions about how to improve student performance and teaching practice. Over the past ten years, educators have learned the distinction between summative and formative assessments, and they know full well that not all measures must be high stakes in nature or that all judgments need be derived from multiple-choice tests. While it will always be important to know how well schools are teaching foundational skills in English language arts and mathematics, the pursuit of deeper learning will require a much greater emphasis on formative assessments that *signal* what students must do to become ready for college and careers and how teachers can best help them, including measures of metacognitive learning skills—about which selected response tests provide no information at all.

Skills such as persistence, goal focus, attention to detail, investigation, and information synthesis are more likely to be the most important for success in the coming decades. It will become increasingly critical for young people to learn how to cope with college assignments or work tasks that do not have one right answer, that require them to gather new information and make judgments about the information they collect, and that may have no simple or obvious solution. Such integrative and applied skills can be assessed, and they can be assessed most usefully by way of performance assessments. They neither can nor should be measured at the granular level that is the focus of most standardized tests.

A final, though by no means trivial, question is whether the nation's postsecondary institutions, having relied for so many decades on multiple-choice tests to help them make admission and placement decisions, can or will use information from assessments of deeper learning, if such sources of data exist. Although an increasing number of postsecondary institutions are either not using SAT or ACT tests or are making them optional and many already employ a form of portfolio review, most still rely on a few numbers that do not represent well the full set of knowledge and skills necessary for postsecondary success.

Will the next generation of state assessments help address this shortcoming by providing more valid and comprehensive information on college and career readiness? As noted, the Common Core tests will generate a cut score that states may or may not use. Beyond that, most states are not giving much thought to how to provide postsecondary institutions with additional, new, and richer information on the deeper learning skills associated with postsecondary success, many of which are included in the Common Core State Standards but not measured by the assessments. The most that is being done with the new assessment data is to exempt students from remedial education requirements if they meet the cut score. This practice is problematic in and of itself, in part because it tends to validate current remedial practices that are demonstrably not effective, and because it sends a message to students regarding their readiness that may not be at all accurate.

One of the side effects of emphasizing the Common Core assessments as readiness indicators is that many (but not all) postsecondary institutions are doing little to signal any interest in more complex information on readiness nor to work with secondary education to develop the data collection and interpretation systems necessary to use results from profiles, portfolios, and performance tasks to gain more interesting and potentially useful insights into student readiness.

Again, it is worth noting that PARCC and SBAC assessments can be a useful step forward if their results are combined thoughtfully with other information. At the same time, though, these assessments should not be mistaken for the kind of bold leap that will be required

in order to capture the full range of student knowledge, skills, abilities, and strategies associated with postsecondary readiness and success.

The postsecondary community seems to be spread along a continuum from being resigned to having to accommodate more information to being eager to gain greater insight into student readiness. While concerns always exist at larger institutions about how to process more diverse data for thousands of applicants, innovative campuses and systems are already gearing up to make decisions more strategically and to learn how to use something more like a profile of readiness rather than just a cut score for eligibility.

An integrated system of state and local assessments, then, consists of a variety of measures that serve different purposes. State assessments should best be viewed more like a thermometer that gives a quick, one-dimensional reading on overall health. Local assessments fill a more focused and situational role, shedding light on what's happening with learning more in real time and providing information on how to improve learning. In a multi-measure system, more assessment is diagnostic, even at the state level (an example would be NAEP-like sample tests administered statewide), and some assessments can be both formative (if they are rich in information and used to inform next steps in classroom instruction) and summative (if they give information about what has been learned—such as the New York Performance Standards Consortium described earlier). The problem with selected response tests is that they are not really able to be diagnostic, and it does not make sense to take up most of the available bandwidth for testing in terms of time and resources with testing that doesn't yield information that is useful for multiple purposes.

Recommendations

Many issues will need to be addressed in order to bring about the fundamental changes in assessment practice necessary to promote and value deeper learning. The recommendations offered here are meant to serve as a starting point for a process that likely will unfold over many years, perhaps even decades. The question is: Can policymakers sustain their attention to this issue long enough to enact the policies necessary to bring about necessary changes? For that matter, can educators follow through with new programs and practices that turn policy goals into reality? And will the secondary and postsecondary systems be able to cooperate in creating a system of assessments and focusing instruction on deeper learning?

To move toward these goals, education policymakers will need to:

1. **Define college and career readiness comprehensively.** States need clear definitions of college and career readiness that highlight the full range of knowledge, skills, and dispositions that research shows to be critical to students' success beyond high school (including not only key content knowledge but also cognitive strategies, learning skills and dispositions, and knowledge and skills necessary for successful transition to college and the workforce).
2. **Take a hard look at the pros and cons of current state accountability systems.** If policymakers agree that college and career readiness entails far more than just a narrow set of academic skills and knowledge, then they should ask themselves how well—or poorly—existing state assessments measure the full range of things that matter to students' long-term success. Further, policymakers should take stock of the real-world impacts that the existing assessment models have had on teaching and learning. For well over a decade, proponents of high-stakes testing have asserted that the prevailing model of accountability creates strong incentives for teachers and schools to improve. However, high-stakes testing is past due for an assessment of its own. State leaders should ask themselves: Are the existing tests as they are being used to

evaluate teacher and school performance truly having the desired impact? In reality, what changes in instruction do teachers make in response to summative testing and to test scores being used to evaluate their performance and their school's performance? How much time and money is currently devoted to such tests, at what opportunity cost? That is, to what extent could high-stakes testing be crowding out other, more useful ways of assessing student progress? How might high-stakes testing be closing educators off to more useful forms of assessment?

3. **Support the development of new assessments of deeper learning.** Across the country, many efforts are now underway to create assessments that address a wide range of knowledge and skills, going well beyond reading and mathematics, and these efforts need to be encouraged and nurtured. However, several key problems will need to be resolved if assessments of deeper learning are to be scalable, reliable, and useful enough to justify their expense. In particular, when it comes to measures that require students to report on their own progress, or that require teachers to rate students in some way, strategies will have to be developed by which to triangulate these reports against other data sources in order to ensure a reasonable level of consistency. Further, it will be extremely important to institute safeguards to protect students' privacy and ensure that this sort of information is not used inappropriately. And, finally, policymakers and educators will have to be careful to distinguish between assessment tools that are meant to generate information that can be used to improve teaching and learning, those that can only be used as the basis for summative judgments about students' learning or teachers' performance, and those that might serve both purposes to some degree.
4. **Learn from past efforts to build statewide performance assessment systems.** States' pioneering efforts to develop performance assessments in the 1990s and early 2000s yielded a wealth of lessons that can inform current attempts to expand assessment beyond a limited set of selected-response tests. Most important is the need to proceed slowly at first, in order to develop systems by which to manage the sometimes-complex mechanics of collecting, analyzing, interpreting, and reporting information from these types of richer assessments. Educators, especially, must have sufficient time to learn how to work with new assessments, not only how to score them but how to teach to them successfully. The technological infrastructure also needs to be in place to store, organize, and retrieve data in ways that makes information from performance assessments useful to educators and other constituents.
5. **Take greater advantage of advances in information technology.** Many of the challenges that confronted states 25 years ago, when they first adopted performance assessment systems, can be addressed better today through the use of vastly more sophisticated technology for information analysis, storage, and retrieval. Online storage is plentiful and cheap, and it's far easier and cheaper to transmit documents to scorers, for example, than it was to convene scorers or mail them documents. The technological literacy level of educators is higher, and the capabilities of postsecondary institutions to receive information electronically have greatly increased. If districts and states take advantage of this new capacity to manage complex data in useful and user-friendly ways, they should find it much easier than in past decades to store student data in digital portfolios and access that information to meet the needs of audiences such as educators, admissions officers, parents, students themselves, and perhaps potential employers.
6. **Make it clear that federal education policy allows greater flexibility in the types of data that can be used to demonstrate student learning and growth.** The U.S. Department of Education's waiver process has introduced some flexibility with respect to the measures of student learning that states—and, in at least one case, a consortium of school districts—can use to meet federal accountability requirements. However, states seem to be largely unaware that

they can propose more complex systems of assessment, or they are portraying their unwillingness to entertain such options as the fault of NCLB requirements. The U.S. Department of Education should take the initiative to publish a white paper that encourages greater diversity of assessment and a more thoughtful mix of state and local measures, with particular attention to those that measure deeper learning and other dimensions of college and career readiness. In addition, any eventual reauthorization of the Elementary and Secondary Education Act and its NCLB provisions should go much further to encourage the use of multiple forms of assessment and to make clear to states that such models can pass federal muster or that the federal government simply will not interfere in their use.

7. **Consider using an improved and upgraded version of the National Assessment of Educational Progress as a baseline measure of student problem solving capabilities.** NAEP has many features that make it attractive as a potential measure of problem solving. The design of NAEP, particularly the fact that not all test-takers are asked to complete the entire battery of NAEP items, allows it to include fairly complex and time-intensive tasks. This design characteristic can be used both to field-test more complex performance items as well as to generate a better national metric of student problem-solving skills in the areas NAEP assesses. Having a baseline that is consistent across states can help determine which states are making the most progress with their statewide systems of assessment of deeper learning. PISA, too, could be used in this fashion, but the implementation challenges would be much greater than building upon NAEP's existing infrastructure. The challenge will be to make NAEP reading and math tests better measures of more cognitively complex tasks of the type associated with problem solving. However, NAEP has been experimenting with new types of items and tasks that look promising as potential measures of deeper learning.
8. **Build a strong base of support for a comprehensive system of assessments**. The process of developing a more complex system of assessments must not exclude any major group of stakeholders. Teachers in particular need to be centrally involved in designing, scoring, and determining how data from rich assessments of student learning will be used. State policymakers, too, have a compelling interest in finding ways to make sure that those assessments are both valid and reliable. And postsecondary and business leaders must have a seat at the table, as well, if they will be expected to make use of any new sources of information about students' college and career readiness. Including stakeholders in the design of a system of assessments is one more way to enhance the overall validity of results as well as to encourage a wider range of measures and methods.
9. **Determine the professional learning, curriculum, and resource needs of educators.** Currently, few states do much, if anything, to gauge the capacity of individual schools to provide meaningful opportunities for professional learning. As a result, most schools are unable to help their teachers acquire essential new skills, such as teaching more cognitively complex material to all students and measuring it with in-class assignments and assessments. In order to implement a system of assessments successfully, it will be absolutely critical to determine—early on in the process—what resources will be necessary to ensure that all teachers are assessment literate, are able to use the information generated by multiple sources, are capable of developing assignments that lead to deeper learning, and are able to teach the full range of content and skills that prepare students to succeed in college and careers. It's worth noting that few state education departments or intermediate service agencies currently have the capacity to offer the level of guidance and support most schools, particularly those in smaller districts, need to undertake the type of professional learning program necessary to implement and use a system of assessments approach to instructional improvement. These agencies must be strengthened as well in their

ability to provide essential support services to schools around professional learning related to assessment.

10. **Look for ways to improve the Common Core State Standards and their assessments so that they become better measures of deeper learning.** This may be a tall order at a time when Common Core implementation is undergoing sustained challenge. However, the surest way to undermine the credibility of the standards and the assessments would be to refuse to improve them in response to feedback from the field. Such a stance hardens the opposition and leads educators to view them as just another mandate to be complied with, rather than as a dynamic source of professional guidance and growth. Already, the standards are almost five years old, and it is past time to begin the lengthy process of designing and initiating a careful and systematic review process. Similarly, even though PARRC and SBAC are only just now completing their field testing, their designers must continue to seek out criticism, keep a close eye on their roll-out, communicate more frankly and vocally the limitations of these assessments, while simultaneously suggesting ways to get at the various aspects of college and career readiness that these assessments currently overlook. In particular, the test developers should push back against forces that seek to decrease the number of performance tasks and the use of data from them. Instead, they should be laying out a path to increased integration of such tasks for states that wish to do so eventually. Such a path would pay more attention to how performance tasks administered in the classroom would be incorporated into overall results or used for formative purposes, or both.

Ideally, the educational assessment system of the future will be analogous to a thorough, high-quality medical diagnostic procedure, rather than the cursory check-up described at the beginning of this article. Educators and students alike will have at their disposal far more sophisticated and targeted tools and techniques to determine where they are succeeding, to show where they are falling short, and to point out how and what to improve. They will receive rich, accurate information about the cause of any learning problems, and not just the symptoms or the effects. Policymakers will understand that improved educational practice, just like improved health, is rarely achieved by compelling people to follow uniform practices or using data to threaten them, but, rather, by creating the right mix of incentives and supports that motivate and reward desired actions, and that help all educational stakeholders to understand which outcomes are in their mutual best interests.

Research and experience make it clear that educational systems that can get students to demonstrate deeper learning must incorporate assessments that honor and embody these goals. New systems of assessment, connected to appropriate resources, learning opportunities, and productive visions of accountability, comprise a critical foundation for enabling students to meet the challenges that face them throughout their education and careers in the 21st century.

References

Achieve, The Education Trust, & Thomas B. Fordham Foundation. (2004). The American Diploma Project: Ready or not: Creating a high school diploma that counts. Washington, D.C.: Achieve, Inc.

ACT. (2011). ACT College Readiness Standards. Retrieved November 23, 2011, from http://www.act.org/standard/

ACT. (2014). In a nutshell- ACT Aspire. Retrieved September 5, 2014, from http://www.discoveractaspire.org/in-a-nutshell.html

American Educational Research Association [AERA], American Psychological Association [APA], & National Council for Measurement in Education [NCME]. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Baldwin, M., Seburn, M., & Conley, David T. (2011). *External validity of the College-Readiness Performance Assessment System (C-PAS)*. Paper presented at the 2011 American Educational Research Association (AERA) Annual Conference New Orleans, LA Roundtable Discussion: Assessing College Readiness, Innovation, and Student Growth retrieved from

Bill & Melinda Gates Foundation. (2014). Summit Public Schools. *Next Generation Learning Challenges*. Retrieved September 5, 2014, from http://nextgenlearning.org/grantee/summit-public-schools

Block, J. H. (1971). *Mastery learning: Theory and practice*. New York: Holt, Rinehart, & Winston.

Bloom, B. (1971). *Mastery learning*. New York: Holt, Rinehart, & Winston.

Brandt, R. (1992/1993). On outcome-based education: A conversation with Bill Spady. *Educational Leadership, 50*(4), 66-70.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (2000). *How people learn: Brain, mind, experience, and school. Expanded edition*: National Academy of Sciences - National Research Council Washington DC. Commission on Behavioral and Social Sciences and Education, US Department of Education, Washington, DC.

Brown, R. S., Conley, David T. (2007). Comparing state high school assessments to standards for success in entry-level university courses. *Journal of Education Assessment, 12*(2), 137 - 160. http://dx.doi.org/10.1080/10627190701232811

California Office to Reform Education (CORE). (2014). CORE NCLB waiver request: School Quality Improvement System. Retrieved September 5, 2014, from http://coredistricts.org

Cawelti, G. (2006). The side effects of NCLB. *Educational Leadership, 64*(3), 64-68.

Center for Educational Testing & Evaluation (CETE). (2014). Career pathways collaborative offers new assessments for career and technical education. Retrieved September 5, 2014, from http://careerpathways.us/news/career-pathways-collaborative-offers-new-assessments-career-and-technical-education

Cherry, K. (2014). History of intelligence testing: The history and development of modern IQ testing. *About Education*. Retrieved September 9, 2014, from http://psychology.about.com/od/psychologicaltesting/a/int-history.htm

Conley, D. T. (2003). *Understanding university success*. Eugene, OR: Center for Educational Policy Research, University of Oregon.

Conley, D. T. (2005). Proficiency-based admissions. In W. Camara & E. Kimmell (Eds.), *Choosing students: Higher education admission tools for the 21st Century*. Mahwah, New Jersey:: Lawrence Erlbaum.

Conley, D. T. (2007). The College-readiness Performance Assessment System (C-PAS). Eugene, Oregon: Educational Policy Improvement Center.

Conley, D. T. (2011). The Texas College and Career Readiness Initiative: Overview & summary report. Eugene, OR: Educational Policy Improvement Center.

Conley, D. T. (2014a). The Common Core State Standards: Insight into their development and purpose. Washington, DC: Council of Chief State School Officers.

Conley, D. T. (2014b). *Getting ready for college, careers, and the Common Core: What every educator needs to know*. San Francisco, CA: Jossey-Bass.

Conley, D. T. (2014c). Learning strategies as metacognitive factors: A critical review. In Prepared for the Raikes Foundation (Ed.). Eugene, Oregon: Educational Policy Improvement Center.

Conley, D. T., Aspengren, K., & Stout, O. (2006). Advanced Placement Best Practices Study: Biology, Chemistry, Environmental Science, Physics, European History, US History, World History. Eugene, Oregon: Educational Policy Improvement Center.

Conley, D. T., Aspengren, K., Gallagher, K., Nies, K. (2006a). College Board validity study for math. Eugene: College Board.

Conley, D. T., Aspengren, K., Gallagher, K., Nies, K. (2006b). College Board validity study for science. Eugene: College Board.

Conley, D. T., Aspengren, K., Gallagher, K., Stout, O., Veach, D., Stutz, D. (2006). [College Board Advanced placement best practices course study].

Conley, D. T., Brown, R. (2003). *Analyzing state high school assessments to determine their relationship to university expectations for a well-prepared student*. Paper presented at the American Educational Research Association, Chicago, IL.

Conley, D. T., Drummond, K. V., DeGonzalez, A., Rooseboom, J., & Stout, O. (2011). Reaching the Goal: The applicability and importance of the Common Core State Standards to college and career readiness. Eugene, OR: Educational Policy Improvement Center.

Conley, D. T., McGaughy, C., Brown, D., van der Valk, A., & Young, B. (2009). Texas Career and Technical Education career pathways analysis study. Eugene, Oregon: Educational Policy Improvement Center.

Conley, D. T., McGaughy, C., Brown, D., van der valk, A., & Young, B. (2009). Validation study III: Alignment of the Texas College and Career Readiness Standards with courses in two career pathways. Eugene, OR: Educational Policy Improvement Center.

Conley, D. T., McGaughy, C., Cadigan, K., Flynn, K., Forbes, J., & Veach, D. (2008). Texas College and Career Readiness Initiative Phase II: Examining the alignment between the Texas College and Career Readiness Standards and entry-level college courses at Texas postsecondary institutions. Eugene, OR: Educational Policy Improvement Center.

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

Conley, D. T., McGaughy, C., O'Shaughnessy, T., & Rivinus, E. (2007). College-readiness Performance Assessment System (C-PAS) conceptual model. Eugene.

Council for Aid to Education. (2014). CWRA+ overview. Retrieved September 5, 2014, from http://cae.org/participating-institutions/cwra-overview/

Council of Chief State School Officers, & National Governors Association. (2010a). Common Core State Standards for English Language Arts & literacy in history/social studies, science, and technical subjects. Retrieved September 6, 2010, from Author http://www.corestandards.org/assets/CCSSI_ELA Standards.pdf

Council of Chief State School Officers, & National Governors Association. (2010b). Common Core State Standards for mathematics. Retrieved September 6, 2010, from Author http://www.corestandards.org/assets/CCSSI_Math Standards.pdf

Criterion-referenced test. (2014, April 30). Retrieved September 9, 2014, from http://edglossary.org/criterion-referenced-test/

Donovan, M. S., Bransford, J. D., & Pellegrino, J. W. (Eds.). (1999). *How people learn: Bridging research and practice*. Washington DC: National Academy Press.

Duckworth, A. L., & Peterson, C. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology, 92*(6), 1087-1101. http://dx.doi.org/10.1037/0022-3514.92.6.1087

Duckworth Lab. (2014). Research & measures: Scales and measures. Retrieved September 5, 2014, from https://sites.sas.upenn.edu/duckworth/pages/research

Dudley, M. (1997). The rise and fall of a statewide assessment system. *English Journal, 86*(1), 15-20. http://dx.doi.org/10.2307/820774

Dweck, C. S., Walton, G. M., & Cohen, G. L. (2011). Academic tenacity: Mindsets and skills that promote long-term learning. Seattle: Gates Foundation.

Educating for the 21st century: Data report on the New York Performance Standards Consortium. (n/d): Performance Standards Consortium.

Educational Policy Improvement Center. (2014a). National Assessment of Educational Progress grade 12 preparedness research college course content analysis study final report. Eugene, Oregon: Author.

Educational Policy Improvement Center. (2014b). ThinkReady: An innovative, formative assessment of student cognitive abilities! *College & Career Readiness: A Comprehensive Approach.* Retrieved September 5, 2014, from https://collegeready.epiconline.org/info/thinkready.dot

Farrington, C. A. (2013). Academic mindsets as a critical component of deeper learning. University of Chicago: Consortium on Chicago School Research.

Gewertz, C. (2013a). Assessment consortia: Who's in and who's out? *Education Week*. Retrieved from Curriculum Matters website: http://blogs.edweek.org/edweek/curriculum/2013/10/missouri_chooses_vendor_for_20.html?qs=utah+common+core+assessment

Gewertz, C. (2013b). Teachers' union president: Halt all high stakes linked to Common Core. *Education Week*. Retrieved from Curriculum Matters website: http://blogs.edweek.org/edweek/curriculum/2013/04/halt_high_stakes_linked_to_common_core.html

Gewertz, C. (2014). Opting out of testing: A rising tide for states and districts? *Education Week*. http://blogs.edweek.org/edweek/curriculum/2014/03/opting_out_of_testing.html?qs=parents+common+core+testing

Goodlad, J., & Oakes, J. (1988). We must offer equal access to knowledge. *Educational Leadership, 45*(5), 16-22.

Guskey, T. R. (1980a). Individualizing within the group-centered classroom: The mastery learning model. *Teacher Education and Special Education, 3*(4), 47-54. http://dx.doi.org/10.1177/088840648000300408

Guskey, T. R. (1980b). Mastery learning: Applying the theory. *Theory Into Practice, 19*(2), 104-111. http://dx.doi.org/10.1080/00405848009542882

Guskey, T. R. (1980c). What is mastery learning? And why do educators have such hopes for it? *Instructor, 90*(3), 80-82,84,86.

Hambleton, R. K., Impara, J., Mehrens, W., & Plake, B. S. (2000). Psychometric review of the Maryland School Performance Assessment Program (MSPAP). Annapolis, Maryland: Maryland State Department of Education.

Hinton, C., Fischer, K. W., & Glennon, C. (2012). Mind, brain, and education. Boston: Jobs for the Future/Nellie Mae Education Foundation.

Horton, L. (1979). Mastery learning: Sound in theory, but... *Educational Leadership*.

Jennings, J., & Rentner, D. S. (2006). Ten big effects of the No Child Left Behind Act on public schools. *Phi Delta Kappan, 82*(2), 110-113. http://dx.doi.org/10.1177/003172170608800206

Jobs for the Future. (2005). Education and skills for the 21st century: An agenda for action. Boston, MA: Jobs for the Future.

Kirst, M. W., & Mazzeo, C. (1996). The rise, fall, and rise of state assessment in California, 1993-96. *Phi Delta Kappan, 78*(4), 319-323.

Knecht, D. (2007). The Consortium and the Commissioner: A grass roots tale of fighting high stakes graduation testing in New York. *Urban Review: Issues and Ideas in Public Education, 39*(1), 45-65. http://dx.doi.org/10.1007/s11256-007-0043-0

Koretz, D., Stecher, B., & Deibert, E. (1993). The reliability of scores from the 1992 Vermont Portfolio Assessment Program: CRESST; Center for the Study of Evaluation, University of California, Los Angeles.

Linn, R. L. (2005). Conflicting demands of No Child Left Behind and state systems: Mixed messages about school performance. *Education Policy Analysis Archives, 13*(33).

Linn, R. L., Baker, E. L., & Betenbenner, D. W. (2002). Accountability systems: Implications of requirements of the No Child Left Behind Act of 2001. Los Angeles: Center for The Study of Evaluation & National Center for Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles.

MacCann, C., Duckworth, A. L., & Roberts, R. D. (2009). Empirical Identification of the Major Facets of Conscientiousness. *Learning and Individual Differences, 19*(4), 451-458. http://dx.doi.org/10.1016/j.lindif.2009.03.007

Medina, J. (2008). *Brain rules: 12 principles for surviving and thriving at work, home, and school.* Seattle: Pear Press.

mindsetworks. (2014). Movitate students to grow their minds! Retrieved September 5, 2014, from http://www.mindsetworks.com

Mintrop, H., & Sunderman, G. L. (2009). Predictable failure of federal sanctions-driven accountability for school improvement--and why we may retain it anyway. *Educational Researcher, 38*(5), 353-364. http://dx.doi.org/10.3102/0013189X09339055

National Research Council. (2002). Learning and understanding: Improving advanced study of mathematics and science in U.S. high schools. Washington, DC: National Academy Press.

New Hampshire Department of Education. (2014). New Hampshire Performance Assessment Network. Retrieved September 5, 2014, from http://www.education.nh.gov/assessment-systems/

Oakes, J. (1985). *Keeping track: How schools structure inequality*. New Haven, CT: Yale University Press.

Ohio Department of Education. (n.d.). Ohio Performance Assessment Pilot Project (OPAPP). Retrieved September 5, 2014, from http://education.ohio.gov/Topics/Testing/Next-Generation-Assessments/Ohio-Performance-Assessment-Pilot-Project-OPAPP

Oregon Department of Education. (2006). Career-related learning standards and extended application standard: Guide for schools to build relevant and rigorous collections of evidence (pp. 66).

Oregon Department of Education. (2014). Oregon's Essential Skills. Retrieved September 5, 2014, from http://www.ode.state.or.us/search/page/?id=2042

Partnership for Assessment of Readiness for College and Careers. (2014). Item and task prototypes. Retrieved September 5, 2014, from http://www.parcconline.org/sample-assessment-tasks

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.

Rothman, R. (1995). The Certificate of Initial Mastery. *Educational Leadership, 52*(8), 41-45.

Sawchuk, S. (2014). New York union encourages districts to boycott field tests. *Education Week*. Retrieved from Teacher Beat website: http://blogs.edweek.org/edweek/teacherbeat/2014/05/new_york_union_encourages_dist.html?qs=new+york+common+core

Seburn, M., Frain, S., & Conley, D. T. (2013). Job training programs curriculum study. Washington, DC: National Assessment Governing Board, WestEd, Educational Policy Improvement Center.

Secretary's Commission on Achieving Necessary Skills (SCANS). (1991). What work requires of schools: A SCANS report for American 2000: U.S. Department of Labor.

Slavin, R. E. (1987). Mastery learning reconsidered. *Review of Educational Research, 57*, 175-213. http://dx.doi.org/10.3102/00346543057002175

Smarter Balanced Assessment Consortium. (2014). Sample items and performance tasks. Retrieved September 5, 2014, from http://www.smarterbalanced.org/sample-items-and-performance-tasks/

Soland, J., Hamilton, L. S., & Stecher, B. M. (2013). Measuring 21st century competencies: Guidance for educators *Global Cities Education Network*: Asia Society/RAND Corporation.

Stanford Center for Assessment Learning and Equity. (2014). Envision Schools college success portfolio. Retrieved September 5, 2014, from https://scale.stanford.edu/content/envision-schools-college-success-portfolio

Stecher, B. M., Rahn, M. L., Ruby, A., Alt, M. N., & Robyn, A. (1997). Using alternative assessments in vocational education: Appendix B: Kentucky Instructional Results Information System (KIRIS). Berkeley, CA: National Center for Research in Vocational Education.

Structural inequality in education. (2014). Retrieved September 9, 2014, from http://en.wikipedia.org/wiki/Structural_inequality_in_education

Texas Higher Education Coordinating Board, & Educational Policy Improvement Center. (2009). Texas College and Career Readiness Standards. Austin, TX: Authors.

The College Board. (2006). *Standards for College Success*. New York, NY: Author.

The College Board. (2014). Redesigned SAT. Retrieved September 5, 2014, from https://http://www.collegeboard.org/delivering-opportunity/sat/redesign

Tough, P. (2012). *How children succeed: Grit, curiosity, and the hidden power of character* New York: Houghton Mifflin Harcourt.

Tucker, M. S. (2014). Fixing our national accountability system. Washington, D. C.: National Center on Education and the Economy.

Tyack, D. B. (1974). *The one best system: A history of American urban education*. Cambridge: Harvard University Press.

Tyack, D. B., & Cuban, L. (1995). *Tinkering toward Utopia: A century of public school reform*. Cambridge, MA: Harvard University Press.

U. S. Department of Education. (2001). No Child Left Behind. Washington, D.C.: Author.

Ujifusa, A. (2014). Florida picks Common-Core test from AIR, not PARCC. *Education Week*. Retrieved from Curriculum Matters website: http://blogs.edweek.org/edweek/state_edwatch/2014/03/florida_picks_common-core_test_from_air_not_parcc.html

Wei, R. C., Schultz, S. E., & Pecheone, R. (2012). Performance assessments for learning: The next generation of state assessments. Stanford, California: Stanford Center for Assessment, Learning, and Equity.

About the Author

David Conley
Educational Policy Improvement Center (EPIC)
conley@uoregon.edu
David T. Conley is the founder, chief executive officer, and chief strategy officer of EPIC. Conley also serves as President of CCR Consulting LLC, Professor of Educational Policy and Leadership, and founder and director of the Center for Educational Policy Research (CEPR) at the University of Oregon.

About the Guest Series Editor

Linda Darling-Hammond
Stanford University
ldh@stanford.edu
Linda Darling-Hammond is Charles E. Ducommun Professor of Education at Stanford University where she is Faculty Director of the Stanford Center for Opportunity Policy in Education. Her latest book is *Beyond the Bubble Test: How Performance Assessments Support 21st Century Learning* (Wiley, 2014).

SPECIAL SERIES
A New Paradigm for Educational Accountability:
Accountability for Meaningful Learning

education policy analysis archives

Volume 23 Number 8 February 2nd, 2015 ISSN 1068-2341