Criteria for High-Quality Assessment

By Linda Darling-Hammond, Joan Herman, James Pellegrino, Jamal Abedi, J. Lawrence Aber, Eva Baker, Randy Bennett, Edmund Gordon, Edward Haertel, Kenji Hakuta, Andrew Ho, Robert Lee Linn, P. David Pearson, James Popham, Lauren Resnick, Alan H. Schoenfeld, Richard Shavelson, Lorrie A. Shepard, Lee Shulman, Claude M. Steele

Published by:

Stanford Center for Opportunity Policy in Education,
Stanford University;

Center for Research on Student Standards and Testing,
University of California at Los Angeles; and

Learning Sciences Research Institute,
University of Illinois at Chicago

June 2013

LEARNING SCIENCES
UIC Learning Sciences Research Institute

scope
Stanford Center for Opportunity Policy in Education

CRESST
National Center for Research on Evaluation, Standards, & Student Testing

The authors gratefully acknowledge the support of the Hewlett Foundation for this work.

Suggested citation: Darling-Hammond, L., Herman, J., Pellegrino, J., et al. (2013). *Criteria for high-quality assessment*. Stanford, CA: Stanford Center for Opportunity Policy in Education.

Table of Contents

Abstract

States and school districts across the nation are making critical decisions about student assessments as they move to implement the Common Core State Standards (CCSS), adopted by 45 states. The Standards feature an increased focus on deeper learning, or students' ability to analyze, synthesize, compare, connect, critique, hypothesize, prove, and explain their ideas. States are at different points in the CCSS transitions, but all will be assessing their K–12 students against these higher standards in the 2014–15 school year.

Based on the changing demands of today's workforce, advances in other nations, and original analysis, this report provides a set of criteria for high-quality student assessments. These criteria can be used by assessment developers, policymakers, and educators as they work to create and adopt assessments that promote deeper learning of 21st-century skills that students need to succeed in today's knowledge-based economy.

The five criteria include:

1. **Assessment of Higher-Order Cognitive Skills** that allow students to transfer their learning to new situations and problems.

2. **High-Fidelity Assessment of Critical Abilities** as they will be used in the real world, rather than through artificial proxies. This calls for performances that directly evaluate such skills as oral, written, and multimedia communication; collaboration; research; experimentation; and the use of new technologies.

3. **Assessments that Are Internationally Benchmarked**: Assessments should be evaluated against those of the leading education countries, in terms of the kinds of tasks they present as well as the level of performance they expect.

4. **Use of Items that Are Instructionally Sensitive and Educationally Valuable**: Tests should be designed so that the underlying concepts can be taught and learned, rather than depending mostly on test-taking skills or reflecting students' out-of-school experiences. To support instruction, they should also offer good models for teaching and learning and insights into how students think as well as what they know.

5. **Assessments that Are Valid, Reliable, and Fair** should accurately evaluate students' abilities, appropriately assess the knowledge and skills they intend to measure, be free from bias, and be designed to reduce unnecessary obstacles to performance that could undermine validity. They should also have positive consequences for the quality of instruction and the opportunities available for student learning.

Criteria for High-Quality Assessment

I am calling on our nation's Governors and state education chiefs to develop standards and assessments that don't simply measure whether students can fill in a bubble on a test, but whether they possess 21st-century skills like problem-solving and critical thinking, entrepreneurship and creativity.

— President Barack Obama, March 2009

Responding to President Obama's call, policymakers in nearly every state have adopted new standards intended to ensure that all students graduate from high school ready for college and careers. Achieving that goal will require a transformation in teaching, learning, and assessment so that all students develop the deeper learning competencies that are necessary for postsecondary success.[1]

The changing nature of work and society means that the premium in today's world is not merely on students' acquiring information, but on their ability to analyze, synthesize, and apply what they've learned to address new problems, design solutions, collaborate effectively, and communicate persuasively.[2]

This transformation will require an overhaul in curriculum and assessment systems to support deeper learning competencies. Ministries of education around the world have been redesigning curriculum and assessment systems to emphasize these skills. For example, as Singapore prepared to revamp its assessment system, then Education Minister, Tharman Shanmugaratnam, noted:

> [We need] less dependence on rote learning, repetitive tests and a 'one size fits all' type of instruction, and more on engaged learning, discovery through experiences, differentiated teaching, the learning of life-long skills, and the building of character, so that students can…develop the attributes, mindsets, character and values for future success.[3]

Reforms in Singapore, like those in New Zealand, Hong Kong, a number of Australian states and Canadian provinces, and other high-achieving jurisdictions, have introduced increasingly ambitious performance assessments that require students to find, evaluate, and use information rather than just recalling facts. In addition, these assessments—which call on students to design and conduct investigations, analyze data, draw valid conclusions, and report findings—frequently call on students to demonstrate what they know in investigations that produce sophisticated written, oral, mathematical, physical, and multimedia products.[4] (See Appendix A for examples.) These assessments, along with other investments (in thoughtful curriculum, high-quality teaching, and equitably funded schools, for example) appear to contribute to their high student achievement.[5]

The United States is poised to take a major step in the direction of curriculum and assessments for this kind of deeper learning with the adoption of new Common Core State Standards (CCSS) in more than 40 states. These standards are intended to be "fewer, higher, and deeper" than previous iterations of standards, which have been criticized for being a "mile wide and an inch deep."[6] They aim to ensure that students are prepared for college and careers with deeper knowledge and more transferable skills in these disciplines, including the capacity to read and listen critically for understanding; to write and speak clearly and persuasively, with reference to evidence; and to calculate and communicate mathematically, reason quantitatively, and design solutions to complex problems.

The Common Core Standards will require a more integrated approach to delivering content instruction across all subject areas.[7] The Common Core Standards in English language arts are written to include the development of critical reading, writing, speaking, and listening skills in history, science, mathematics, and the arts, as well as in English class. The Common Core Standards in mathematics are written to include the use of mathematical skills and concepts in fields like science, technology, and engineering. These standards emphasize the ways in which students should use literacy and numeracy skills across the curriculum and in life. As states seek to implement these standards, they must also examine how their assessments support and evaluate these skills and create incentives for them to be well taught.

Two consortia of states—the Partnership for Assessment of Readiness for College and Careers (PARCC) and the Smarter Balanced Assessment Consortium (SBAC)—have been formed to develop next-generation assessments of these standards. As states are increasingly able to work collaboratively on problems of policy and practice, other initiatives, such as the Innovation Lab Network (ILN) of states and districts, coordinated by the Council for Chief State School Officers, are also developing strategies to create more intellectually ambitious assessments that are more internationally comparable.

Undoubtedly, there will be many initiatives to rethink assessments that accompany these reforms. Thus, it is timely to consider what the features of high-quality assessment systems that meet these new goals should include.

The recently released report of the Gordon Commission, written by the nation's leading experts in curriculum, teaching, and assessment, described the most critical objectives this way:

> To be helpful in achieving the learning goals laid out in the Common Core, assessments must fully represent the competencies that the increasingly complex and changing world demands. The best assessments can accelerate the acquisition of these competencies if they guide the actions of teachers and enable students to gauge their progress. To do so, the tasks and activities in the assessments must be models worthy of the

attention and energy of teachers and students. The Commission calls on policy makers at all levels to actively promote this badly needed transformation in current assessment practice...[T]he assessment systems [must] be robust enough to drive the instructional changes required to meet the standards...and provide evidence of student learning useful to teachers.

New assessments must advance competencies that are matched to the era in which we live. Contemporary students must be able to evaluate the validity and relevance of disparate pieces of information and draw conclusions from them. They need to use what they know to make conjectures and seek evidence to test them, come up with new ideas, and contribute productively to their networks, whether on the job or in their communities. As the world grows increasingly complex and interconnected, people need to be able to recognize patterns, make comparisons, resolve contradictions, and understand causes and effects. They need to learn to be comfortable with ambiguity and recognize that perspective shapes information and the meanings we draw from it. At the most general level, the emphasis in our educational systems needs to be on helping individuals make sense out of the world and how to operate effectively within it. Finally, it is also important that assessments do more than document what students are capable of and what they know. To be as useful as possible, assessments should provide clues as to why students think the way they do and how they are learning as well as the reasons for misunderstandings.[8]

What Should High-Quality Assessment Systems Include?

No single assessment can evaluate all of the kinds of learning we value for students, nor can a single instrument meet all of the goals held by parents, practitioners, and policymakers. It is important to envision a coordinated system of assessment, in which different tools are used for different purposes—for example, formative and summative, diagnostic vs. large-scale reporting. Within such systems, however, all assessments should faithfully represent the standards, and all should model good teaching and learning practice.

Five major features define the elements of assessment systems than can fully measure the Common Core State Standards and support the evaluation of deeper learning:

1. Assessment of Higher-Order Cognitive Skills: Most of the tasks students encounter should tap the kinds of cognitive skills that have been characterized as "higher-level"—skills that support transferable learning, rather than emphasizing only skills that tap rote learning and the use of basic procedures. While there is a necessary place for basic skills and procedural knowledge, it must be balanced with attention to critical thinking and applications of knowledge to new contexts.

2. High-Fidelity Assessment of Critical Abilities: In addition to key subject matter concepts, assessments should include the critical abilities articulated in the standards, such as communication (speaking, reading, writing, and listening in multi-media forms), collaboration, modeling, complex problem solving, planning, reflection, and research. Tasks should measure these abilities directly as they will be used in the real world, rather than through a remote proxy.

3. Standards that Are Internationally Benchmarked: The assessments should be as rigorous as those of the leading education countries, in terms of the kind of content and tasks they present, as well as the level of performance they expect.

4. Use of Items that Are Instructionally Sensitive and Educationally Valuable: The tasks should be designed so that the underlying concepts can be taught and learned, rather than reflecting students' differential access to outside-of-school experiences (frequently associated with their socioeconomic status or cultural context) or depending on tricky interpretations that mostly reflect test-taking skills. Preparing for and participating in the assessments should engage students in instructionally valuable activities, and results from the tests should provide instructionally useful information.

5. Assessments that Are Valid, Reliable, and Fair: In order to be truly valid for a wide range of learners, assessments should *measure well* what they purport to measure, *accurately evaluate* students' abilities, and do so *reliably* across testing contexts and scorers. They should also be *unbiased* and *accessible* and used in ways that support positive outcomes for students and instructional quality.

Standard 1: Assessment of Higher-Order Cognitive Skills

As suggested above, the Common Core State Standards, along with the Next Generation Science Standards, call for the development of many more complex skills than those that have been typically assessed in U.S. tests over the past decade. If these are to be developed in classrooms, the assessments should represent the critical skills and abilities that are outlined in the standards, rather than measuring only what is easiest to assess.

In particular, assessments should strike a much more productive balance between evaluating basic skills and those capacities that students can use to *transfer* their learning to novel contexts. As the National Research Council noted in its recent study, *Education for Life and Work*:

> We define "deeper learning" as the process through which an individual becomes capable of taking what was learned in one situation and applying it to new situations (i.e., transfer)...The goals included in the new [Common Core] Standards and the NRC Framework reflect each discipline's desire to promote deeper learning and develop transferable knowledge and skills within that discipline. For example, both the mathematics standards and the science framework include a "practices" dimension, calling for students to actively use and apply—i.e., to transfer—knowledge, and the English language arts standards call on students to synthesize and apply evidence to create and effectively communicate an argument.[9]

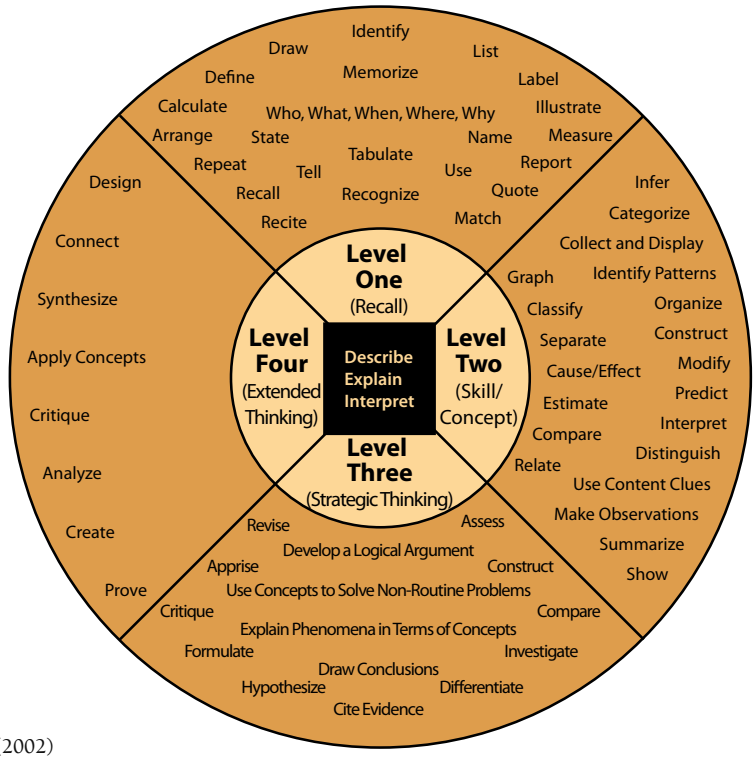
There are many ways to conceptualize the knowledge and skills represented in curriculum, teaching, and assessment. One widely used approach—though by no means the only useful one—is Webb's *Depth of Knowledge* (DOK) taxonomy (see figure 1, page 6).[10] Using the DOK framework as a guide, if assessments are to reflect and encourage transferable abilities, a substantial majority of the items and tasks (at least two-thirds) should tap conceptual knowledge and abilities (level 2, 3, or 4 in the DOK taxonomy). At least one-third of the total in mathematics—and at least half of the total in English language arts—should tap the kinds of cognitive skills that have been characterized as "higher-level," such as the abilities to assess, compare, evaluate, hypothesize, and investigate (level 3), as well as the abilities to analyze, synthesize, design, and create (level 4 in the DOK taxonomy below).

A number of studies have found that current state tests tend not to measure the more intellectually ambitious expectations set out in state standards, settling for recall, recognition, and implementation of procedures more often than analysis, evaluation, and production of ideas.[11]

According to a recent study of tests in 17 states, selected because they were reputed to have higher standards and more ambitious assessments than many others, fewer than 2% of mathematics items and only 21% of English language arts items reached the higher levels (DOK levels 3 and 4). These higher-level skills expect students to hypothesize, critique, analyze, synthesize, compare, connect, prove, or explain their ideas.[12] This study found that the level of cognitive demand was severely constrained by the extent of multiple-choice questions, which were rarely able to assess these higher-order skills.

Another recent study of 19 state tests, using a different method for classifying items, had very similar findings. The researchers found that only 7% of mathematics items and one-third of English language arts (reading) items required students to use higher-order skills of analysis, synthesis, or problem solving. Fully 80% of mathematics items and 52% of reading items tapped only lower-level skills such as memorization, recognition of information, and use of routine procedures.[13]

Figure 1: Depth of Knowledge Levels



Source: N. Webb (2002)

The plans for the new Consortia assessments could increase cognitive expectations by many orders of magnitude. An analysis of the Content Specifications for the Smarter Balanced Assessment Consortium found, for example, that 68% of the targets in English language arts and 70% of those in mathematics intend to tap these higher-level skills.[14] A more qualitative analysis of the item specifications for PARCC found comparable levels of intended intellectual rigor.[15]

This represents very significant initial progress that should be at least matched by any other assessments states select or develop to serve as their next-generation assessments. A reasonable standard would be that at least half of the assessment items and tasks would address higher-order skills. This would also suggest that at least half of the assessment items would call for students to respond in formats that require an elaborated response.

Percentage of Items at Different Levels of Cognitive Demand on 19 State Tests					
	Memorize/ Recognize/ Identify	Implement Procedures	Demonstrate Understanding	Conjecture; Prove/Analyze	Solve Novel Problems / Draw Connections
Mathematics	16%	63%	13%	6%	1%
English Language Arts (Reading)	31%	21%	15%	29%	4%

Standard 2: High-Fidelity Assessment of Critical Abilities

The Common Core State Standards identify a number of areas of knowledge and skills that are clearly so critical for college and career readiness that they should be targeted for inclusion in new assessment systems. As described in the standards, these include:

- **Research:** Conduct sustained research projects to answer a question (including a self-generated question) or solve a problem, narrow or broaden the inquiry when appropriate, and demonstrate understanding of the subject under investigation. Gather relevant information from multiple authoritative print and digital sources, use advanced searches effectively, and assess the strengths and limitations of each source in terms of the specific task, purpose, and audience.

- **Analysis and Synthesis of Information:** Integrate and synthesize multiple sources of information (e.g., texts, experiments, simulations) presented in diverse formats and media (e.g., visually, quantitatively, orally) in order to address a question; make informed decisions; understand a process, phenomenon, or concept; and solve problems while evaluating the credibility and accuracy of each source and noting any discrepancies among the data.

- **Experimentation and Evaluation:** Follow precisely a complex multistep procedure when carrying out experiments, taking measurements, or performing technical tasks; analyze the specific results based on explanations in the text. Evaluate hypotheses, data, analysis, and conclusions, verifying the data when possible and corroborating or challenging conclusions with other sources of information.

- **Communication in Oral, Written, Graphic, and Multi-Media Forms:** Use oral and written communication skills to learn, evaluate, and express ideas for a range of tasks, purposes, and audiences. Develop and strengthen writing as needed by planning, revising, editing, and rewriting while considering the audience. Present information, findings, and supporting evidence, making strategic use of digital media and visual displays to enhance understanding. Use technology, including the Internet, to research, produce, publish, and update individual or shared products in response to ongoing feedback, including new arguments or information.

- **Collaboration and Interpersonal Interaction:** Develop a range of interpersonal skills, including the ability to work with others and to participate effectively in a range of conversations and collaborations.

- **Modeling, Design, and Complex Problem Solving:** Use quantitative reasoning to solve problems arising in everyday life, society, and the workplace—e.g., to plan a school event or analyze a problem in the community, to solve a design problem, or to examine relationships among quantities of interest. Plan solution pathways, monitoring and evaluating progress and changing course if necessary, and find relevant external resources, such as experimental and modeling tools, to solve problems. Interpret and evaluate results in the context of the situation and improve the model or design as needed.

If the skills discussed above are to be well measured, it will be important that the tasks students are asked to tackle directly measure these complex skills, rather than evaluating a remote proxy for the skills. For example, where the standards ask that students conduct research; find, evaluate, and synthesize information; weigh and balance evidence; and communicate a logical argument that is well defended with evidence, assessments should call for the demonstration of these skills in authentic tasks.

While it is possible to argue that the ability to locate evidence for an argument could be demonstrated by a student examining an already-provided text and pointing out where evidence for a particular claim could be found, this will not be sufficient to demonstrate that the student knows how to conduct research by finding, evaluating, and using appropriate evidence to build an argument or evaluate a situation. Thus, one of the sample performance tasks in the planned SBAC assessment requires that students take up a social scientific topic about which there are different views, engage in a Google search to find and weigh scientific and historical evidence, evaluate the credibility of the evidence, and develop a cogent essay that takes a position on that topic. Students are also expected to revise their work before it is final.

Similarly, while it is possible to ask students to answer multiple-choice questions that identify possible corrections to errors that have been identified in a text, this does not demonstrate whether the student could independently write or revise a text. New assessments should require students to revise their own written texts, as well as to revise a selection that is provided for them.

And while it is possible to ask students to select an answer to a mathematical problem that is given in a familiar format, this will not demonstrate whether the student could take a real-world problem and identify the kind of mathematics needed to solve it, develop and apply their own solution strategy, select and use appropriate tools, and explain their conclusions. This requires tasks that present the problem in its real-world form and allow students to figure out how to approach it and show their thinking.

As Webb (2005) suggests in illustrating how skills can be displayed and assessed at each Depth of Knowledge level, the activities needed to develop and evaluate these skills become more complex at deeper levels. (See figure 2, page 9.)

Figure 2: Web Alignment Tool

Level One Activities	Level Two Activities	Level Three Activities	Level Four Activities
Recall elements and details of story structure, such as sequence of events, character, plot and setting. Conduct basic mathematical calculations. Label locations on a map. Represent in words or diagrams a scientific concept or relationship. Perform routine procedures like measuring length or using punctuation marks correctly. Describe the features of a place or people.	Identify and summarize the major events in a narrative. Use context cues to identify the meaning of unfamiliar words. Solve routine multiple-step problems. Describe the cause/effect of a particular event. Identify patterns in events or behavior. Formulate a routine problem given data and conditions. Organize, represent, and interpret data.	Support ideas with details and examples. Use voice appropriate to the purpose and audience. Identify research questions and design investigations for a scientific problem. Develop a scientific model for a complex situation. Determine the author's purpose and describe how it affects the interpretation of a reading selection. Apply a concept in other contexts.	Conduct a project that requires specifying a problem, designing and conducting an experiment, analyzing its data, and reporting results/solutions. Apply mathematical model to illuminate a problem or situation. Analyze and synthesize information from multiple sources. Describe and illustrate how common themes are found across texts from different cultures. Design a mathematical model to inform and solve a practical or abstract situation.

Webb, Norman L. and others. "Web Alignment Tool" 24 July 2005. Wisconsin Center of Educational Research. University of Wisconsin-Madison. 2 Feb. 2006. (*www.wcer.wisc.edu/WAT/index.espx*)

The Consortium assessments include items and tasks that will measure some of these key attributes—including skills like listening and writing with revision, as well as mathematical modeling and applied problem solving—that have been neglected in most state tests over the last decade. However, the initial versions of the Consortia tests will not include long-term research and investigation tasks or assessments of multi-modal communications, such as spoken, visual, and technology-supported presentations. Some of these aspects of the standards will likely be tackled in later versions of one or both of the Consortia's tests.

States and districts should include these capacities in other aspects of their assessment systems, as many did during the 1990s, when states like Connecticut and Vermont had students design and conduct scientific experiments, often collaboratively, and analyze and present their results; Kentucky and Vermont engaged in writing portfolios that required students to plan, write, and revise extended pieces of work; and Wyoming and Wisconsin created profiles of students' learning through sets of performance tasks.

There are also a number of networks of high schools that engage students in demonstrating these abilities through structured performance assessments for graduation. These include public school networks like the Performance Standards Consortium in New York City, the International High Schools Network, New Tech High Schools, and Envision Schools, all of which require graduation portfolios that demonstrate research, presentation, and technology skills developed to a high standard.

New statewide efforts to create assessments that will evaluate these more complex abilities are underway in the U.S. through the Council for Chief State School Officers' Innovation Lab Network. This group of 10 states is planning to augment the Consortium assessments with more extended performance tasks that replicate, to the extent possible, the ways in which these kinds of abilities will be used in college and career contexts. That such initiatives are not only desirable but feasible has been demonstrated both in U.S. states and many countries abroad, which include such tasks routinely, at scale, as part of their examination systems.[16]

Standard 3: Standards that Are Internationally Benchmarked

The assessments should be as rigorous as those of the leading education countries, in terms of the kinds of tasks they present, as well as the level of performance they expect.

On the Program in International Student Assessments (PISA) tests, assessments typically require constructed responses to questions that require analysis and applications of knowledge to novel problems or contexts. From Finland to Singapore and Australia to New Zealand, students write even more extended responses to questions that require them to evaluate and analyze texts, data, and problems, rather than bubbling in responses to multiple-choice questions.

In addition to open-ended, "sit-down" tests, project components are now used in the examination systems of Hong Kong; Queensland and Victoria, Australia; New Zealand; and Singapore, as well as the International Baccalaureate program, which is used in more than 100 countries around the world as an arbiter of international educational standards.[17] These projects require students to investigate problems and design solutions, conduct research, analyze data, write extended papers, and deliver oral presentations describing their results. Some of the tasks also include collaboration among students in both the investigations and the presentations.[18] These assessments are scored by teachers, who are supported with moderation and auditing systems to ensure consistency, and are included in overall examination results. (See Appendix A for examples.)

Jurisdictions that use such assessments understand that problem-solving abilities are growing increasingly important in the modern economy, as indicated by shifts in the most valued skills identified by Fortune 500 companies. (See figure 3, page 11.) In 1970, companies were calling for reading, writing, and arithmetic.

However, by the turn of the century, the top five attributes these companies sought were, in order of importance: 1) teamwork, 2) problem solving, 3) interpersonal skills, 4) oral communications, and 5) listening skills. Thus, it will ultimately be important for high-quality assessment systems to include these abilities, setting performance standards that are comparable to those in high-achieving nations around the world.

Figure 3: Fortune 500 Most Valued Skills [19]

	1970	1999
1	Writing	***Teamwork***
2	Computational Skills	***Problem Solving***
3	Reading Skills	***Interpersonal Skills***
4	Oral Communications	Oral Communications
5	Listening Skills	Listening Skills
6	Personal Career Development	Personal Career Development
7	Creative Thinking	Creative Thinking
8	Leadership	Leadership
9	Goal Setting/Motivation	Goal Setting/Motivation
10	***Teamwork***	Writing

Sources: Cassel & Kolstad (1999); Creativity in Action *(1990).*

To underscore the worldwide responses to these new realities, in 2015, PISA will add assessment of collaborative problem solving to its assessments of reading, mathematics, and scientific literacy. Assessment of computer literacy is on the horizon as well.

If the United States wants to be internationally competitive with respect to preparing students for 21st-century occupations, its assessment strategies will also need to focus more explicitly on developing students' capacities to think, problem solve, collaborate, and communicate in many forms and using a wide range of technologies.

Standard 4: Use of Items that Are Instructionally Sensitive and Educationally Valuable

Assessment tasks should also be *instructionally sensitive* and *educationally useful.* That is, they should 1) represent the curriculum content in ways that respond to instruction, and 2) have value for guiding and informing teaching.

Instructionally sensitive items are designed so that the underlying concepts can be taught and learned, rather than reflecting students' differential access to outside-of-school experiences (frequently associated with their socioeconomic status or cultural context) or depending mostly on test-taking skills.[20] Although test-taking skills can be taught,

it is not a good use of valuable instructional time to spend hours teaching students to "psych out" the tests rather than to develop the critically important skills they will need to use in the real world.

Unfortunately, researchers have found that some of the current generation of state basic skills tests are instructionally insensitive.[21] As James Popham observed, these tests may fail to reveal how well a group of students has been taught, because they measure what students bring to school, but not what they learn from school.[22] This is particularly problematic when judgments about the success of individuals and groups of teachers and administrators are based on students' test performance. Only through the use of instructionally sensitive items—that is, items capable of distinguishing between students who have been taught relatively more or less well—can valid inferences about educators' contributions be drawn.

Studies have also found that, even when instruction sometime registers on tests like these, students have had to learn the standards in ways that replicate the format of the items on the tests, failing to generalize to the content domain, and thus under-representing the intent of the standards.[23]

However, there are ways to both test and improve instructional sensitivity.[24] Among these, assessments should be constructed so that an appropriate response requires the student to employ key, enabling knowledge and skills. The knowledge and/or skills represented by the test are clearly described so that teachers will have an understanding of the cognitive demands required for students' successful performance.

In addition, some researchers have found that "performance-based testing maximize[s] sensitivity by better integrating instruction and assessment."[25] Because performance-based testing is more congruent with the ways in which skills are taught and used, it can more easily measure the effects of instruction. Furthermore, such assessments can both guide and inform teaching in educationally valuable ways.[26]

In many countries, assessments *of, as*, and *for* learning are a goal.[27] Assessments are intended not only to measure learning, but also to improve learning by offering good models for teaching and learning and insights into how students think and what they know. Evidence shows that assessments that provide these insights, used to guide instruction and revision of work, can be powerful influences on learning and achievement.[28] As described below, this is an aspect of consequential validity.

Standard 5: Assessments that Are Valid, Reliable, and Fair

All large-scale assessments are expected to meet standards of validity, reliability, and fairness. Associated with these standards are expectations for accuracy and access.[29] Assessments can only be validated in relation to the specific purposes they are designed to serve. This requires the assembly of evidence about several kinds of validity claims.

To be valid for any purpose, an assessment should be a good representation of the knowledge and skills it intends to measure. This premise is generally captured by the overarching idea of *construct validity. Construct validity evidence* can be of several kinds. We have argued above that this kind of validity requires authentic representations of the knowledge and skills an assessment purports to measure using high-fidelity tasks. Other evidence may examine *content relationships* (whether there is a demonstrable relationship between the content specifications intended to be evaluated and the set of items and tasks on the test) and *concurrent relationships* (whether scores are reasonably related to those on other previously validated measures).

Predictive evidence is a form of construct validation that examines whether performance on an assessment is strongly related to real-world success in the domain that the assessment is meant to reflect. For example, to what extent does performance on a college admissions test predict a student's actual success in college? Does a cut score on a particular test used to allow students to take credit-bearing courses in college accurately predict students' abilities to succeed in such courses? And so on. This is particularly important for the assessment of new standards that are intended to lead to and reflect college- and career-readiness.

Consequential evidence refers to the kinds of consequences an assessment and its uses have for learners and for instruction. As Herman and Choi note: "Results should be useful and used for intended purposes and not carry serious unintended or negative consequences."[30] Assessments can positively influence instruction through their diagnostic value, as well as by communicating important learning goals and modeling appropriate pedagogy. They can guide helpful interventions and teaching decisions. However, assessments can also have negative consequences if they are designed or used in ways that distort teaching, deny students access to learning opportunities from which they could benefit, or create incentives for schools to underserve or exclude students with particular needs. Thus, both the assessments themselves, and the decisions related to their interpretation and use, must be subjected to scrutiny.

In order to have assessments that are truly valid for a wide range of learners, they should also be demonstrably *accurate* in evaluating students' abilities and do so *reliably* across testing contexts and scorers. They should also be *fair* and *accessible*: They should be free from bias and designed to reduce construct-irrelevant obstacles to performance

that otherwise would undermine validity for some subgroups (for example, language complexities not related to the construct being measured that impede the performance of English learners). Use of the principles of universal design, together with the design of accommodations and modifications, should create maximum access to the assessment for a wide range of learners. And they should sufficiently cover the continuum of achievement that they enable a wide range of students to show what they know and how they've progressed. Finally, they should be transparent enough to support opportunities to learn relative to the expected content and cognitive demands.

Conclusion

If schools are to enable the kind of transferable learning described in the Common Core State Standards and required of young people in contemporary society, assessments will need to support curriculum and teaching focused on such learning, along with traditional basic skills. New assessment systems, grounded in new standards, should include the features we have described here.

We recognize that our criteria for assessment systems should be rigorous and ambitious, while taking account in the near term of what should be achievable financially, logistically, technologically, and scientifically. The path to reaching more ambitious goals is likely to traverse distinct phases rather than occur in one giant leap. Given where we are today and what should be feasible in the near term, we suggest a set of indicators that can be used to evaluate the criteria. (See page 15.)

States should evaluate the set of assessments they select and develop against these standards, and should use them in ways for which they have been appropriately validated and in contexts that ensure positive consequences for students and for instruction.

INDICATORS OF QUALITY IN A SYSTEM OF NEXT GENERATION ASSESSMENTS

1) Assessment of Higher-Order Cognitive Skills

√ A large majority of items and tasks (at least two-thirds) evaluate the conceptual knowledge and applied abilities that support transfer (e.g., Depth of Knowledge levels 2, 3, or 4 in Webb's Taxonomy or the equivalent).

√ At least one-third of the assessment content in mathematics, and at least one-half in English language arts, evaluate higher-order skills that allow students to become independent thinkers and learners (DOK levels 3 or 4).

2) High-Fidelity Assessment of Critical Abilities

Critical abilities outlined in the standards are evaluated using high-fidelity tasks that use the skills in authentic applications:

√ Research, including analysis and synthesis of information

√ Experimentation and evaluation

√ Oral communications—speaking and listening

√ Written communications—reading and writing

√ Use of technology for accessing, analyzing, and communicating information

√ Collaboration

√ Modeling, design, and problem solving using quantitative tools

3) Standards that Are Internationally Benchmarked

√ Calibration to PISA, International Baccalaureate, or other internationally comparable assessment (based on evaluation of content comparability, performance standards, and analysis of student performance on embedded items)

4) Items that Are Instructionally Sensitive and Educationally Valuable

√ Research that confirms instructional sensitivity

√ Rich feedback on student learning and performance

√ Tasks that reflect and can guide valuable instructional activities

5) Assessments that Are Valid, Reliable, and Fair

√ Evidence that the intended knowledge and skills are well measured

√ Evidence that scores are related to the abilities they are meant to predict

√ Evidence that the assessments are well-designed and valid for each intended use—and that uses are appropriate to the test purposes and validity evidence

√ Evidence that the assessments are unbiased and fairly measure the knowledge and skills of students from different language, cultural, and income backgrounds, as well as students with learning differences

√ Evidence that the assessments measure students' learning accurately along a continuum of achievement, consistent with the purposes the assessments are intended to serve

Appendix A: Assessments Around the World

PROJECT WORK IN SINGAPORE

In Singapore, **Project Work** (PW) is an assessment that is compulsory for all pre-university students. There is dedicated curriculum time for students to carry out their collaborative interdisciplinary project tasks over an extended period. The assessment tasks, which are set by the Singapore Examinations and Assessment Board (SEAB), are designed to be sufficiently broad to allow students to carry out a project that they are interested in while meeting the task requirements.

In groups formed by the teacher, students agree on the project that the group will undertake, brainstorm and evaluate each other's ideas, and decide on how the work should be allocated. Project Work tasks result in:

- a **Written Report** that shows evidence of the group's ability to generate, analyze, and evaluate ideas for the project;
- an **Oral Presentation** in which each individual group member is assessed on his/her fluency and clarity of speech, awareness of audience, as well as response to questions. The group as a whole is also assessed in terms of the effectiveness of the overall presentation;
- a **Group Project File** in which each individual group member submits three documents related to "snapshots" of the processes involved in carrying out the project. These documents show the individual student's ability to generate, analyze, and evaluate (I) preliminary ideas for a project, (II) a piece of research material gathered for the chosen project, and (III) insights and reflections on the project.

The SEAB specifies task-setting, conditions, assessment criteria, achievement standards, and marking processes. Classroom teachers carry out the assessment of all three components of PW using the assessment criteria provided by the Board. All schools are given exemplar material that illustrates the expected marking standards. The Board provides training for assessors and internal moderators. Like all other assessments, the grading is both internally and externally moderated to ensure consistency in scoring.

In carrying out the PW assessment task, students are intended to acquire self-directed inquiry skills as they propose their own topic, plan their timelines, allocate individual areas of work, interact with teammates of different abilities and personalities, and gather and evaluate primary and secondary research material. These PW processes reflect life skills and competencies such as knowledge application, collaboration, communication, and independent learning, which prepare students for the future workplace.

EXTENDED EXPERIMENTAL INVESTIGATIONS IN QUEENSLAND

In Queensland (Australia) science courses, like those in Singapore, Hong Kong, and other Australian states, students must complete an extended experimental investigation that they design, conduct, and evaluate. In Queensland, the task is defined as follows:

> Within this category, instruments are developed to investigate a hypothesis or to answer a practical research question. The focus is on planning the extended experimental investigation, problem solving, and analysis of primary data generated through experimentation by the student. Experiments may be laboratory- or field-based. An extended experimental investigation may last from four weeks to the entirety of the unit of work. The outcome of an extended experimental investigation is a written scientific report. For monitoring, the discussion/conclusions/evaluation/recommendations of the report should be between 1500 and 2000 words.

To complete such an investigation the student must:

- develop a planned course of action;
- clearly articulate the hypothesis or research question, providing a statement of purpose for the investigation;
- provide descriptions of the experiment;
- show evidence of modification or student design;
- provide evidence of primary and secondary data collection and selection;
- execute the experiment(s);
- analyze data;
- discuss the outcomes of the experiment;
- evaluate and justify conclusion(s); and
- present relevant information in a scientific report.

GRADUATE CERTIFICATE IN SECONDARY EDUCATION (GCSE) TASK IN INTERACTIVE COMPUTER TECHNOLOGY, ENGLAND

In England, students choose a number of domains in which to be examined as part of the high school assessment system. Most of these examinations, which are linked to high school courses, include a project-based component that typically counts for 60% of the total examination score. The project below has been used as part of the Interactive Computer Technology examination.

> Litchfield Promotions works with over 40 bands and artists to promote their music and put on performances in England. The number of bands they have on their books is gradually expanding. Litchfield Promotions needs to be sure that each performance will make enough money to cover all the staffing costs and overhead, as well as make a profit. Many people need to be paid: the bands, sound engineers, and lighting technicians. There is also the cost of hiring the venue. Litchfield Promotions needs to create an ICT solution to ensure that it is all necessary information and that it is kept up to date. Its solution will show income, outgoings, and profit.
>
> Candidates need to: 1) Work with others to plan and carry out research to investigate how similar companies have produced a solution. The company does not necessarily have to work with bands and artists or be a promotions company. 2) Clearly record and display your findings. 3) Recommend a solution that will address the requirements of the task. 4) Produce a design brief, incorporating timescales, purpose, and target audience.

Produce a solution, ensuring that the following are addressed:

1. It can be modified for use in a variety of situations.
2. It has a user-friendly interface.
3. It is suitable for the target audience.
4. It has been fully tested.

You will need to:

1. Incorporate a range of: software features, macros, modeling, and validation checks—used appropriately.
2. Obtain user feedback.
3. Identify areas that require improvement, recommending improvement with justification.
4. Present information as an integrated document.
5. Evaluate your own and others' work.

Endnotes

[1] National Research Council. (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century.* Washington DC: National Research Council.

[2] Baker, E. L. (2008). *Measuring 21st century skills*. Invited paper presented at the Universidad Complutense de Madrid, Madrid, Spain.

[3] Ng, P. T. (2008). Educational reform in Singapore: From quantity to quality. *Education Research on Policy and Practice, 7*, 5–15.

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

[5] Darling-Hammond, L. (2010). *The flat world and education: How America's commitment to equity will determine our future*. New York: Teachers College Press.

[6] Schmidt, W. (2004). *Mathematics and science initiative*. Retrieved January 1, 2013, from http://www2.ed.gov/rschstat/research/progs/mathscience/schmidt.html

[7] For an analysis of the Common Core State Standards and the Next Generation Science Standards, see National Research Council (2012).

[8] The Gordon Commission on Future Assessment in Education. (2013). *A public policy statement*. Princeton, NJ: Educational Testing Service, p. 7.

[9] NRC. (2012). 4-5.

[10] Webb, N.L. (2002). *Depth-of-knowledge levels for four content areas*. Retrieved August 24, 2011, from http://facstaff.wcer.wisc.edu/normw/All%20content%20areas%20%20DOK%20levels%2032802.doc

[11] See for example, Resnick, L. B., Rothman, R., Slattery, J. B., & Vranek, J. L. (2003). Benchmark and alignment of standards and testing. *Educational Assessment, 9*(1&2), 1–27.; Webb, N. L. (2002). *Alignment study of language arts, mathematics, science, and social studies of state standards and assessments in four states*. Washington, DC: Council of Chief State School Officers; Wixson, K. K., Fisk, M. C., Dutro, E., & McDaniel, J. (2002). *The alignment of state standards and assessments in elementary reading.* Ann Arbor, MI: Center for the Improvement of Early Reading Achievement.

[12] Yuan, K., & Le, V. (2012). *Estimating the percentage of students who were tested on cognitively demanding items through the state achievement tests.* Santa Monica, CA: RAND Corporation.

[13] Polikoff, M.S., Porter, A.C., & Smithson, J. (2011). How well aligned are state assessments of student achievement with state content standards? *American Educational Research Journal, 48*(4), 965–995.

[14] Herman, J. L. & Linn, R. L. (2013). On the road to assessing deeper learning: The status of Smarter Balanced and PARCC assessment consortia (*CRESST Report 823*). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

[15] Herman, J. L. & Linn, R. L. (2013).

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

[17] It is important to note that these different assessments are used for different purposes. PISA is a cross-national assessment that samples students and does not report individual scores. The International Bac calaureate, like the A-level examinations in some countries, is designed for an elite, college-going population of students. Finally, some state or national systems are used for virtually all students (e.g., the combined qualifications examinations systems in Queensland and Singapore). These are arguably the most relevant to the assessment context discussed here.

[18] Darling-Hammond, L., & Wentworth, L. (2010).

[19] Cassel, R. N., and Kolstad, R. (1999). The critical job-skills requirements for the 21st century: Living and working with people. *J. Instructional Psychology, 25*(3), 176-180; Creativity in Action (1990). *Skills desired by Fortune 500 companies (in order of importance)*. Buffalo, NY: Creative Education Foundation.

[20] Baker, E. L., O'Neil, H. F., Jr., & Linn, R. L. (1993). Policy and validity prospects for performance-based assessment. *American Psychologist, 48*(12), 1210-1218; Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15-21.

[21] Pham, V. H. (2009). Computer modeling of the instructionally insensitive nature of the Texas Assessment of Knowledge and Skills (TAKS) exam. PhD Dissertation, the University of Texas at Austin; Popham, W. J. (2007a). Accountability tests' instructional insensitivity: The time bomb ticketh. *Education Week*. Retrieved, January 1, 2013, from http://www.edweek.org/ew/articles/2007/11/14/12popham.h27.html

[22] Popham, W. J. (2010). Instructional sensitivity. In W. J. Popham (Ed.), *Everything school leaders need to know about assessment*. Thousand Oaks, CA: Sage.

[23] Ibid.

[24] Methods for identifying insensitive items and tasks so that they can be eliminated are reviewed in Popham, W. J., & Ryan, J. P. (2012). *Methods for identifying a high-stakes test's instructional sensitivity.* A paper presented at the annual meeting of the National Council on Measurement in Education, April 12–16, 2012. Vancouver, B.C., Canada.

[25] D'Agostino, J. V., Welsh, M. E., & Corson, N. M. (2007). Instructional sensitivity of a state's standards-based assessment. *Educational Assessment*, *12*(1), 1–22,; Yoon, B., & Resnick, L. B. (1998). Instructional validity, opportunity to learn and equity: New standards examinations for the California mathematics renaissance (*CSE Tech. Rep. No. 484*). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

[26] Darling-Hammond, L., & Adamson, F. (2010).

[27] Ibid.

[28] Black, P., & William, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80(2)*, 139-148. Retrieved from: www.pdkintl.org /kappan/kbla9810.htm

[29] AERA, APA, & NCME. (1999). *The standards for educational and psychological testing.*Washington, DC: AERA Publications.

[30] Herman, J. L., & Choi, K. (2012). *Validation of ELA and mathematics assessments: A general approach,* Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Author Biographies

Jamal Abedi, *Professor of Education, University of California, Davis*. Specializing in educational and psychological assessments, Abedi's research focuses on testing for English language learners. From 2010 to present, Abedi is a member of the Technical Advisory Committee of the SMARTER Balanced Assessment Consortium. Before then, he served on the expert panel of the U.S. Department of Education, LEP Partnership and as co-founder and chair of AERA's Special Interest Group on Inclusion & Accommodation in Large-Scale Assessment. In 2003, he received the AERA Award for: Outstanding Contribution Relating Research to Practice; in 2008, the California Educational Research Association for Lifetime Achievement Award; and in 2013, the Outstanding Contribution to Educational Assessment Award from The National Association of Test Directors.

J. Lawrence Aber, *Distinguished Professor of Applied Psychology and Public Policy, New York University*. Lawrence Aber is an internationally recognized expert whose basic research examines the influence of poverty and violence, at the family and community levels, on the social, emotional, behavioral, cognitive, and academic development of children and youth. In 2006, Dr. Aber was appointed by the Mayor of New York City to the Commission for Economic Opportunity, an initiative to help reduce poverty and increase economic opportunity in New York City. He is also Chair of the Board of Directors of the Children's Institute, University of Cape Town, South Africa; and serves as consultant to the World Bank on their new project, "Children and Youth in Crisis."

Eva Baker, *Distinguished Research Professor; Director, Center for the Study of Evaluation; Co-Director, Center for Research on Evaluation, Standards, and Student Testing; Director, Center for Advanced Technology in Schools, UCLA*. Dr. Baker's research focuses on the integration of instruction and measurement, including design and empirical validation and feasibility of complex human performance, particularly in technology. She was the President of the World Education Research Association, President of the American Educational Research Association, President of the Educational Psychology Division of the American Psychological Association and is a member of the National Academy of Education. Dr. Baker co-chaired the committee that produced Standards for Educational and Psychological Testing published in 1999 and was Chair of the Board on Testing and Assessment and well as a Committee member on other NRC panels.

Randy Bennett, *Norman O. Frederiksen Chair in Assessment Innovation in the Research & Development Division, Educational Testing Service*. Since the 1980s, Dr. Bennett has conducted research on integrating advances in cognitive science, technology, and measurement to create new approaches to assessment. He was given the ETS Senior Scientist Award in 1996 and the ETS Career Achievement Award in 2005. From 1999 through 2005, Bennett directed the NAEP Technology Based Assessment project, which explored the use of computerized testing in NAEP, conducting the first nationally representative administrations of performance assessments on computer. Since 2007, Bennett has directed an integrated research initiative attempting to create a balanced system of K-12 assessment that provides accountability information and supports classroom learning.

Linda Darling-Hammond, *Charles E. Ducommun Professor of Education at Stanford University & Faculty director of the Stanford Center for Opportunity Policy in Education (SCOPE)*. Linda Darling-Hammond is a former president of the American Educational Research Association and member of the National Academy of Education. Her research, teaching, and policy work focus on issues of school restructuring, teacher quality, and educational equity. She has been deeply engaged in developing and researching performance assessments for both students and teachers, and she serves as a Technical Advisor to the Smarter Balanced Assessment Consortium.

Edmund Gordon, *John M. Musser Professor of Psychology, Emeritus at Yale University & Richard March Hoe Professor, Emeritus of Psychology and Education and Director of the Institute of Urban and Minority Education (IUME) at Teachers College, Columbia University*. Professor Gordon is concerned with issues associated with increasing the number of high academic achieving students who come from African American, Latino, and Native American families. Currently, Professor Gordon is the Senior Scholar and Advisor to the President of the College Board where he developed and co-chaired the Taskforce on Minority High Achievement. He has been elected as Fellow of the American Psychological Association (among others) and was elected to membership in the National Academy of Education in 1968.

Edward Haertel, *Jacks Family Professor of Education, Emeritus, Stanford University*. Dr. Haertel is an expert in the area of educational testing and assessment, focusing on test-based accountability and related policy uses of test data. His recent work has examined standard setting methods, limitations of value-added models for teacher and school accountability, impacts of testing on curriculum, students, and educational policy, test reliability, and generalizability theory. Dr. Haertel serves on the California Advisory Committee for the Public Schools Accountability Act of 1999 and is Chair of the Subcommittee on the Academic Performance Index. He is also chair of the National Research Council's Board on Testing and Assessment (BOTA) and the Vice-President for Programs of the National Academy of Education.

Kenji Hakuta, *Lee L. Jacks Professor of Education, Stanford University*. Kenji Hakuta is an experimental psycholinguist by training, best known for his work in the areas of bilingualism and the acquisition of English in immigrant students. He is the co-chair of Understanding Language, an initiative focusing on the role of language in subject-area learning, with a special focus on helping English Language Learners meet the new Common Core State Standards and Next Generation Science Standards. He has chaired a National Academy of Sciences report Improving Schooling for Language-Minority Children (National Academy Press) and testified to Congress and other public bodies on many topics, including language policy, the education of language minority students, affirmative action in higher education, and improvement of quality in educational research.

Joan Herman, *Senior Scientist and former director, National Center for Research on Evaluation, Standards, and Student Testing (CRESST), UCLA*. A former teacher and school board member, Dr. Herman is past president of the California Educational Research Association; has held a variety of leadership positions in the American Educational Research Association, National Organization of Research Centers, and Knowledge Alliance; and is a frequent contributor at the National Academy of Science's National Research Council. Dr. Herman is current editor of Educational Assessment, serves on the Joint Committee for the Revision of Standards for Educational and Psychological Testing, and chaired the Board of Education for Para Los Niños.

Andrew Ho, *Associate Professor, Harvard Graduate School of Education*. Andrew Ho is a psychometrician interested in educational accountability metrics. He has studied the consequences of "proficiency"-based accountability metrics, the validation of high stakes test score trends with low stakes comparisons, and the potential for alternative accountability structures—like "growth models" and "index systems"—to improve school- and classroom-level incentives. Dr. Ho has his Ph.D. in Educational Psychology and his M.S. in Statistics from Stanford University. He has been a postdoctoral fellow at the National Academy of Education and Spencer Foundation, received the Jason Millman Promising Measurement Scholar Award from the National Council on Measurement in Education, and is a member of the National Assessment Governing Board.

Robert Lee Linn, *Distinguished Professor Emeritus of Education, University of Colorado at Boulder*. Robert Linn's research explores the uses and interpretations of educational assessments, with an emphasis on educational accountability systems. He has received the National Council on Measurement in Education (NCME) Career Award and the American Educational Research Association (AERA) Award for Distinguished Contributions to Educational Research. Dr. Linn is a member of the National Academy of Education (NAEd) and a Lifetime National Associate of the National Academies. He is a past president of the National Council on Measurement in Education (NCME), a past president of the American Educational Research Association (AERA), a past editor of the Journal of Educational Measurement and served as chair of the National Research Council's (NRC) Board on Testing and Assessment.

P. David Pearson, *Professor of Language and Literacy and Human Development, University of California, Berkeley*. P. David Pearson's current research focuses on reading, writing, and language as tools to foster the development of knowledge and inquiry in science. His assessment work includes the design of state assessments in Illinois and Minnesota, directing the ELA performance assessment for the New Standards Project, and conducting validity studies of the National Assessment of Educational Progress as a member of the NAEP Validity Studies Panel. In 1989, the NRC awarded him the Oscar Causey Award for contributions to reading research and AERA presented him Distinguished Contributions to Research in Education Award in 2010. He is the founding editor of the Handbook of Reading Research and served as Dean of the Graduate School of Education at the University of California, Berkeley, from 2001-2010.

James Pellegrino, *Liberal Arts and Sciences Distinguished Professor and Distinguished Professor of Education, Co-Director, Learning Sciences Research Institute (LSRI), University of Illinois at Chicago*. Jim Pellegrino uniquely blends expertise in cognitive science, psychometrics, educational technology, instructional practice, and educational policy. He has served as head of several National Academy of Science/National Research Council study committees and currently co-chairs the NRC/NAS Study Committee on Developing Assessments of Science Proficiency in K-12. Pellegrino is an AERA fellow, a lifetime National Associate of the National Academy of Sciences, and a past member of the Board on Testing and Assessment of the National Research Council. In 2007, he was elected to lifetime membership in the National Academy of Education.

James Popham, *Professor emeritus at the Graduate School of Education and Information Studies, UCLA*. W. James Popham has taught courses for nearly 30 years, mainly on instructional methods for prospective teachers and on evaluation and measurement. In January 2000, he was recognized by *UCLA Today* as one of UCLA's top 20 professors of the 20th century. In 1978, Dr. Popham was elected to the presidency of the American Educational Research Association (AERA). He was also the founding editor of *Educational Evaluation and Policy Analysis*. In 2002, the National Council on Measurement in Education presented him with its Award for Career Contributions to Educational Measurement. In 2009, he was appointed to be a board member of the National Assessment Governing Board.

Lauren Resnick, *Distinguished University Professor, Co-Director, Institute for Learning, University of Pittsburgh*. Lauren Resnick is a senior scientist on at the Institute for Learning (IFL) and the Pittsburgh Science of Learning Center (PSLC). She is a Lifetime National Associate of the National Academies of Science, Engineering and Medicine, a Phi Beta Kappa Visiting Scholar (1993-1994), and the Founding Editor of Cognition and Instruction (1982-1993). She received the Oeuvre Award for Outstanding Contributions to the Science of Learning and Instruction in 1999, and the Edward L. Thorndike Award for Distinguished Psychological Contributions to Education from the American Psychological Association in 1998.

Alan H. Schoenfeld, *Elizabeth and Edward Conner Chair in Education & Professor of Mathematics (by Affiliation), University of California, Berkeley*. Alan H. Schoenfeld's research deals with thinking, teaching, and learning, with an emphasis on mathematics and on issues of equity and diversity in mathematics education. Schoenfeld has served as President of the American Educational Research Association and as Vice President of the National Academy of Education. He is a Fellow of the American Association for the Advancement of Science, a Laureate of Kappa Delta Pi, and served as a senior advisor to the Educational Human Resources Directorate of the National Science Foundation. Schoenfeld was lead author for grades 9-12 of the National Council of Teachers of Mathematics' Principles and Standards for School Mathematics.

Richard Shavelson, *Margaret Jacks Professor Emeritus of Education & Professor of Psychology (by courtesy), Stanford University*. Shavelson's current work focuses on the design of assessments and assessment systems that measure college students learning, both their development of competence/achievement and so-called "soft-skills" such as perspective taking. He co-created the Collegiate Learning Assessment with Steve Klein and built statistical models for estimating value added for the CLA and other college-level assessments. He was the I. James Quillan Dean of Stanford's School of Education from 1995-2000. He is a board member at the Spencer Foundation and the BSCS (formerly Biological Sciences Curriculum Study) and chairs the Education Advisory Council at NatureBridge.

Lorrie A. Shepard, *Dean for the School of Education and University Distinguished Professor, University of Colorado at Boulder*. Shepard's research focuses on psychometrics and the use and misuse of tests in educational settings. In the field of educational measurement, she has made contributions to validity theory, standard setting, and statistical models for detecting test bias. Her studies evaluating test use have addressed the identification of learning disabilities, readiness screening for kindergarten, grade retention, teacher testing, effects of high-stakes accountability testing, and most recently the use of classroom formative assessment to support teaching and learning. Shepard has served as President of the National Council on Measurement in Education and as President of the American Educational Research Association. She is the immediate past president of the National Academy of Education.

Lee S. Shulman, *President Emeritus of the Carnegie Foundation for the Advancement of Teaching and Charles E. Ducommun Professor of Education Emeritus at Stanford University*. Shulman's research laid the foundation for the approaches to teacher assessment used by the National Board for Professional Teaching Standards and the more recent portfolio-based methods of assessment for teacher licensure. He is a past president of the American Educational Research Association and of the National Academy of Education.

Claude Steele, *I. James Quillen Dean for the School of Education, Stanford University*. Claude M. Steele is the new Dean for the School of Education at Stanford University. Previously, he served as the 21st Provost of Columbia University, as well as a professor of psychology. His research focuses on the psychological experience of the individual and, particularly, on the experience of threats to the self and the consequences of those threats. He has been elected to the National Academy of Sciences, the National Academy of Education, the American Academy of Arts and Sciences, and the American Philosophical Society. He is a member of the Board of the Social Science Research Council and of the John D. and Catherine T. MacArthur Foundation Board of Directors. He has also received numerous fellowships and awards.

National Center for Research
on Evaluation, Standards, & Student Testing

Stanford Center for Opportunity Policy in Education
Barnum Center, 505 Lasuen Mall
Stanford, California 94305
Phone: 650.725.8600
scope@stanford.edu
http://edpolicy.stanford.edu
@scope_stanford