From No Child Left Behind to Every Child a Graduate

MEANINGFUL MEASUREMENT

The Role of Assessments in Improving High School Education in the Twenty-First Century

ALLIANCE FOR EXCELLENT EDUCATION

From No Child Left Behind to Every Child a Graduate

MEANINGFUL MEASUREMENT

The Role of Assessments in Improving High School Education in the Twenty-First Century

June 2009

ALLIANCE FOR EXCELLENT EDUCATION

Suggested citation:
L. M. Pinkus, ed., *Meaningful Measurement: The Role of Assessments in Improving High School Education in the Twenty-First Century* (Washington, DC: Alliance for Excellent Education, 2009).

Ordering information:
Copies of *Meaningful Measurement: The Role of Assessments in Improving High School Education in the Twenty-First Century* can be downloaded from the Alliance's website at www.all4ed.org. To request print copies of the report, please visit http://www.all4ed.org/publication_material/order_form. The first copy of the report is complimentary. Additional copies are available at a charge of $1 per copy to cover shipping and handling costs.

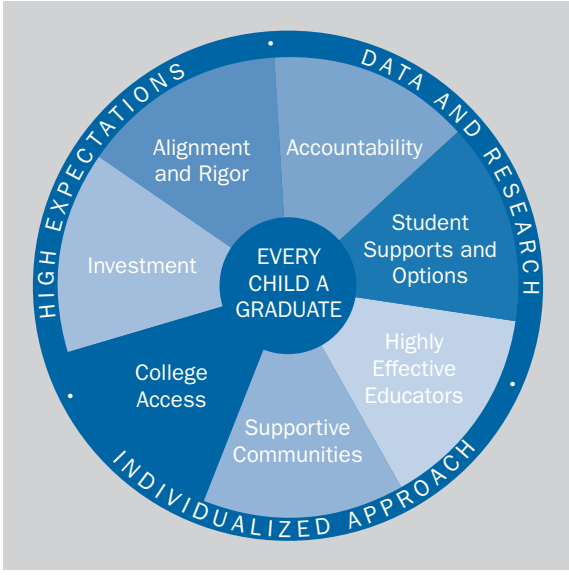
About the Alliance for Excellent Education

The mission of the Alliance for Excellent Education is to promote high school transformation to make it possible for every child to graduate prepared for postsecondary learning and success in life.

The Alliance for Excellent Education is a national policy and advocacy organization, based in Washington, DC, working to improve national and federal policy so that all students can achieve at high academic levels and graduate high school ready for college, careers, and citizenship in the twenty-first century.

The Alliance has developed a "Framework for Action to Improve Secondary Schools" that informs a set of federal policy recommendations based on the growing consensus of researchers, practitioners, and advocates about the challenges and solutions for improving secondary student learning.

The framework, shown graphically here, encompasses seven policy areas that represent key leverage points in ensuring a comprehensive, systematic approach to improving secondary education. The framework also captures three guiding principles that apply to all of the policy areas. Although the appropriate federal role varies from one issue area to another, they are all critically important to reducing dropouts and increasing college and career readiness.



About the Editor

Lyndsay M. Pinkus is director of strategic initiatives at the Alliance for Excellent Education. Since joining the Alliance in January 2002, she has served in a variety of research, coordination, and advocacy roles, where her work has included managing policy and grant work on a range of issues including graduation rates, data, secondary school accountability, and secondary school improvement, and authoring a number of publications for the Alliance. Prior to rejoining the staff in January 2006, Ms. Pinkus served as a legislative associate at Washington Partners, LLC, providing government relations and policy research and analysis for a variety of clients, including the Alliance. She is a graduate of the School of Public Affairs at American University as a presidential scholar; the Public Affairs and Advocacy Institute at the Center for Congressional and Presidential Studies; and the Institute for Educational Leadership's Education Policy Fellowship program.

Acknowledgments

The Alliance for Excellent Education is greatly appreciative of the authors for sharing their time and expertise in writing the following chapters, as well as of the multiple Alliance staff members and advisors whose dedication contributed significantly to this volume.

The Alliance for Excellent Education is also grateful to Carnegie Corporation of New York for the financial support that made this publication possible.

The views expressed in this volume are those of the authors and do not necessarily represent those of the Alliance for Excellent Education or the funder.

Contents

Introduction

"[D]ropping out of high school is no longer an option. It's not just quitting on yourself, it's quitting on your country—and this country needs and values the talents of every American. That is why we will provide the support necessary for you to complete college and meet a new goal: by 2020, America will once again have the highest proportion of college graduates in the world."

—President Obama, February 24, 2009

College and Career Readiness for All: The Dual Challenge

Success in today's global and entrepreneurial economy increasingly requires some form of postsecondary education or training. Yet too many students—particularly poor and minority students—leave the K–12 system without the knowledge and skills necessary for success after high school. The long-term implications of an inadequate education have social and economic consequences for individuals, the communities in which they live, and the nation as a whole. As a result, the country is beginning to embrace a new goal for the public education system: graduate every child ready for college and careers in the twenty-first-century global economy. And at the highest levels of national leadership there is a call to action to dramatically increase the number of American students going on to success in college.

The challenge ahead is twofold. First, the mission of our public education system must shift from "educate some students and prepare them for the twentieth-century American economy" to "educate all students and prepare them for the twenty-first-century global economy." The system goals must be radically altered. These new expectations must be made clear at all levels of the system—from federal and state policies establishing standards, accountability systems, and graduation requirements to the culture established in individual schools. The second part of the challenge is to fundamentally improve the education system's performance in delivering this twenty-first-century education to all students. This will require improvements in the delivery of instruction, the allocation of human, financial, and other resources, and efforts to address the nation's chronically lowest-performing high schools, among other things. Ultimately, the nation demands that the education system not only aspire to higher performance for all students, but that it deliver that result.

The federal government has traditionally taken action in the education arena for three specific reasons: (1) to reduce poverty, increase equity, and serve the most disadvantaged; (2) to ensure national security and economic and competitive position; and (3) to advance research that supports state and district innovation, policies, and practices. Given the economic, societal, and civil rights imperatives of ensuring that the public education system adequately prepares our students—the nation's future workers, consumers, and leaders—there is clearly a federal role in addressing the current weaknesses in the system.

The Role of Assessments in Addressing the Dual Challenge

Assessments can be vital tools in addressing the dual challenge described above—changing and raising expectations and improving the education system's capacity to meet those expectations. Assessments both clarify expectations and measure progress toward meeting them. Assessment results have consequences for students in the form of grades, promotion, graduation, and college admission. Assessments also play a meaningful role in improving the delivery of education. Classroom assessments help inform educators' classroom instruction on a daily basis. The results of summative assessments—large-scale assessments designed to measure student learning at the end of a period of time, such as a course or a grade—play an important role in holding the system responsible for student outcomes, particularly when they are shared publicly and transparently as part of accountability and improvement systems. And assessment data—from a variety of assessment sources—can help inform systemic improvement efforts at the school, district, and state levels, guiding decisions about professional development, resource allocation, and program effectiveness.

Federal policymakers have long recognized the power of summative assessments in playing these various roles, primarily through the Elementary and Secondary Education Act of 1965 (ESEA) and the Individuals with Disabilities Act (IDEA). Over the course of the standards-based movement, federal policy has increasingly required states to develop, administer, and report the results of statewide assessments. The focus has shifted over time from a narrow concentration on measurement to monitor specific program implementation (for example, measuring the academic achievement of

students served by a specific program, such as ESEA's Title I) to monitoring the academic achievement of all students.*

Today, the current version of ESEA, known as the No Child Left Behind Act (NCLB), requires that states administer annual reading, math, and science assessments to all students in grades 3–8 and once in grades 10–12, and assessments of English language proficiency to all English language learners in grades K–12. Through IDEA and NCLB, states are required to include students with disabilities in these assessments, with or without accommodations, and to develop an alternate assessment for students with the most significant cognitive disabilities. Through the NCLB accountability system, these results are reported publicly and used to trigger mandated actions in low-performing schools. States must also participate in the National Assessment of Educational Progress, also known as the Nation's Report Card.

Current Assessments and Assessment Policies Do Not Support the Dual Challenge

Unfortunately, there is a general consensus that current assessment policies and practices are not designed to support the dual challenge: they neither establish the goal of college and career readiness for all students nor support improved practices that will help educators achieve this goal. There are oft-articulated criticisms of the quality of current summative assessments, those assessments' lack of usefulness to educators in improving instruction, and the unintended consequences created by accountability systems that rely so heavily on them. Concerns also exist about the lack of incentives or policies to promote assessments that can inform teaching and learning, such as formative assessments (classroom assessment practices that inform daily instruction) and performance assessments (those that give students opportunities to demonstrate their knowledge and skills through real-world tasks that represent the key aspects of their learning). However, these challenges are not insurmountable, and promising practices from across the globe demonstrate ways forward. In the chapters that follow, leading experts

* W. J. Popham, *The Role of Assessment in Federal Education Programs* (Los Angeles: University of California, Los Angeles, 2008).

Chapter Synopses

- In "College and Work Readiness as a Goal of High Schools: The Role of Standards, Assessments, and Accountability," John Tanner of the Center for Innovative Measures at the Council of Chief State School Officers establishes why, in the twenty-first century, the nation needs standards, assessments, and accountability systems aligned to college and career readiness, and offers recommendations for federal policymakers to support such efforts.
- In "Reframing Accountability: Using Performance Assessments to Focus Learning on Higher-Order Skills," Ray Pecheone and Linda Darling-Hammond of the School Redesign Network at Stanford University discuss how performance assessments can help evaluate students' ability to apply their knowledge and encourage teaching and learning of higher-order skills. They describe what performance assessments are and how they can benefit instruction, how they are being used to support policy goals in the United States and abroad, the major challenges and considerations that must be addressed to use performance assessments well, and how federal policy can support the development and implementation of high-quality assessments that both support and evaluate more complex knowledge and skills.
- In "Formative Assessment and Assessment *for* Learning," Jan Chappuis, Stephen Chappuis, and Richard Stiggins of the ETS Assessment Training Institute describe the characteristics of formative assessment, with a particular focus on those formative assessment practices that engage and empower students in their own learning, or assessments for learning. They also describe challenges related to the effective use of formative assessment and recommended actions for policymakers.
- In "The Role of Interim Assessments in a Comprehensive Assessment System," Judy Wurtzel, of the Aspen Institute, and Marianne Perie, Scott Marion, and Brian Gong of the National Center for the Improvement of Education Assessment, differentiate between true classroom formative assessment and the interim assessments currently in the marketplace. They then provide a framework for considering the appropriate role of interim assessments.
- In "International Assessments of Student Learning Outcomes," Andreas Schleicher of the Organisation for Economic Co-operation and Development provides a brief introduction of the history of international assessments and describes the potential benefits of international assessments for educational policy and practice. He discusses some of the methodological challenges faced in providing valid, comparable, and reliable evidence, and offers recommendations to U.S. policymakers.
- In "Measuring Student Achievement Growth at the High School Level," Joseph Martineau of the Michigan Department of Education explains the technical underpinnings of growth models, describes the various types

of growth models, articulates challenges inherent to measuring growth at the high school level, and explores implications for policymakers interested in moving toward the widespread use of growth models.

- In "Assessing High School English Language Learners," Jamal Abedi of the University of California at Davis describes the challenges inherent in assessing the English proficiency and content knowledge of the diverse high school English language learner (ELL) population and offers recommendations to federal policymakers for creating reliable, valid, and accessible assessments for ELL students.
- In "Students with Disabilities: Expectations, Academic Achievement, and the Critical Role of Inclusive Standards-Based Assessments in Improving Outcomes," Rachel Quenemoen of the National Center on Educational Outcomes describes issues concerning the assessment of high school students with disabilities in a standards-based accountability system, ways to evaluate assessments that are inclusive of all students in the accountability system, and recommendations for policymakers.
- In "Assessments and Technology: A Powerful Combination for Improving Teaching and Learning," Erin Martin Gohl, Daniel Gohl, and Mary Ann Wolf of the State Educational Technology Directors Association describe how the use of technology to assess students and to record and analyze performance can result in timely, appropriate, and individualized instruction for all students. They highlight some of the innovative approaches in using technology to assess student progress, address current challenges in the use of technology, and provide recommendations to federal policymakers to overcome those challenges.

describe some of the assessment challenges in greater detail and provide federal recommendations on how to address them.

Rethinking Assessments and the Federal Role

Meeting the dual challenge of raising the bar for high school graduation to represent college and career readiness while simultaneously helping to ensure educators and students clear that bar will require rethinking the assessments and the federal role in supporting them.

Current federal policy does nothing to establish college and career readiness as the goal for all students or to ensure that standards and assessments are both aligned to this goal and comparable across states. Today, the nation relies on more than fifty sets of state standards and assessments that define

expectations and proficiency in fifty different ways. As a result, expectations about what students should learn are dependent on their state of residence, zip code, and curriculum track rather than on a common understanding of the skills, content, and competencies necessary for college, careers, and life. Meanwhile, current federal policy mandates how educators should address low-performing schools by requiring a specific sequence of one-size-fits-all interventions that are not informed by more specific data about the challenges that are unique to the schools themselves.

This approach should be reversed. Federal policy should establish college and career readiness as the goal for all students and support collaborative state-led efforts to define those expectations through a set of common standards and assessments. Federal policy should require that policymakers, administrators, and educators use information from these assessments to inform decisionmaking around teaching, learning, and student outcomes and ensure improvements. However, it should leave those decisions—about what to do, when, and how—to the educators who are closest to students and schools.

With this approach in mind, federal policy should do the following to support the development and use of assessments to establish college and career readiness as the goal for all students, and to improve the education system's capacity to meet that goal:

Support the development of common standards and assessments. Federal policymakers should support state-led efforts to develop common standards and assessments that are aligned to college and career readiness and reflect global best practice. This should be accompanied by incentives for states to adopt these standards and assessments, to use them as part of their K–12 accountability systems, and to better align secondary and postsecondary education. Federal policy should continue to require states to include all students, including students with disabilities or limited English proficiency, in the assessment process, through the development of high-quality, appropriate accommodations and modifications for those assessments.

Federal policy should also ensure full U.S. participation at the national and state levels in international assessments of student performance, including

the Programme for International Student Achievement (PISA) and the Trends in International Mathematics and Science Study (TIMMS). These opportunities to compare our performance and the quality of our standards and assessments internationally are critical to efforts to improve policies, practices, and student outcomes.

Support the timely and transparent communication and use of assessment results. Assessment data is only actionable if it is accessible. NCLB made significant progress on this front by requiring the public reporting of test results for every school and each of its student subgroups. Federal policymakers should continue this progress by supporting the timely and transparent communication of all assessment results for use by educators, parents, policymakers, and researchers. Federal policy should require the use of information about students' college and career readiness, along with other important data, to inform decisionmaking about school improvement efforts and allocation of resources, such as distribution of teachers and targeting of professional development. Federal policy should support development of the technical infrastructure necessary for communication and use of assessment results. It should also provide incentives for the development of coherent data systems at the state, district, and school levels that support the collection, communication, analysis, and use of assessment data.

Improve educators' capacity to use data to improve teaching and learning. The assessment landscape is broad and complex. As described in various chapters in this report, educators at the school, district, and state levels are using innovative tools such as performance assessments and formative assessment processes that engage students in their learning and give educators valuable information that can be immediately used to improve instruction. Federal policy should help build educators' assessment literacy: this includes both their ability to embed assessment practices in instruction and their capacity to use multiple kinds of assessment data to make informed decisions about instructional practice and program design. Beginning with pre-service education and continuing through induction programs and professional development practices, educators must be prepared and supported to effectively use a wide range of assessments that

inform instruction and student interventions. Federal policies and funding streams designed to help make educators more effective—such as ESEA's Title II, the Higher Education Act's professional development programs, or the Enhancing Education through Technology program—should be targeted to support these practices.

Invest in research and development to improve our collective knowledge about the development and use of assessments in ways that improve teaching, learning, and student outcomes. There are a number of assessment issues that need further exploration, such as the impact of interim assessments on students' learning, the development of appropriate assessment options for some students with disabilities, and the development of sound growth models at the high school level, to name a few. Federal policymakers should dedicate some of their research and development agenda to exploring these key assessment issues. Also, educators are demanding increased information about how to effectively use assessments and assessment data to improve teaching, learning, and student outcomes. Federal policy should support the collection and dissemination of best practices related to assessment use.

CHAPTER 1

College and Work Readiness as a Goal of High Schools: The Role of Standards, Assessments, and Accountability

John Tanner
Center for Innovative Measures, Council of Chief State School Officers

Much has been made lately of the idea represented by the phrase "college and work ready" as a goal of the educational system. The purpose of this chapter is twofold: to give policymakers a perspective on the subject through the eyes of those tasked with the standards, assessment, and accountability work, and to provide information for policymaking so the system can produce the intended result. First and foremost, the goal assumed in this chapter is that it is the responsibility of high schools to produce college- and work-ready students, and that our systems of standards, measures, and accountability must support that effort. This chapter will begin with this notion, and then describe the systems that need to be in place to support the goal.

The College- and Work-Ready Student

Research has clearly identified college readiness as containing academic elements related to the ability to succeed in college courses, but also the capacity to tackle the culture, intellectual norms, and social environment of

the college setting.[1] ACT research suggests that preparing students for the worlds of work and college requires a range of factors other than academic ones—so-called "habits of mind"—including traits like persistence, cooperation, and teamwork.[2] Thus the concept of college and workplace readiness contains both academic and nonacademic elements, and the absence of either set diminishes the ability of students to succeed beyond high school.

In addition to this, standards that are representative of college readiness must go beyond typical academic content, since some amount of skill and knowledge in a particular domain is not enough; what is needed in all cases is the *right* skills and knowledge. For example, the ability to write expository, descriptive, and persuasive text and to perform effective research is necessary for college success, since students will be required to use those skills in almost every postsecondary course they take.[3] But proficiency in these areas is not always measured before graduation. Within the core academic subjects—which research suggests should include English, math, science, social studies, world languages, and the arts—only English and math are currently part of every state's standard set.

The Next Generation of Standards

If standards are to support the current notions of what it means to be ready for college and work, they must do far more than simply articulate expectations in reading, writing, and math. Without the inclusion of elements from other academic and nonacademic areas, the definition of readiness is incomplete. Along those same lines, there really is no such thing as college- and work-ready English language arts standards, or college- and work-ready mathematics standards; instead, both are part of a larger, more complete definition of readiness.

For policymakers, this means that while it is fine to mandate academic standards in core content areas, doing *only* this falls short of describing what college and work readiness really looks like. Standards that do not incorporate habits of mind, various work attitudes, and motivation are not educating students to be college and workplace ready with any degree of reliability.

It should be noted that, as of this writing, an effort is under way for states to come together as a first step toward generating standards that meet some of these "next-generation" criteria. Led by two of the leading state education advocacy groups, the Council of Chief State School Officers and the National Governors Association, the project seeks to produce standards in English language arts and mathematics—referred to as the "common core"—that states could voluntarily adopt as part of their regularly occurring revision cycles. In the process of determining what should make up that core, there has been much focus on notions of college and workplace readiness.

While this is an extremely important and exciting development, work remains to be done. While English language arts and math are indeed critical elements of any definition of post–high school success, they are only two of the domain areas identified as necessary for *college* success, and students' habits of mind need to be addressed as well. The common core activity should be treated as a critical first step and be offered full support, but other steps will need to be taken as well.

It is crucial that the common core effort remain a state-led and voluntary activity. Federal policy could certainly promote adoption of the common core standards and encourage collaboration among states in other ways, but states have expressed great interest in this work precisely because it is something they are choosing to do, not something they are required to do. As a result, the chances for successful implementation are much higher, and states will be able to retain their traditional role with regard to education.

Measuring College and Work Readiness

The tendency with reading, writing, and mathematics assessment, which occurs in virtually every state, is to pick a score on each test that signifies that students at or above that level are "ready" for college and the workplace. This may seem perfectly logical, but there are several reasons why it fails to accomplish its goal.

1. The scores on traditional tests (other than writing tests that ask students to respond to a prompt) are derived from test items that generally come from the lower end of the standards, meaning that

success on the tests is not the same as the level of achievement needed to succeed beyond school. That is, the tested content tends to come from the material in the standards that can be easily tested, which often does not include the skills and competencies that are necessary for college and workplace success. Success on tests that assess the "wrong" portion of the standards when it comes to success beyond school does not equate with success on the "right" portion of the standards, even if the passing score is set extremely high.

2. Tests given in states are usually limited to a few core content areas. Even if they consisted of a new generation of measures capable of addressing the full range of the standards, they would still not assess important content domains and nonacademic elements. Any declaration of college and work readiness from only a portion of the requirement risks a very high probability of being wrong, either in failing to identify students who may be closer to readiness than the existing data elements suggest, or in implying that students are ready when in fact they fall short on other measures. Such misidentification is even more probable when the tests, as indicated above, do not directly measure what is required.

3. As long as no measurement exists for nonacademic aspects, no declaration of complete success against the goal of college or work readiness is valid; any such declaration can be made only against the full range of requirements. Certainly, students should not be expected to meet all the criteria in exactly the same way, but knowing how prepared they are increases the likelihood that resources can be correctly deployed for those who need more support.

The Next Generation of Measures

If the measures are to have relevance in terms of the goal of college and workplace readiness, the academic measures must change, not just to incorporate the full range of the standards, but to focus on the issues that really matter to the definition. If this does not happen—and if the measures continue to be drawn from the lower end of the standards—then no amount of success on those measures can be said to equal eventual college or workplace success.

This is more than simply covering the full range of the standards in the test. Consider the way that cut scores work. When a student takes a test and answers 70 percent of the items correctly, he or she will generally have answered correctly the easiest 70 percent (with, of course, some anomalies). On a test that covers the full range of a set of standards, it is likely that the passing score occurs somewhere below the realistic threshold for everything students need to know and be able to do in that domain area to succeed beyond school. On current tests, even a perfect score may fail to reflect many of those things that are most important, because they were not included on the test.

When it comes to nonacademic elements of success—such as intellectual openness, inquisitiveness, analysis, reasoning, argumentation and proof, interpretation, precision and accuracy, and problem solving[4]—measurement needs to fundamentally change. Research has proved that these "soft skills" are imperative to the success of students beyond school.[5] In saying this, one runs the risk of being accused of dumbing down existing standards, but nothing could be farther from the truth. As Lauren Resnick, a well-known researcher and longtime standards advocate, said some time back, "The common idea that we can teach thinking without a solid foundation of knowledge must be abandoned. So must the idea that we can teach knowledge without engaging students in thinking."[6]

But assessing these habits of mind accurately is quite another matter. The generation of measures that policymakers need to support in the name of college and workplace readiness will have to take on a very different form than current ones. The assessments may continue to involve the use of domain-based tests—improved, of course, from the current versions—but they also must include measures that show students being aware of and developing the habits of mind that provide them with a complete portfolio of what is necessary for life after high school.

Consider, however, that an assessment of "intellectual openness" would be quite silly if it were done using traditional testing practices. However, all of the elements listed above are behaviors that can become part of a student's repertoire if inserted properly into the educative process. And assessing those behaviors is doable. Consider a theoretical system, implemented in addition

to traditional accountability testing, that students managed themselves in the course of their education. This could be online, and arranged in a format similar to Facebook or MySpace accounts. Ideally, it would

- assess the domain knowledge necessary for success in college and work by asking students to complete online tasks and activities that directly measure the skills and abilities identified as important;
- include scores from accountability testing to show whether scores on tasks and other activities are reasonable and within the realm of expectation given other performances;
- include observations from teachers and other adults with regard to traits that cannot be easily assessed via traditional measures; observations would be against established rubrics and require multiple observations from multiple teachers to be considered valid, and the purpose would be to always help the student grow and progress;
- give students the opportunity to manage their work in terms of their own desires and goals; and
- give school counselors and administrators sufficient access to support and encourage students in a meaningful way with regard to their goals.

Policymakers could encourage this new type of assessment system through targeted research and development (R&D) that focuses on identifying goals and outcomes rather than just the means. An R&D approach is the right one, because technology is changing so quickly that there is a risk of creating an overly bureaucratic system that might fail the audiences it is attempting to serve. The right system would encourage the use of innovation, open sourcing, and social networking, all in the name of encouraging and supporting education in a particular direction. In this sense, the system should not be "federal" or expected to function dynamically, but rather should follow a similar pattern as the common core, wherein states lead and adopt autonomously.

On a simpler note, policymakers should also recognize that in their enthusiasm to have test results come back quickly, they are directly responsible for states turning away from more innovative assessment

practices. This enthusiasm is driven by the misplaced assumption that state accountability testing has such enormous diagnostic power that the results should be driven into classrooms at the earliest possible moment. Unfortunately, two things are wrong with this thinking. First, only tests designed explicitly for diagnostic purposes can serve this diagnostic role. Accountability tests—which sample out from the lower end of the domain—do not contain sufficient items to do more than ascertain students' strengths and weaknesses at the broadest possible levels. As a result, instructional decisions based on the results of accountability tests may miss or misdiagnose underlying problems, and are likely to be unhelpful or possibly even harmful.

Second, test scores are indicative only of the material tested. As a result, instructional changes initiated in response to test scores on assessments are limited to the areas covered by the assessment. Unfortunately, accountability tests do not represent the full range of content and skills that students need for success beyond school. This means that decisions based on accountability tests—which do not contain the most significant, relevant material—will also be limited.

Policies that are insensitive to the appropriate uses of data from test instruments—however well intentioned—must accept that a nondiagnostic instrument used inappropriately can damage the system as likely as help it. More sensitivity to having the policy dictate outcomes without dictating the means would be extremely helpful.

Systems Versus Students When It Comes to Accountability

No argument that includes assessment and standards has a chance of succeeding unless it also addresses the issue of accountability. Examining the assumptions underlying current accountability systems can help illuminate where policymakers need to make changes if the notion of college and work readiness is to be supported in the distinction of what it means to hold schools accountable.

First, accountability should continue to be about students "meeting the standard," but it is important to understand what that means now and what it might mean in a new system.

Written standards attempt to break down broad goals—such as numeracy and literacy—into specific, manageable pieces that define what students should know and be able to do. Prior to the accountability movement, traditional, short, paper-and-pencil tests that measured the component parts of the standards were intended to function as efficient proxy measures for far more complex and broad material. However, the intent was not to lose the focus on the broader goals; "meeting the standard" was intended to take on the more global meaning of reaching the broader goals. For example, a forty-item reading test administered as a research tool did not directly measure literacy, but because of the correlation between success on the test and actual literacy, the assessment could serve as a proxy measure for that important educational goal. These types of methods were cheap and effective, and thus came to be used widely.

When policymakers recognized that success on such tests was indicative of the broader goal, the reaction was to create an accountability structure around these tests. After all, if students who did well on the tests could indeed be shown to have met the goal of literacy, then holding schools and students accountable to such tests made sense and the broader goal seemed within reach. Policymakers failed to recognize that these tests are a sample of the larger domain—generally sampling from the portions that are least relevant to college and workplace readiness—and that placing accountability on only a sample of what is deemed important could have some fairly serious unintended consequences.

As accountability was placed in the less sophisticated, less relevant proxy, the message that was sent to teachers and schools was that what mattered was not the standard as shown in the written documents that represented a domain, but rather what was necessary for students to pass the test. In the current system, students who "meet the standard" are deemed successful, as are the teachers and systems that helped them do so, but the standard represented in the tests is a far cry from written standards, and preparation for these tests is not adequate to prepare students for college and work.

Grounding accountability in tests designed to serve as proxies means that the proxies now define what matters. This mistake is exacerbated when policymakers demand tighter and tighter turnaround times for state tests,

which increases the use of assessments that can be scored quickly, which in turn increases the likelihood that states will continue to allow simpler and simpler proxy-based tests—which no longer function as proxies—to be the drivers of what it means to meet the standard. This is made even worse by budget cuts to testing programs, which force states to use even simpler testing methods for cost reasons.

Finally, as long as the *status* measure is the important one—the measure that compares, say, the percentage of this year's fourth graders who meet the standard with last year's fourth graders—the material that is most relevant does not need to be taught in order for the system to succeed. As in the example earlier, if a passing grade on a test is 70 percent, consider what gets communicated when accountability is added to that passing score: it suggests that any teacher or system that can get a student to that level has "succeeded." But what meaning is attributed to the most difficult material on the test, regardless of whether the test measures the full richness of the standards or just the lesser pieces, in the form of a proxy? The answer is that for the purpose of success of the accountability system—with the indicator being the percentage of students who achieved the passing score—it does not matter. The *system* could be deemed successful even if the most relevant material was not even taught.

This means that teachers wind up with a very unhealthy tension when it comes to doing their jobs. Do they teach so that students can pass the test, and help the public and school administrators whose jobs are on the line, which may translate into "teaching to the test" and neglecting the most relevant material? Or do they concentrate on the needs of each and every student and on what it will take to get those students to the next level of knowledge? This places teachers in an unfair and difficult situation, where the needs of the system and the needs of the students do not always match, and, indeed, are often in direct conflict.

Policymakers can remove this tension by changing from a system focused on status to one focused on *growth*. However, it is important that the growth models be selected carefully, since the notion of growth as represented by test scores is one that requires real technical expertise in order to avoid

another round of unintended consequences.* This is not to suggest that simply switching to a growth model automatically alleviates the concerns of a status model, but consider the following:

1. As mentioned above, one of the concerns of the status model is that teachers are put in the untenable situation of negotiating between the needs of the students and the needs of the system. A growth model stands an excellent chance of aligning those needs in a way the status model cannot.

2. Another concern of the status model is that it focuses teaching around the passing score and not necessarily on the material most relevant to any of the possible definitions for college and workplace readiness. If growth is required for all students from one year to the next, the opportunity for the full range of material to be learned increases.

3. The notion of growth fits with the view that teaching is about meeting a student where they happen to be and then doing whatever is necessary to move that student to the next level. It fits with the commonsense approach that, while schools cannot control what happens outside or beyond schools, they certainly should be held accountable for what happens *in* schools. At the teaching level, it fits with the notion that a teacher cannot be held accountable for what occurred prior to a student coming into their classroom, but a teacher should be held accountable for his or her own efforts with that student. A status model in which this year's students are compared to a completely different set of students from a previous year just does not make sense to a teacher.

---

* When it comes to growth models, the actual results can seem counterintuitive to a nontechnical audience. A quick example may be helpful to show why care must be taken in determining the model. Consider the concept of "growth to standards," which has been indicated as a desirable trait for growth models. In practice, many such models wind up replicating the status models. This happens because if the achievement goal for each student becomes the passing score, then students who were farther from the mark in the previous year must take a bigger step in the current year than students who are closer to the mark—but schools are given credit for growth only when a student meets the goal, not when he or she makes progress toward it. Students who are closer to the bar are more likely to meet the goal, and a school gets credit for the growth of those students but not for the growth of students who had to take a larger step. Because the students who were closer to meeting the standard in previous years are also the students most likely to meet the standard in the current year, the growth results tend to look remarkably similar to the status results—but at much greater effort and cost and with no appreciable benefit. And while schools often do generate tremendous growth for their lowest-achieving students, if that growth misses the target it is considered unsuccessful by the model.

Next-Generation Accountability

Clearly, policymakers should consider the movement to growth models as a necessary—but not sufficient—step toward a better system. But a growth model against test instruments is still incomplete, even if those instruments address the full range of the material determined appropriate to definitions of college and work readiness.

There still exists the possibility that for those areas deemed to be completely necessary for college or workplace success (such as writing and the ability to do research), schools could be declared successful when students do not leave possessing such skills. In this case, the accountability measure could be rather simple, but with a profound impact: a school would need to state something to the effect that "no student shall leave this high school without having written well at least once," and then commit to aligning resources around such a goal, beginning in the freshman year and commencing only when the student had met the goal. Although this is a paradigm shift from current practice, it has numerous benefits for the student and the school:

1. It galvanizes the student and the school to focus on the most significant, most relevant goals, and makes those goals a priority.

2. It makes the standard have meaning for all students without discounting the fact that students enter and leave high school with a variety of talents and skills. Some students would complete the requirement almost upon entering, while others might take four or more years. Regardless of timing, all students would leave with the same skill set.

3. It reinforces the idea that the institution as a whole, not just the individual domain teacher, is responsible for the teaching and learning of a given set of skills.

In short, while not all requirements could or should be given that kind of treatment, galvanizing student attention and system efforts would be useful for the critical tasks.

As to the various domain areas, care must be given within accountability measures for schools to address what research suggests is necessary for continued success. To dictate this from a federal level seems counterproductive; it should, instead, be supported and encouraged at the local level.

Finally, any accountability system that purports to support college and workplace readiness cannot stop at the academic level, since the definition of college and workplace readiness includes nonacademic elements as well. It was suggested earlier in this chapter that policy support for tools that would enable students and school staff to provide documentation and evidence for these other elements would call much-needed attention to them, but some level of accountability needs to be attached to the effort as well. This clearly cannot be done in the same manner as the domain areas, given both the highly subjective nature of many such observations and the fact that holding schools accountable for some habits of mind clearly biases the judgment. After all, if a school administrator is in charge of evaluating the habits of mind of the school's students, and that administrator also is held accountable for scores going up, those scores are suspect, no matter how honest and objective the administrator is in assigning them.

On the other hand, simply requiring schools to fill out questionnaires or check items off a list is just as likely to produce an invalid result, since it creates even more of a bureaucracy than currently exists and risks being done for the wrong reasons. The resulting information would fail any test of reasonableness and create one more meaningless hoop for students and administrators to jump through.

Rather, if the goal is to prepare students for the world of college and work, their preparedness once they reach those arenas should be considered part of the accountability measure. As students leave high school and embark on the next step, are they, in fact, prepared? Do those students who move on to college coursework require remediation? Do they drop out after their freshman year, and, if so, is it because they were not adequately prepared? Do students enter the world of work ready for workplace challenges? What do their employers have to say about their level of preparedness?

This remains—and will likely remain for some time—a policy challenge. If the skills identified as necessary to succeed beyond school are the skills that are important, and schools effectively support students in obtaining those skills and abilities, then that ought to be reflected beyond schooling. A set of accountability indices that enable schools to view the results of their efforts and feed such information back into their practice would be powerful, but it would have to function as a carrot, not a stick. Schools should be held accountable for careful planning in light of such indices, for showing evidence of improvements and the steps taken, and for thoughtfully pursuing improvements. It would be a mistake to simply assign an index to a school and then demand improvement. Such a strategy dismisses the complex nature of what will happen once students move beyond high school, and would likely cause far more harm than good.

Accountability to a thoughtful process that considers the results from such surveys and resulting indices treats educators as the professionals they are, and gives them a chance to act upon information they do not currently have but would find immensely valuable. It also demands that they do this as part of their regular practice in the name of the goals of education.

Conclusion

The most important point for policymakers to take from this chapter is that definitions of college and workplace readiness are not contained in a test score, but rather in a number of domain areas and habits of mind that together form the basis of what students need to succeed beyond schooling.

The second most important point is that in the creation of college- and work-ready standards and assessments, the nation must move beyond traditional notions of both, particularly when it comes to ensuring that students possess the appropriate work habits. Tests can certainly be a part of that assessment system, but a policy that stops at simple testing will always fall short.

Finally, while accountability can still invoke the phrase "meets the standard," it needs to take a very different form than in the current system. Policymakers should recognize, accept, and gear policy toward allowing this to happen. Much remains to be done to create a new accountability

paradigm, but if it is done properly, it stands a much-improved chance of success over current systems.

Policymakers should consider the following actions (in addition to others not listed here), to encourage and support the idea of college and workplace readiness.

1. **Support the common core effort as one that should be led by states and voluntarily adopted state by state.** Included in this should be the notion that while reading, writing, and mathematics are critical, research has suggested that stopping at those domain areas for the notion of college and workplace readiness causes the definition to be incomplete.

2. **Encourage definitions of college and workplace readiness that include the habits of mind elements**, but that will also allow new research to fine-tune and improve the definition.

3. **Support the move to next-generation assessment.** This may include much more flexibility in terms of the return of test scores to allow for more meaningful measures, multiple measurement systems that include more than just traditional test scores, and online systems that allow for the proper data to be collected and shared appropriately. A way for policymakers to consider their work is to stop thinking about assessment and start thinking about measurement. By asking "What are we trying to measure?" rather than "What needs to be tested?" policymakers put themselves in a position to make better policy decisions in this arena.

4. **Fund next-generation assessment.** Much of the reason for a lack of real innovation in the assessment space is due to government reluctance to accept something that looks and feels different. When it comes to assessment, this means the continued use of outmoded and outdated methodologies. The fear of change is tremendous in schools, largely because of compliance issues, but this could be alleviated if policy were seen as supporting innovative practices that have been proven to be successful.

5. **Support growth models for accountability purposes in domain areas, but recognize that not all growth models are created equal, and that dictating the technical elements of the work has created some unintended consequences.** To this end, policymakers will need to keep an open mind and trust the research that has produced excellent information on understanding and measuring growth.

6. **Support accountability definitions that include information from a variety of sources**, including observations, test scores, portfolios, and so on. Policymakers should acknowledge that many elements of the college- and work-ready definition cannot be assessed in traditional formats. They should also be willing to assign accountability to schools that are carrying out a careful and thoughtful process. If this can be done in light of the goal—that is, what actually happens to students who leave school prepared according to the definition—schools can be held accountable for the thoughtfulness of their planning and strategic processes with regard to emerging data.

The views expressed in this chapter are those of the author and do not necessarily represent those of the Alliance for Excellent Education.

About the Author

John Tanner is director of the Center for Innovative Measures at the Council of Chief State School Officers. Mr. Tanner oversees the council's efforts in the areas of standards, assessment, and accountability, where his work focuses primarily on supporting states in getting to the next generation in all three areas.

[1] D. Conley, *Toward a More Comprehensive Conception for College Readiness* (Eugene, OR: Educational Policy Improvement Center, 2007).

[2] ACT, *Impact of Cognitive, Psychosocial, and Career Factors on Educational and Workplace Success* (Iowa City, IA: Author, 2007).

[3] Conley, *Toward a More Comprehensive Conception for College Readiness.*

[4] Ibid.

[5] ACT, *Impact of Cognitive, Psychosocial, and Career Factors on Educational and Workplace Success.*

[6] L. B. Resnick, "Making America Smarter," *Education Week Century Series* 18, no. 40 (1999): 38–40.

CHAPTER 2

Reframing Accountability: Using Performance Assessments to Focus Learning on Higher-Order Skills

Linda Darling-Hammond and Raymond Pecheone
School Redesign Network, Stanford University

Over the past decade, educators, policymakers, and the public have begun to forge a consensus that our public schools must focus on better preparing all children for the demands of citizenship in the twenty-first century. This has resulted in states developing "standards-based" educational systems and assessing the success of districts and schools in meeting these standards through more systematic testing. Most of these tests are multiple-choice, standardized measures of achievement. While these assessments offer the benefits of ease of administration and inexpensive scoring, practitioners and researchers have found that they also have a number of less desirable side effects. These include narrowing of the academic curriculum and experiences of students (especially those in low-income communities); a focus on recognizing right answers to lower-level questions rather than on developing higher-order thinking, reasoning, and performance skills; and growing dissatisfaction among parents and educators with the school experience.

The sharp differences between the forms of testing used in the United States and the performance-based assessments used in other higher-achieving countries also suggest that low international rankings may be related, in part, to overreliance on these narrow conceptions of standardized testing in the United States.

In large part for cost reasons, reliance on multiple-choice tests rather than on more open-ended assessments of performance has increased in response to the annual testing requirements of the No Child Left Behind Act (NCLB), despite the fact that language in NCLB calls for "multiple up-to-date measures of student academic achievement, including measures that assess higher-order thinking skills and understanding."[1] Changing what counts as assessment evidence, along with related changes in NCLB's accountability structure, could contribute substantially toward school improvement.

This chapter discusses how performance assessments can help evaluate what students can actually do with what they know and encourage the teaching and learning of higher-order skills. It describes what performance assessments are and how they can benefit instruction, how they are used in policy settings in the United States and abroad, what the major challenges and considerations are that must be addressed to use performance assessments well, and how federal policy can support the development and implementation of high-quality assessments that both support and evaluate more complex knowledge and skills.

What Is Performance Assessment?[2]

Almost every adult in the United States has experienced at least one performance assessment: the driving test that places new drivers in an automobile with a DMV official for a spin around the block and a demonstration of a set of driving maneuvers, including, in some parts of the country, the dreaded parallel-parking technique. Few Americans would be comfortable handing out licenses to people who have only passed the multiple-choice written test also required by the DMV; we understand the value of the performance assessment as a real-world test of whether a person can actually handle a car on the road. Not only does the performance assessment tell us some important things about potential drivers' skills, it

also helps improve those skills, as potential drivers practice to get better. (What parent doesn't remember the hair-raising outings with sixteen-year-olds wanting to practice taking the car out over and over again?) The test sets a standard toward which everyone must work. Without it, society would have little assurance about what people can actually *do* with what they know about cars and road rules, and little leverage to improve actual driving abilities.

Performance assessments are used in bar examinations for lawyers, where they must write briefs and analyze cases; in the medical boards for doctors, where they must diagnose patient cases and, in fields like psychiatry, interview patients under the watchful eye of evaluators; and in registration exams for architects, where candidates must submit a portfolio of their designs.

Performance assessments in education are similar. They are opportunities for students to show how they can apply their knowledge and skills in real-world tasks that represent the key aspects of their learning. Performance assessments may include science experiments that students design, carry out, analyze, and write up; computer programs that students create and test; or research inquiries that they pursue, seeking and assembling evidence about a question, which they may present in written and oral form.

Whether the skill or standard being measured is writing, speaking, scientific literacy, mathematical reasoning, or social science research, with a performance assessment students perform tasks involving these skills and teachers score the performance based on a set of predetermined criteria. As in our driving test example, these assessments typically consist of four parts: performance standards, a task, a scoring guide or rubric, and a set of administration guidelines. The development, administration, and scoring of these tasks requires teacher training and development to ensure quality and consistency.

Illinois's assessments provide a good example of the contrast between classroom performance assessment and a state multiple-choice test. The state's eighth-grade science learning standard 11B reads, "Technological design: Assess given test results on a prototype; analyze data and rebuild and

retest prototype as necessary." The multiple-choice example on the state test simply asks what "Josh" should do if his first prototype sinks, with the wanted answer, "Change the design and retest his boat." This, however, gives the assessor no idea whether Josh would have any idea *how* to change the design productively and to systematically test the design, holding some features constant while changing others.

The classroom assessment allows evaluation of these critical questions. The prompt states, "Given some clay, a drinking straw, and paper, design a sailboat that will sail across a small body of water. Students can test and retest their designs." In the course of this activity, students can explore significant physics questions such as displacement in order to understand how a ball of clay can be made to float. Such activities combine hands-on inquiry with reasoning skills, have visible real-world applications, are more engaging, and enable deeper learning. They also allow the teacher to assess student learning along multiple dimensions, including the ability to frame a problem, develop hypotheses, reflect on outcomes and make reasoned and effective changes, demonstrate scientific understanding, use scientific terminology and facts, persist in problem solving, and organize information, as well as develop sound concepts regarding the scientific principles in use.

The assessment systems of most of the highest-achieving nations in the world emphasize local in-school performance assessment throughout the elementary and middle school years. At the high school level, jurisdictions like the UK, Hong Kong, Singapore, Finland, Sweden, and Victoria, Australia, among others, use a combination of centralized assessments that use primarily open-ended and essay questions and local assessments given by teachers which are factored into the final examination scores.

The centralized assessments are often developed jointly by high school and college faculty and scored using common criteria by teachers. The classroom-based assessments—which include research papers, applied science experiments, presentations of various kinds, and projects and products that students construct—are mapped to the syllabus and the standards for the subject, and are selected because they represent critical skills, topics, and concepts. They are often suggested and outlined in the curriculum, and may be designed centrally or locally. They are administered and scored by teachers.

While not all performance assessments are locally developed—Hong Kong offers a bank of tasks teachers can draw upon, while teachers in Finland create their own—all of these systems include some rich assessment tasks at the classroom level that can be used as formative or benchmark assessments, helping teachers to gauge ongoing progress. Local scoring guided by standardized protocols allows immediate feedback to teachers and students. This enables results to be used to improve instruction and student learning immediately, something that standardized examinations with long lapses between administration and results cannot do. In addition, as teachers use and evaluate these tasks, they become more knowledgeable about the standards and how to teach to them, and about what their students' learning needs are. This process improves their teaching. Scoring is often subject to moderation, auditing, or calibration processes, as described later.

Performance assessments often provide several ways to view student learning. For example, multiple samples of actual writing taken over time can best reveal to a teacher the progress a student is making in the development of composition skills. This provides ongoing feedback to learners as well, as they see how they are developing as writers and what they have yet to master. In addition, different kinds of writing tasks—persuasive essays, research papers, journalistic reports, responses to literature—encourage students to develop the full range of their writing and thinking skills in ways that answering multiple-choice questions about writing or even writing a five-paragraph essay over and over again do not.

Locally managed performance assessments that provide multiple sources of evidence about what people can actually *do* with what they know are often characterized as "tests worth teaching to," because they help focus effort on developing important skills. Let's think back to the state driver's license exam. This involves both a written test and a performance assessment on the road. Everyone knows precisely what to expect in terms of the skills to be demonstrated—for example, whether or not the applicant can manage a car safely and (at least on the East Coast) parallel-park skillfully—as the examination is not a total secret. Most performance assessments challenge students to address issues and problems in real life.[3] Moreover, a number of studies associate performance assessment with a positive influence over teaching and learning.[4]

The fact that the assessment is open and transparent is not a problem, because the point is to see whether drivers have developed these real-world abilities; this is not undermined by the drivers knowing what they need to learn to do. The performance is scored by the instructor, working from a rubric, and if the driver is sufficiently successful in all aspects of the examination (as determined by a state cutoff score), a license is conferred. The task is so well defined that instructional programs (driver's education) that include both hands-on and classroom instruction clearly demonstrate their effectiveness in preparing students to perform. (This is reflected in the reduced insurance rates granted to graduates of driver's education programs.) Imagine what life on the roads would be like if prospective drivers did not have to demonstrate what they know before taking the wheel on their own. And imagine what life in classrooms would be like if the nation *did* require students to demonstrate that they can express and defend their ideas, develop and analyze data, and apply their knowledge in problem-solving situations.

Benefits of Performance Assessment

Research and experience have uncovered a number of benefits, challenges, and criteria for making such assessment systems successful. Among the benefits of well-designed performance assessment systems are that they can

- elevate the focus of instruction to include higher-order thinking skills;
- provide a more comprehensive assessment of what students know and can do;
- provide clearer information to parents, teachers, and the public as to student development, accomplishments, and needs;
- allow instruction to be altered in a timely fashion to meet student learning needs;
- lead to more student engagement in both the learning and assessment processes;
- invite more teacher buy-in and encourage collaborative work; and
- support standards-based instruction and improvement of teaching practices.

Considerable research suggests that performance assessments are essential tools for showing the extent to which students have developed higher-order thinking skills, such as the abilities to analyze, synthesize, and evaluate information. Studies have found that the use of such assessments has improved teaching quality and increased student achievement, especially in areas requiring complex reasoning and problem solving.[5] Evaluations of reading and writing portfolios in Vermont and Kentucky, for example, found that the assessments—along with the professional development opportunities associated with them—influenced instruction in positive ways, especially in encouraging much more complex mathematical tasks and more extensive and higher-quality student writing.[6] These assessment systems also stimulated school improvement through curriculum reforms and supports for teacher learning.

Researchers have noted that assessment systems in which teachers look at student work with other teachers and discuss standards in very explicit ways appear to help schools develop shared definitions of quality. Evaluating work collaboratively rather than grading students in isolation helps teachers make their standards explicit, gain multiple perspectives on learning, and think about how they can teach to produce the kinds of student work they want to see. Where teachers do this, studies find that changes in teaching and schooling practices almost invariably occur—especially for students who are not as consistently successful at schoolwork.[7]

Performance assessments are more sensitive to instruction and of more immediate use to teachers than most current standardized tests, while providing richer evidence of student learning that can be used to solve learning problems as they occur. When teachers see their students' written responses and reasoning, they can diagnose *how* students are learning and *why* they may be struggling, rather than just what they know. Typically, standardized test information is not available to schools for six to nine months after the testing date, often in the subsequent school year, and far too late and far too thin on information to provide usable data to teachers about their students' learning needs.

Perhaps the most important benefit to using performance assessments is that they assist in learning and teaching. They are *formative*, in that they

provide teachers and students with the feedback they need from authentic tasks that reveal students' mastery of content, and can guide future teaching. They can also be *summative*, in that they can serve as a final assessment of student capabilities with respect to state and local standards. As summative measures, performance assessments are useful because they organize teaching around the kinds of tasks that support the transfer of learning to new contexts, helping students learn more of what they will need to do in the world outside of school. In addition to acquiring and demonstrating in-depth knowledge of content, this may include the ability to plan an inquiry and organize their time, develop self-discipline and perseverance as well as intellectual discipline, define problems and determine strategies for how to pursue answers, organize and display data, evaluate findings, draw conclusions, and express and defend their ideas according to standards of evidence.

Where and How Performance Assessments Are Used

As noted above, most high-achieving nations and many states in the United States—including Connecticut, Kentucky, Maine, Nebraska, New Hampshire, New Jersey, New York, Rhode Island, Vermont, and Wyoming—have developed and used state and local performance assessments as part of their testing systems. Indeed, the National Science Foundation provided millions of dollars for states to develop hands-on science and math assessments as part of its Systemic Initiative in the 1990s, and prototypes exist all over the country. Additionally, twenty-seven states use multiple approaches for high school graduation decisions, including many that combine state requirements with local performance assessments and other measures (e.g., grades, student work samples, portfolios of work, and senior projects).[8] In this section we briefly describe performance assessment models in both the United States and abroad.

One common model in several U.S. states and in a number of other countries is to combine an external reference exam, which includes open-ended questions that measure aspects of performance such as analysis and expression, with classroom-managed assessments that ask students to tackle more complex, extended tasks that cannot be completed in a couple of hours on a sit-down test. Some states (such as Nebraska, Rhode Island, and Wyoming) and countries (such as Finland, Scotland, and Wales, and

Queensland and ACT, Australia) rely much more heavily on school-based performance assessments. Both approaches are described below.

U.S. examples of performance assessment systems

Connecticut developed a performance task approach during the 1990s as part of its state assessment and accountability system. Connecticut test items include a range of test formats: multiple choice, constructed responses, short essays, mini experiments, and performance tasks to measure how students can apply what they know.[9] Teachers are involved in all areas of test development, including task development, scoring, and standard setting. At the high school level, the Connecticut Academic Performance Test (CAPT), administered in the tenth grade, reports on student performance in four areas: mathematics, reading across the disciplines (focusing on response to literature and reading for information), writing across the disciplines, and science. The CAPT uses classroom-embedded tasks as part of its statewide assessment system. For example, students design and conduct science experiments that are embedded in the science curriculum around a unit of study on specific topics. Students are asked to formulate hypotheses, conduct the experiment, analyze the data, and report their results to prove their ability to engage in scientific reasoning. They also critique experiments and evaluate the soundness of findings and are tested on their findings as part of the CAPT on-demand science assessment.

While the CAPT is required of all public high schools students in Connecticut, the state legislature specifies that the test cannot be used as the sole basis for graduation or promotion. As part of its official policy (2000), the state board of education stated that "the CAPT results alone do not provide a comprehensive picture of student accomplishment. There is a danger that overemphasizing state test scores to evaluate a student's school or district performance can result in an inappropriate narrowing of the curriculum and inappropriate classroom instructional practice."[10] As a consequence, districts are required to use the CAPT assessment in combination with local assessments, which must include performance assessments.

Maine, Vermont, New Hampshire, and Rhode Island also have developed performance assessment components as part of their accountability systems,

but with more participation on the part of the state in helping local districts implement their assessments. These New England states combine a jointly constructed reference exam—the New England Common Assessment Program (NECAP)—with locally developed assessments that provide evidence of student work from performance tasks and portfolios.

Vermont was an early leader, developing in the late 1980s and early '90s both on-demand performance tasks and portfolios that are used throughout the school year, so teachers and students can learn from the results of the assessments and continually improve their work. The writing and mathematics portfolios, developed by the state department of education with the engagement of teachers, include both common tasks to be completed by all students and locally selected work samples that reflect particular kinds of work to be represented in the portfolios.

As the system was phased in, teachers learned how to develop and evaluate assessments and how to teach toward the standards through support networks that sponsored professional development sessions and summer institutes across the state. Teachers from different schools convened to score assessment tasks together, moderating their scoring to gain consistency. While evaluations found that the early, nonstandardized portfolios were not scored very reliably, revisions brought common structures to the portfolios and performance assessments, which resulted in much higher levels of reliability, comparable to those achieved on AP exams.[11]

The state's involvement of large numbers of teachers in designing and scoring the assessments created substantial focus on the quality of student work, providing a powerful form of professional development. Harvard professor Richard Murnane described the conversations of Vermont teachers who gathered in the summer to evaluate portfolios: "Often heated, the discussions focused on what constitutes good communication and problem-solving skills, how first rate work differs from less adequate work, and what types of problems elicit the best student work."[12]

For more than a decade, the Vermont portfolios were the primary assessments for support and accountability in the state. They are now a voluntary adjunct to the annual standardized tests at each grade level

required by NCLB, and many districts and schools continue to use them to obtain a comprehensive assessment of student learning.

Maine's assessment system was designed to include the use of the NECAP reference exam and the Maine Education Assessment, both of which include many open-ended items and a writing assessment, plus locally developed performance assessments. The local assessments are organized around Maine's Learning Results in eight areas (English language arts, mathematics, science, social studies, health/physical education, career preparation, modern and classical language, and visual and performing arts).With extensive professional development provided by the state, local districts developed common performance tasks, classroom-based portfolios, observations, and exhibitions of student work. With the advent of NCLB, which introduced new standardized tests at each grade level, the performance components are now used voluntarily by districts to support instructional decisions.

As part of a state high school redesign initiative, **Rhode Island** has also developed a performance-based graduation system. Starting in 2008, all Rhode Island graduates had to show evidence of success across three elements of the performance-based graduation requirement: a standardized reference exam, course performance, and state-approved performance assessments such as portfolios, senior projects, and/or end-of-course exams. The performance outcomes for each of these data elements must be authentic and aligned to state standards, and must demonstrate meaningful content knowledge. Commissioner Peter McWalters emphasized that there are three non-negotiables in this work: "We have to educate every child; we have to hold high standards; and we have to provide differentiated learning and instruction." In its first year of implementation, the new system was reported to engender greater student engagement and participation in school, with graduation rates increasing from 70 percent to 74 percent, rather than declining, as is frequently the case when new state graduation assessments are introduced.[13]

New Hampshire is moving to a competency-based system for graduation that will no longer use Carnegie units. The state will base graduation on a competency-based credit system using a "mastery of learning" approach to

assess student learning, which relies on performance assessments to evaluate mastery of content and skills and allows students to earn credits both in school and during out-of-school time. The state has already introduced a technology portfolio, which all students must complete to demonstrate their technology competence in high school.

Ohio is also developing a set of standards-based performance tasks measuring core knowledge and skills in the content areas of math, English, science, and history to become part of the state's high school assessment system. These tasks represent the skills of disciplinary inquiry necessary for college readiness and success in the workplace and will support instructional decisions as well as accountability reporting.

Nebraska utilizes a system of performance assessments created and scored by local educators trained to score reliably. These systems are peer reviewed by measurement and assessment experts and include a check on validity through the use of a statewide writing examination and the administration of one norm-referenced test. **Wyoming** uses a "body of evidence" approach that is locally developed in order to determine whether students have mastered standards required for graduation. **Oregon** uses both online diagnostic assessments and performance assessments in multiple subject areas that are state designed and locally scored using state rubrics as the basis for a Certificate of Mastery.

Some well-developed performance assessment systems were created and are used by consortia of local schools and/or districts. In New York, for example, the **New York Performance Assessment Consortium** is a network of forty-seven schools in the state that rely upon a set of performance tasks assembled in a portfolio to determine graduation. These include a major task in each disciplinary area: a scientific investigation, a historical research paper, a literary response, an applied mathematical problem or model, an arts exhibition, and an analysis of an internship experience. These are defended before a panel that includes outside experts as well as teachers and parents and scored according to common rubrics. Because of the quality of their work, the consortium schools have a state waiver from some of the Regents Examinations. Research shows that New York City students who graduate from these schools (which have a much higher graduation rate

than the City even though they serve more low-income students, students of color, and recent immigrants) are more successful in college than students with a traditional Regents diploma, which relies upon standardized tests.

Among other notable performance-based systems under development nationally is the **College Readiness Performance Assessment System (C-PAS)**, developed by David Conley at the University of Oregon. The C-PAS is designed to track the development of five generic cognitive strategies that represent the thinking skills necessary for college readiness and success: problem solving, research, interpretation, reasoning, and precision. The C-PAS assessment is a series of performance tasks that teachers administer and score with a common scoring guide.

The **Collegiate Learning Assessment (CLA)**, developed by Richard Shavelson at Stanford University, Stephen Klein at RAND, and colleagues, is a collegiate assessment that is being adapted for secondary schools. The CLA uses real-world performance tasks that elicit critical thinking, analytic reasoning, problem solving, and communication skills. Students are typically faced with a problem that requires them to collect and evaluate evidence, then frame and defend a solution. They may use a variety of documents and resources provided in an "in-basket" to learn about aspects of the problem that is posed. The CLA uses a matrix sampling approach to assess student performance at the beginning and the end of college (not all students perform all tasks), and develops institutional reports focusing on students' college-level competencies.

Performance assessments abroad

School-based performance assessment is the dominant mode of assessment in most high-achieving countries.[14] (See Table 1.) At the high school level, a number of countries use a blended approach that combines school-based tasks that measure specific subject-matter concepts and skills with a common examination, often developed by teachers in collaboration with university faculty, featuring primarily open-ended questions requiring written or oral responses.

Table 1: Summary of International Assessment Systems Using Performance Assessment

Country/ organization	What assessments are used?	Who grades the assessments?	Who designs the assessments?	How are the results used?
Victoria, Australia	**School-based** •Projects, labs, papers, essays, presentations	**School-based** •50%+ of grade •Graded by the teacher	**School-based** •Teacher designed based on state syllabi curriculum	Use scores to guide admission to a university and workplace apprenticeship programs
	National •Multiple choice, short answer, essays, oral exams	**National** •Administered by the teacher •Graded by the teacher	**National** •Designed by teachers, professors through VCAA	
Sweden	**School-based** •Coursework, projects, essays, test	**School-based** •To 30% of grade •Graded by the teacher	**School-based** •Teacher designed based on national curriculum	Uses scores to compare coursework grades and local assessment results to national standards
	National •National syllabi, open-ended questions, material given in advance	**National** •Included in grading, but not the sole factor in grading	**National** •Educational research institution designed, teacher input	
Finland	**National** •Short problems that ask students to apply their thinking	**National** •Graded by teachers and re-checked by the board of education	**National** •Originally developed by University of Helsinki	Uses scores to inform instruction and student self-reflection, and in some cases used for placement in a university
	School-based •Presentations, plays, demos	**School-based** •Graded by the teacher	**School-based** •Teachers design with national themes	

United Kingdom: England	**School-based** • Coursework, tests, projects, essays	**School-based** • To 30% of grade • Graded by the teacher	**School-based** • Teacher designed based on national curriculum	Use scores to select upper-secondary coursework and gain admission to a university
	National • Essays	**National** • To 80% of grade • By exam group	**National** • Designed by examining group	
United Kingdom: Wales	**School-based** • Student investigations, presentations	**School-based** • Graded by the teacher	**School-based** • Teacher designed based on national curriculum	Reported to parents and government, meant to encourage better teaching, more student engagement
	National • Essays (only upper secondary)	**National** • By exam group	**National** • Designed by examining group	
International Baccalaureate	**School-based** • Speeches, projects, portfolio, presents, investigates, labs	**School-based** • 20–50% of grade • Graded by the teacher	**School-based** • Designed by the classroom teacher	Used for awarding IB Diploma; giving college credit in some cases
	National • Multiple choice, essay, short answer	**National** • Administered and graded by IB	**External** • Designed by IB	
Hong Kong	**School-based** • Closely aligned with national assessments	**School-based** • Graded by the teacher	**School-based** • Designed by the classroom teacher	Uses scores to judge whether students may advance to the next level
	National • Projects, portfolio, observations, exam	**National** • Graded by the teacher	**National** • Designed by teachers	

The **International Baccalaureate** (**IB**) program, which enrolls 650,000 worldwide, including a growing number in the United States, exemplifies the syllabus-based approach to classroom assessment used in many countries in Europe and Asia. Designed for students in grades eleven and twelve, it assesses student learning using school-based performance assessments and external exams at the end of each course. Both types of assessments measure students' performance on the objectives specified in the "subject outlines" written by the IB organization. School-based performance assessments —such as oral exercises in language subjects, projects, student portfolios, practical laboratory work, mathematical investigations, and artistic performances—contribute 30 to 50 percent of the final examination grade. The external exam consists largely of essays, constructed responses, and data response questions, case study questions, and text response questions, with a limited use of multiple-choice items. A typical essay question students might choose among several options on the exam would be the following:

> Acquiring material wealth or rejecting its attractions has often been the base upon which writers have developed interesting plots. Compare the ways the writers of two or three works you have studied have developed such motivations.

This blended approach also characterizes the **General Certificate of Secondary Education** (**GCSE**) examinations in **Great Britain**, as well as the high school examinations in **Finland, Sweden, Hong Kong, Singapore,** and **Victoria, Australia**. (Most of these countries now use primarily local performance assessments in the elementary and middle school years.) The school-based performance assessments typically comprise from 30 to 50 percent of the total examination score in these assessment systems.

In **Sweden**, schools offer nationally approved examinations in the upper-secondary years in several subjects.[15] Teachers work with university faculty to help design the tasks and questions, and they weight information from these exams, their own assessments, and classroom work to assign a grade reflecting how well students have met the objectives of the syllabus.[16] Regional education officials and schools provide time for teachers to calibrate their grading practices to minimize variation across the schools and across the region.[17] Toward the end of their upper-secondary schooling,

Swedish students receive a final grade or "learning certificate" in each area that acts as a compilation of all of these sources of evidence, including projects completed by the student as well as grades awarded for courses.

In **Victoria, Australia**, the Victoria Curriculum and Assessment Authority (VCAA) establishes courses in a wide range of studies, develops the external examinations, and ensures the quality of the school-assessed component of the Victorian Certificate of Education(VCE). The VCAA conceptualizes assessment as "of," "for," and "as" learning. Teachers are involved in developing assessments, along with university faculty in the subject area, and all prior-year assessments are public, in an attempt to make the standards and means of measuring them as transparent as possible. Before the external examinations are given to students, teachers and academics take the exams themselves, as if they were students. The external subject-specific examinations, given in grades eleven and twelve, include written, oral, and performance elements scored by classroom teachers.

In addition, at least 50 percent of the total examination score is comprised of classroom-based tasks that are given throughout the school year. These required assignments and assessments—lab experiments and investigations on central topics as well as research papers and presentations—are designed by teachers in response to syllabus expectations. These required classroom tasks ensure that students are getting the kind of learning opportunities that prepare them for the assessments they will later take, that they are getting feedback they need to improve, and that they will be prepared to succeed not only on these very challenging tests but in college and in life, where they will have to apply knowledge in these ways.

As in Victoria, assessments in **Great Britain** use a combination of external and school-based tasks based on the national curriculum and course syllabi. Throughout the school years, classroom-based tasks scored by teachers are used to evaluate student achievement of curriculum goals. A mandatory set of assessments at year nine (age fourteen) includes both teacher-created and -administered assessments and, for students who have reached a certain level of achievement, national exams and tasks.[18]

While not mandatory, most students take a set of exams at grade eleven (age sixteen) to achieve their GCSE. Students choose which tests they will take based on their interests and areas of expertise. Most GCSE items are essay questions. The math exam includes questions that ask students to show the reasoning behind their answers, and foreign-language exams require oral presentations. About 25 to 30 percent of the final examination score is based on class work, coursework, and assessments developed and graded by teachers. In many subjects, students also complete a project worked on in class that is specified in the syllabus.

Hong Kong has typically used the British A- and O-Level exams for students in high school. In collaboration with educators from Australia, the United Kingdom, and other nations, Hong Kong's assessment system is evolving from a centralized examination structure to one that increasingly emphasizes school-based formative assessments that expect students to analyze issues and solve problems. The government has decided to gradually replace the Hong Kong Certificate of Education Examinations, which most students sit for at the end of their five-year secondary education, with a new Hong Kong Diploma of Secondary Education that will feature school-based assessments.

In addition, the Hong Kong Territory-wide System Assessment (TSA), which assesses lower-grade student performance in Chinese, English, and mathematics, is developing an online bank of assessment tasks to enable schools to assess their students and receive feedback on their performance on their own timeframes. The formal TSA assessments, which include both written and oral components, occur at primary grades three and six and secondary grade three (the equivalent of grade nine in the United States).

As outlined in Hong Kong's "Learning to Learn" reform plan, the goal of the reforms is to shape curriculum and instruction around critical thinking, problem solving, self-management skills, and collaboration. A particular concern is to develop metacognitive thinking skills, so students may identify their strengths and areas needing additional work.[19] By 2007, Curriculum and Assessment Guides were published for four core subjects and twenty elective subjects, and assessments in the first two subjects—Chinese language and English language—were revised. These became criterion-

referenced, performance-based assessments featuring not only the kinds of essays previously used on the exams, but also new speaking and listening components, the composition of written papers testing integrated skills, and a school-based component that factors into the examination score. Although the existing assessments already use open-ended responses, the proportion of such responses will increase in the revised assessments. Like the existing assessments, the new assessments are developed by teachers with the participation of higher education faculty, and they are scored by teachers who are trained as assessors.

In **Queensland, Australia**, there has been no assessment system external to schools for forty years. Until the early 1970s, a traditional "postcolonial" examination system controlled the curriculum. When it was eliminated—about the same time as in Finland and Sweden—all assessments became school based. School-based assessments are developed, administered, and scored by teachers in relation to the national curriculum guidelines and state syllabi (also developed by teachers), and are moderated by panels that include teachers from other schools as well as professors from the university system.

The syllabi spell out a small number of key concepts and/or skills to be learned in each course, and what kinds of projects or activities (including minimum assessment requirements) students should be engaged in. Each school designs its program to fit the needs and experiences of its own students, choosing specific texts and topics with this in mind. At the end of the year, teachers collect a portfolio of each student's work, which includes the specific assessment tasks, and grade it on a five-point grading scale. To calibrate these grades, teachers put together a selection of portfolios from each grade level—one from each of the five score levels, plus borderline cases—and send these to a regional panel for moderation. A panel of five teachers re-scores the portfolios and confers about whether the grade is warranted, making a judgment on the spread. A state panel also looks at portfolios across schools. Based on these moderation processes, the school is given instructions to adjust grades so they are comparable to others.

Summary

The use of curriculum-embedded assessments in these performance-based systems allows for the testing of more complex skills that cannot be measured in a two-hour test on a single day. They shape the curriculum in ways that ensure stronger learning opportunities. They give teachers timely, formative information they need to help students improve—something that standardized examinations with long lapses between administration and results cannot do. And they help teachers become more knowledgeable about content standards and how to teach to them, as well as about their own students and how they learn. The process of using these assessments can improve their teaching and their students' learning. The processes of collective scoring and moderation that many nations use to ensure reliability in scoring also prove educative for teachers, who learn to calibrate their understanding of the standards to common benchmarks. In these ways, as part of a balanced assessment approach, performance assessments can help ensure that students are ready for college and the workplace.

Challenges and Considerations in Scaling Up Performance Assessments

From the research and evidence on performance assessment, there are a number of lessons learned that should be considered when designing a system that substantially incorporates performance-based assessments.

Calibration of scoring: Perhaps the most complex question surrounding these assessments when they are locally developed or scored is how to ensure comparability. Many of the systems described above, both in the United States and abroad, use common scoring guides and extensive scorer training to achieve consistency in the use of these rubrics. In addition, they use auditing, moderation, and calibration systems of several kinds to maintain the quality of the system over time.

In Victoria, Australia, the quality of the tasks assigned by teachers, the work done by students, and the appropriateness of the grades and feedback given to students are audited through an inspection system, and schools are given feedback on all of these elements. In addition, the VCAA uses statistical moderation to ensure that the same assessment standards are applied to

students across schools. The external exams are used as the basis for this moderation, which adjusts the level and spread of each school's assessments of its students to match the level and spread of the same students' scores on the common external test score. The result is a rich curriculum for students with extensive teacher participation and a comparable means for examining student learning.

In Hong Kong, tests are allocated randomly to scorers, and essay responses are typically rated by two independent scorers.[20] Results of the new school-based assessments are statistically moderated to ensure comparability within the province. The assessments are internationally benchmarked, through the evaluation of sample student papers, to peg the results to those in other countries. Many of the new assessments are also to be scored online, which the Examinations Authority notes is now the common practice in twenty of China's mainland provinces, as well as in the United Kingdom.

Queensland's system, like those in a number of countries, also employs "moderation," a process of bringing samples from different schools to be re-scored, with results sent back to the originating schools. This process leads to stronger comparability across schools and is part of building a strong performance assessment system. Nebraska also supplements extensive scorer training on common rubrics with external validation checks such as comparisons with the statewide writing assessment, the ACT, and other commonly administered standardized tests. Each district's assessment system is evaluated and approved through a review process conducted by measurement experts.

Costs and scoring models: Appropriate, affordable, and educationally supportive scoring models must be developed. Although some methods of managing performance assessments can cost more than machine scoring of multiple-choice tests (i.e., when such assessments are treated as traditional external tests and shipped out to separately paid scorers), the cost calculus changes when assessment is understood as part of teachers' work and learning—built into teaching and professional development time. Much evidence suggests that developing and scoring these assessments is a high-yield investment in teacher learning and a good use of professional development resources.

In most European and Asian systems, and in those used in several U.S. states, scoring of assessments is conducted by teachers and time is set aside for this aspect of teachers' work and learning. While teacher time to create and score the assessments can be substantial, these activities lead to more skilled and engaged teachers. In contrast, most external standardized tests provide teachers with little guidance on how to improve student learning, since they simply receive numerical scores on secret tests months after the students have left school. Hence the professional development that seeks to help teachers improve achievement in this system is less well informed and less effective.

Professional development: Extensive professional development is necessary for educators to learn to build, use, and score assessments that will inform and guide their teaching. Many systems have demonstrated that teachers can develop this knowledge rapidly when given the support. In successful systems, teachers are engaged in curriculum alignment, performance task development, scoring processes, and data analysis so that they understand the system and can teach productively to the standards. The processes include a peer review or moderation system that provides a feedback loop, checks on quality, and directions for staff development. Teachers often report that some of the best professional development of their careers occurs when they have opportunities to examine, score, and discuss student work. Importantly, international assessments have strategically "captured" teacher professional development time to evaluate and validate student work. Capitalizing on this time can both lower costs and establish a common language around curriculum standards and assessment.

Administrative support: Education agency officials and legislators at the state and federal levels must develop targeted assistance to teachers, administrators, and school systems that allows their effective participation in these systems and leverages improvements in teaching. In addition to professional development, this will include widespread information, extensive training in both use and scoring, the redesign of curriculum materials to ensure alignment with and support of new assessments, and the redesign of school schedules to provide in-class time for more in-depth work on the part of students and out-of-class time for teachers' planning, analysis, and scoring of student work, as is common in other countries.

Quality of tasks: Careful attention must be paid to the quality of performance tasks. They should be developed around important disciplinary content so that they measure core concepts and abilities with strong validity, and they should be developed in response to criteria that establish the technical quality of assessments (including checking for bias and fairness), high proficiency standards, consistent administration of assessment (including clear criteria that would certify the quality of an assessment task), and opportunity to learn what is assessed. They should also be constructed to allow students with special needs and those who are learning English opportunities to demonstrate their knowledge appropriately.

Proper use: Productive use of performance assessments, like proper use of standardized tests, should be aimed at revealing areas needing improvement and should lead to curriculum and professional learning supports that can result in powerful learning outcomes for all students. Additionally, tools and protocols, including technological tools, are needed to support the design and use of performance assessment. For example, tools such as task blueprints, rubric specifications, and training and scoring protocols should be developed to support proper use of performance assessments. Finally, as other countries have found, using assessments for information rather than sanctions allows the development of more ambitious tasks aimed at higher standards, and less corruption of the assessment system. This framework for assessment has driven stronger learning and higher achievement in many nations abroad.

Many nations have developed strategies to monitor and improve assessment quality. In Hong Kong, for example, to guide the process of assessment reform, the Education Bureau has implemented a School Development and Accountability Framework which emphasizes school self-evaluation, plus external peer evaluation, using a set of performance indicators. The bureau promotes the use of multiple forms of assessment in schools including projects, portfolios, observations, and examinations, and looks for the variety of assessments in the performance indicators used for school evaluation.[21] For example, the performance indicators ask, "Is the school able to adopt varied modes of assessment and effectively assess students' performance in respect of knowledge, skills, and attitude?" and "How does the school make use of curriculum evaluation data to inform curriculum

planning?"[22] This practice of examining school practices and the quality of assessments through an inspection or peer review process is also used in Australia and Great Britain to improve teaching by using standards as a tool for sharing knowledge and reflecting on practice.

Federal Policy Recommendations

Performance assessment is a key component in a balanced assessment system that responds to fast-paced changes placing greater demands on education and knowledge development in the United States and around the world. Images of what students will need to *do* with their knowledge should help shape formulations of curriculum, instruction, and assessment policy at the national, state, and local levels. As a starting point for the development of the next generation of assessments, we must begin with a vision of our young people as lifelong learners who deeply understand core concepts and modes of inquiry within the disciplines, and who can also work across disciplines to evaluate evidence, frame and solve problems, express and defend their ideas, and create new ideas, technologies, and solutions.

Many efforts are under way to refine standards for learning at the state level and by consortia of states collaborating under the auspices of the Council for Chief State School Officers and Achieve, Inc., a national organization of governors, business leaders, and education leaders. These efforts seek to ensure that standards are internationally benchmarked and are fewer, higher, and deeper. It is critical that new assessments be developed in the context of new standards and in relation to curriculum frameworks that ensure the content and skills can be taught coherently and well. To accomplish this federal policy should:

Fund an intensive development effort that enables states and consortia of states, in collaboration with development experts in federal labs, centers, nonprofit organizations, and universities, to

- develop, validate, and test high-quality performance assessments that are part of balanced assessment systems which are guided by thoughtful, coherent standards and curriculum frameworks;
- train the field of practitioners—ranging from psychometricians to a new generation of state and local curriculum and assessment

specialists to teachers—who can be skillfully involved in the development, administration, and scoring of these assessments in valid and reliable ways; and
- conduct high-quality research on the validity, reliability, instructional consequences, and equity consequences of these assessments.

Encourage improvements in federal, state, and local assessment practice in the following ways:

- Provide incentives and funding for states to refine their existing state assessments and introduce related high-quality locally administered performance assessments that evaluate critical thinking and applied skills. Support states in making such assessments reliable, valid, and practically feasible through teacher professional development and scorer training, moderated and audited scoring systems, and calibration systems, as well as research.

- As part of these efforts, develop more appropriate assessments and accommodations for special education students and English language learners by underwriting efforts to strengthen the validity and reliability of existing performance assessments for these populations, properly adapt new assessments under construction, and create, as needed, new assessments of performance in the content areas for these students, based on professional testing standards that consider principles of universal design as well as specific needs for valid assessment of students in these groups.

- To model high-quality items and better measure the standards, support the further development and implementation of the new blueprints, already under way, for the National Assessment of Educational Practice (NAEP), which include more performance-oriented items that evaluate students' abilities to evaluate evidence, solve problems, and explain and defend their ideas. These kinds of tasks were part of NAEP when it was first launched in the 1960s, and are common in other nations' large-scale assessments, as well as in PISA. Their introduction would need to be incorporated carefully

over time in a planful fashion that maintains existing trend data and continues to enable comparisons among states and over time.

Enable the incorporation of new assessments into the NCLB accountability system in the following ways:

- Replace the current "status model" for measuring school progress with a Continuous Progress Index that sets expectations for schools—and groups of students within them—to show progress on an index of measures that include multiple assessments of student learning, including performance measures, as well as school progression and graduation rates. In such an index, which reports information on multiple indicators and then combines them for tracking overall progress, states could choose to include subject areas beyond reading and mathematics, such as writing, science, and history—which are important in their own right and essential to encourage and evaluate students' literacy skills as they are applied in the content areas. Within a given subject, the index could accommodate assessments of student learning that capture a wider array of skills—including the more complex inquiry and problem-solving skills demanded by twenty-first-century jobs and colleges. Such an index would reduce incentives to narrow the curriculum. It would evaluate students' growth over time across the entire achievement continuum, thus focusing attention on progress in all students' learning, not just those who fall at the so-called "proficiency bubble," reducing ceiling effects, and recognizing schools' gains with students who score well below and above a single cut score. The CPI would also encourage greater inclusion and more appropriate measurement of gains for special education students and English language learners by tracking gains at all points along the continuum and by incorporating the results of appropriate measures.

Conclusion

Current accountability reforms are based on the idea that standards can serve as a catalyst for states to be explicit about learning goals, and the act of measuring progress toward meeting these standards is an important force

toward developing high levels of achievement for all students. However, an on-demand test taken in a limited period of time on a single day cannot measure all that is important for students to know and be able to do. As described by Achieve, Inc., the limitation of traditional on-demand tests is that they cannot measure many of the skills that matter most for success in the worlds of work and higher education:

> States … will need to move beyond large-scale assessments because, as critical as they are, they cannot measure everything that matters in a young person's education. The ability to make effective oral arguments and conduct significant research projects are considered essential skills by both employers and postsecondary educators, but these skills are very difficult to assess on a paper-and pencil test.[23]

Balanced systems of assessment that include performance assessments have the potential to strengthen curriculum and instruction by evaluating the full range of standards in valid and appropriate ways, providing rich information about student learning that is useful to classroom teachers, and providing diverse means for students to demonstrate their learning. Developed carefully and used properly, such assessments can stimulate more thoughtful teaching, become an engine for ongoing improvement and professional development, and create a commitment to standards that shape more powerful learning.

The views expressed in this chapter are those of the authors and do not necessarily represent those of the Alliance for Excellent Education.

About the Authors

Linda Darling-Hammond is Charles E. Ducommun Professor of Education at Stanford University, where she has launched the School Redesign Network and the Stanford Center for Opportunity Policy in Education (SCOPE). Her research, teaching, and policy work focus on issues of school restructuring, teacher quality, and educational equity. Dr. Darling-Hammond has served as faculty sponsor for the Stanford Teacher Education Program, where she helped to introduce performance-based portfolio assessments for pre-service teachers, and cofounded the Performance Assessment for California Teachers with colleagues from eleven other universities. Previously, she served on the

National Board for Professional Teaching Standards and supported its development of a new model of performance assessment for accomplished teachers, and she chaired the standards drafting committee of the Interstate New Teacher Assessment and Support Consortium as it developed new standards for beginning teachers and piloted performance assessments to evaluate the standards. She also chaired the New York State Council on Curriculum and Assessment as it redesigned the state standards and introduced new performance elements to the Regents testing system. Dr. Darling-Hammond is a former president of the American Educational Research Association and a member of the National Academy of Education. Among her more than three hundred publications are *Preparing Teachers for a Changing World: What Teachers Should Learn and Be Able to Do, The Right to Learn: A Blueprint for Schools That Work,* and *Authentic Assessment in Action: Studies of Schools and Students at Work.*

Raymond L. Pecheone is the co–executive director of the Stanford School Redesign Network LEADS Network. LEADS is an executive educational leadership program that builds partnerships between schools of business and education to bring interdisciplinary perspectives and knowledge bases to the work of K–12 district and school reformers. Dr. Pecheone also serves as the director of the Performance Assessment for California Teachers (PACT) program. PACT is a consortium of thirty-two California universities that have joined together to develop a reliable and valid licensure assessment of pre-service teaching. He also leads and directs a performance-based student assessment project that is aligned to college- and workplace-readiness skills and includes work in California (the Stanford Bay Area Assessment Consortium), and nationally in partnerships with the Asia Society and the State of Ohio. Prior to Stanford, Dr. Pecheone was the Connecticut bureau chief for curriculum, research, and assessment. In this role, he directed the First Assessment Development Laboratory for the National Board for Professional Teaching Standards and cofounded the Interstate New Teacher Assessment and Support Consortium, housed at the Council of Chief State School Officers. He supported the redesign of New York State's Regents Examinations, and served as a consultant to ETS in the development and validation of a national performance-based assessment test for school administrators.

---

[1] No Child Left Behind Act of 2001, Public Law 107-110, 107th Cong., 1st sess., Sec. 1111, b, I, vi.

[2] This section draws from L. Darling-Hammond and G. H. Wood, *Refocusing Accountability: Using Performance Assessments to Enhance Teaching and Learning* (Washington, DC: Forum for Education & Democracy, 2008).

[3] S. Messick, "Validity," in *Educational Measurement*, ed. R. L. Linn, 13–103 (Washington, DC: National Council of Measurement in Education and the American Council on Measurement in Education, 1989).

[4] L. A. Shepard, R. J. Flexer, E. H. Heibert, S. F. Marion, V. Mayfield, and T. J. Weston, *Effects of Introducing Classroom Performance Assessments on Student Learning*, CSE Technical Report 394 (Los Angeles: National Center for Research on Evaluation, Standards, and Students Testing [CRESST], Graduate School of Education & Information Studies, University of California, Los Angeles, 1995).

[5] For a summary see L. Darling-Hammond and E. Rustique-Forrester, "The Consequences of Student Testing for Teaching and Teacher Quality," in *The Uses and Misuses of Data in Accountability Testing*, ed. Joan Herman and Edward Haertel, 289–319 (Malden, MA: Blackwell Publishing, 2005).

[6] Appalachia Educational Laboratory, "Five Years of Reform in Rural Kentucky," *Notes from the Field: Educational Reform in Rural Kentucky 5*, no. 1 (February 1996); Charleston, WV: Author; B. M. Stecher, S. Barron, T. Kaganoff, and J. Goodwin, *The Effects of Standards-Based Assessment on Classroom Practices: Results of the 1996–97 RAND Survey of Kentucky Teachers of Mathematics and Writing*, CSE Technical Report (Los Angeles: UCLA National Center for Research on Evaluation, Standards, and Student Testing, 1998); B. L. Whitford and K. Jones, *Accountability, Assessment, and Teacher Commitment: Lessons from Kentucky's Reform Efforts* (Albany: State University of New York, 2000); D. Koretz, B. Stetcher, and E. Deibert, *The Vermont Portfolio Program: Interim Report on Implementation and Impact, 1991–92 School Year* (Santa Monica, CA: RAND, 1992); Shepard et al., *Effects of Introducing Classroom Performance Assessments*.

[7] L. Darling-Hammond, J. Ancess, and B. Falk, *Authentic Assessment in Action* (New York: Teachers College Press, 1995); M. Kornhaber and H. Gardner, *Varieties of Student Excellence* (New York: National Center for Restructuring Education, Schools, and Teaching, Teachers College, Columbia University, 1993).

[8] L. Darling-Hammond, E. Rustique-Forrester, and R. Pecheone, *Multiple Measures Approaches to High School Graduation* (Stanford: Stanford University, School Redesign Network, 2005).

[9] R. Mitchell, *Testing for Learning* (New York: Free Press, 1992).

[10] Connecticut State Board of Education, 2000.

[11] Koretz, Stetcher, and Deibert, *The Vermont Portfolio Program*.

[12] R. Murnane and F. Levy, *Teaching the New Basic Skills* (New York: Free Press, 1996).

[13] J. D. Jordan, "R.I. Graduation Rate Up: What Happened to the 13,163 Ninth Graders of 2004?" *Providence Journal* (March 21, 2009).

[14] This section draws on L. Darling-Hammond and L. McCloskey, "Assessment for Learning Around the World: What Would It Mean to Be Internationally Competitive?'" *Phi Delta Kappan* 90, no. 4 (2008): 263.

[15] Swedish National Agency for Education, *The Swedish School System: Compulsory School*, 2005, http://www.skolverket.se/sb/d/354/a/959 (accessed May 31, 2008).

[16] M. A. Eckstein and H. J. Noah, *Secondary School Examinations: International Perspectives on Policies and Practice* (New Haven: Yale University Press, 1993); S. O'Donnell, *International Review of Curriculum and Assessment Frameworks, Comparative Tables and Factual Summaries—2004* (London: Qualifications and Curriculum Authority, 2004).

[17] Eckstein and Noah, *Secondary School Examinations*, p. 230.

[18] Qualifications and Curriculum Authority, "England: Assessment Arrangements," 2008, http://www.inca.org.uk/1315 (accessed May 27, 2008); http://education.qld.gov.au/corporate/newbasics/html/richtasks/richtasks.html (accessed April 1, 2009).

[19] J. K. Chan, K. J. Kennedy, F. W. Yu, and P. Fok, "Assessment Policy in Hong Kong: Implementation Issues for New Forms of Assessment," *Hong Kong Institute of Education*, 2008, http://www.iaea.info/papers.aspx?id=68 (accessed September 12, 2008).

[20] M. Dowling, "Examining the Exams," http://www.hkeaa.edu.hk/files/pdf/markdowling_e.pdf (accessed September 14, 2008).

[21] Chan et al., "Assessment Policy in Hong Kong"; "Performance Indicators for Hong Kong Schools, 2008 with Evidence of Performance," 2008, http://www.edb.gov.hk/FileManager/EN/Content_6456/pi2008%20eng%205_5.pdf (accessed September 12, 2008).

[22] Quality Assurance Division of the Education Bureau, "Performance Indicators for Hong Kong Schools."

[23] Achieve, Inc., *Do Graduation Tests Measure Up? A Closer Look at State High School Exit Exams*, Executive Summary (Washington, DC: Achieve, Inc., 2004).

CHAPTER 3

Formative Assessment and Assessment *for* Learning

Jan Chappuis, Stephen Chappuis, and Richard Stiggins
ETS Assessment Training Institute

As our nation seeks to improve learning for all students, there is increased demand for assessment information to use in data-driven decisionmaking at all levels of the education system—state, district, and classroom. Educators and policymakers are now called upon to establish balanced assessment systems designed to meet the information needs at each level. Such a system includes *annual assessments* designed to allow schools, school systems, and communities to judge the impact of the educational experiences on student learning, for accountability and other purposes. It includes *interim assessments* used across classrooms to identify standards that students are struggling to master and to provide a focus for instructional and program improvement. And it includes *classroom assessments* designed both to support student learning and to measure their achievement. Within this balanced assessment system are classroom assessment practices that inform daily instruction, known as *formative assessment*, whose purpose is to provide the detailed achievement information that teachers and students can act on every day to improve learning.

In 1998, British researchers Paul Black and Dylan Wiliam published a comprehensive review of research on formative assessment practices, in which they concluded, "Innovations that include strengthening the practice of formative assessment produce significant and often substantial learning gains."[1] Their review examined studies that collectively encompassed kindergartners to college students; represented a range of subject areas including reading, writing, social studies, mathematics, and science; and were conducted in numerous countries throughout the world, including the United States. The gains reported in the studies they describe are among the largest found for any educational intervention: the achievement gains realized by students whose teachers relied on formative assessment practices ranged from 15 to 25 percentile points, or two to four grade equivalents, on commonly used standardized achievement test score scales. In broader terms, this kind of score gain, if applied to performance on international assessments, would move the United States's rank from the middle of the pack of the forty-two nations tested to the top five. An additional outcome common among the studies they analyzed is that certain formative assessment practices greatly increased the achievement of low-performing students, in some cases to the point of approaching that of high-achieving students.

Black and Wiliam's report in large part triggered the current widespread interest in formative assessment: over the last ten years educators and policymakers alike have become aware of the need to support its use. But along with increased awareness has come increased confusion about what "counts" as formative assessment and how to develop educators' capacity to use assessment formatively. This chapter will describe the characteristics of formative assessment, with a particular focus on those formative assessment practices that engage and empower students in their own learning, or *assessments for learning*. It will also describe challenges related to the effective use of formative assessment and recommended actions for policymakers.

Confusion about the Meaning of "Formative Assessment"

Recently a school leader asked an assessment expert for an example of a good test item on a formative assessment and then for an example of how that item might look when used on a summative test. He wanted to explain to his staff the difference between formative and summative assessment.

His end goal was for teachers to develop assessments to measure how well students were mastering the content standards on the state accountability test before the test was given in the spring. But he knew that formative assessment had been shown to improve achievement, so he wanted to make sure they were using formative items.

His question reflects the uncertainty with which many educators, school leaders, and policymakers approach formative assessment. It isn't surprising: the assessment landscape is broad and populated with multiple, sometimes conflicting definitions of formative assessment. As a result, practices labeled as formative assessment in schools today vary widely.

What is formative assessment?

It is helpful to begin with an understanding of what is and what isn't formative assessment. For many experts in the field, formative assessment is not an *instrument* or an *event*, but a collection of practices with a common feature: *they all lead to some action that improves learning*. Well-known educational researchers emphasize this point when they describe what is at the heart of formative assessment:

- "Formative assessment, therefore, is essentially feedback … both to the teachers and to the pupil about present understanding and skill development in order to determine the way forward."[2]

- "[Formative assessment] refers to assessment that is specifically intended to provide feedback on performance to improve and accelerate learning."[3]

- "Formative assessment is defined as assessment carried out during the instructional process for the purpose of improving teaching or learning …What makes formative assessment formative is that it is immediately used to make adjustments so as to form new learning."[4]

The Council of Chief State School Officers, as part of its advocacy for formative assessment, has developed the following definition: "Formative assessment is a process used by teachers and students during instruction

that provides feedback to adjust ongoing teaching and learning to improve students' achievement of intended instructional outcomes."[5]

The common thread woven throughout formative assessment research, articles, definitions, and books bears repeating: it is *not the instrument* that is formative; it is the *use of the information* gathered, by whatever means, *to adjust teaching and learning*, that merits the "formative" label.

At the classroom level, teachers assess formally through tests, quizzes, assignments, performances, projects, and surveys, or informally through questioning and dialogue, observing, and anecdotal note taking. In any of these instances, they may or may not be engaged in formative assessment: the determining factor is not the type of assessment they use, but rather how they and their students use the information.

What is summative assessment?

When the information from an assessment is used solely to make a judgment about the level of competence or achievement, it is a *summative assessment*. In the classroom, an assessment is summative when it is given to determine how much students have learned at a particular point in time, for the purpose of communicating achievement status to others. The communication usually takes the form of a symbol, a letter grade, a number, or a comparison to a standard such as "meets the standard" or "proficient" that is reported to students and eventually to parents.

At the program level, an assessment is summative when results are used to make judgments such as determining how many students are and are not meeting standards in a certain subject, or to evaluate the effectiveness of a particular curriculum or instructional model. The data may be reported to educators within the system, the school board, and the community.

Summative assessments aren't bad or wrong; they're just not formative. They have a different purpose: to report out level of achievement. Mislabeling them as formative, or using summative assessment information in formative ways, will not generate the achievement gains realized in formative assessment research studies.

The growing field of "formative assessment"

Not surprisingly, a plethora of programs and products described as "formative" assessment has surfaced, due in part to the achievement gains and gap-closing powers reported by Black and Wiliam and other researchers.

One cause of growth in this segment of the assessment field is an indirect result of implementation of the No Child Left Behind (NCLB) legislation of 2001. Recent years have seen a dramatic increase in quantity and frequency of student testing—much of it voluntary and well beyond the requirements of federal law or state assessment systems. For example, many schools and districts administer *benchmark, short-cycle,* or *interim assessments* to predict student performance on high-stakes tests, to identify students needing additional help, and to isolate those standards students struggle with most. This use of testing has contributed to the widening scope of what is loosely called formative assessment.

Additionally, testing companies in the K–12 education market, seeking to support and profit from the trend toward more testing, sometimes advertise products as "formative assessments." As a result, the adjective *formative* now appears in the titles of many commercially prepared tests and item banks. This adds to the confusion by implying that it is the test itself that is formative.[6] In reality, these off-the-shelf assessments may be little more than a series of mini summative tests, not always tightly aligned to what was taught in the classroom.

These developments in the assessment field have implications for those seeking to support formative assessment practices. Are all of the tests and practices labeled as "formative" truly formative? Most importantly, what is it about *formative* that gives it its power? What led to the gains these researchers uncovered?

Importance of Assessment Purpose: The Use of Results

Almost any assessment instrument or event can be used for summative or formative purposes. But some assessments are *by design* better suited to summative use and others to formative use. For example, state assessments, although they may have some limited formative use, are constructed to

provide accountability data and to compare schools and districts. Because their primary purpose is summative, by design the results often do not communicate in detail about individual student strengths and weaknesses. Further, the results are often delivered months after the administration of the tests. Therefore, such state tests usually do not function well in a formative way: they are of limited use diagnostically because they cannot contribute meaningfully to guide day-to-day instruction or help determine the next learning steps of the individual students who generated the data.

Benchmark assessments, either purchased by the district from commercial vendors or developed locally, are generally meant to measure progress toward state or district content standards and to predict future performance on large-scale summative tests. Such assessments are sometimes intended for formative use, to guide further instruction for groups or individual students, but teachers' and administrators' lack of understanding of how to use the results can derail this intention. The assessments will produce no formative benefits if teachers administer them, report the results, and then continue with instruction as previously planned—as can easily happen when teachers are expected to cover a hefty amount of content in a given time.

Teachers also select or develop their own summative assessments—those that count for a grade. Compared with state and district tests, these classroom assessments can more readily be adapted to formative use because their results are more immediately available and their learning targets have been more recently taught. When teachers know what specific learning target each question or task on their test measures, they can use the results to select and reteach portions of the curriculum that students haven't yet mastered. Carefully designed common assessments can be used this way as well.

Students, too, can use summative test results to make decisions about further study. If the assessment items are explicitly matched to the intended learning targets, teachers can guide students in examining their right and wrong responses to answer questions such as these:

- What are my strengths relative to the standards?
- What have I seen myself improve or get better at?

- Where didn't I perform as desired, and how might I make those answers better?
- What do these results mean for the next steps in my learning, and how should I prepare for that improvement?

For students to make maximum use of these questions to guide further study, however, teachers must plan and allow time for students to learn the knowledge and skills they missed on the summative assessment and then to retake it. Lack of time for such learning is a big hindrance to formative use of summative classroom assessments.

Necessary conditions

The examples cited above began with summative purposes in mind. And the achievement gains credited to formative assessment practices will not materialize unless certain conditions are met—and at least some of these conditions are often *not* met by assessments whose primary purpose is summative. The conditions are as follows:

1. The assessment instrument or event is designed so that it aligns directly with the content standards to be learned.
2. All of the instrument or event's items or tasks match what has been or will be taught.
3. The instrument or event provides information of sufficient detail to pinpoint specific problems, such as misunderstandings, so that teachers and students can make good decisions about what actions to take.
4. The results are available in time for educators to take action with the students who generated them.
5. Teachers and students do indeed take action based on the results.

If one or more of these conditions is not fulfilled, it is at best an incomplete attempt at formative assessment, with diminishing returns the farther one strays from the conditions. Assessment does not accomplish a formative purpose when "the information is simply recorded, passed on to a third party who lacks either the knowledge or the power to change the outcome, or is too deeply coded (for example, as a summary grade given by the teacher) to lead to appropriate action."[7]

Figure 1: Formative or Summative?				
Type of assessment	**What is the purpose?**	**Who will use the information?**	**How will it be used?**	**Is the use formative or summative?**
State test	Measure level of achievement on state content standards	State	Determine AYP	Summative
		District, teacher teams	Determine program effectiveness	Summative
	Identify percentage of students meeting performance standards on state content standards	State	Comparison of school/ districts	Summative
		District, teacher teams	Develop programs/ interventions for groups or individuals	Formative
District benchmark, interim, or common assessment	Measure level of achievement toward state content standards	District, teacher teams	Determine program effectiveness	Summative
		District, teacher teams	Identify program needs	Formative
	Identify students needing additional help	District, teacher teams, teachers	Plan interventions for groups or individuals	Formative
Classroom assessment	Measure level of achievement on learning targets taught	Teachers	Determine report card grade	Summative
	Diagnose student strengths and areas needing reteaching	Teacher teams, teachers	Revise teaching plans for next year/ semester	Formative
			Plan further instruction/ differentiate instruction for these students	Formative: Assessment *for* Learning
		Teachers, students	Provide feedback to students	Formative: Assessment *for* Learning
	Understand strengths and areas needing work	Students	Self-assess, set goals for further study/ work	Formative: Assessment *for* Learning

Program = curriculum, texts/resources, and pedagogy

Source: Adapted with permission from J. Chappuis, *Seven Strategies of Assessment for Learning* (Portland, OR: ETS Assessment Training Institute, 2009), p. 8.

The table in Figure 1 lists the types of assessments typically present in school systems in the United States, describes their intended uses, and identifies the uses as formative or summative.

What Gives Formative Assessment Its Power?

The studies Black and Wiliam[8] examined represent a diverse array of interventions, all of which featured some formative use of assessment data or processes. Practices yielding the largest achievement gains displayed the following characteristics:

- use of classroom discussions, classroom tasks, and homework to determine the current state of student learning/understanding, with action taken to improve learning/correct misunderstandings;
- provision of descriptive feedback, with guidance on how to improve, during the learning; and
- development of student self- and peer-assessment skills.

Drawing from their analysis of these studies, Black and Wiliam[9] make the following recommendations about key components of formative assessment:

- "Opportunities for students to express their understandings should be designed into any piece of teaching, for this will initiate the interaction through which formative assessment aids learning."

- "The dialogue between pupils and teachers should be thoughtful, reflective, focused to evoke and explore understanding, and conducted so that all pupils have an opportunity to think and to express their ideas."

- "Feedback to any pupil should be about the particular qualities of his or her work, with advice on what he or she can do to improve, and should avoid comparison with other pupils."

- "Feedback on tests, seatwork, and homework should give each pupil guidance on how to improve, and each pupil must be given help and an opportunity to work on the improvement."

- "If formative assessment is to be productive, pupils should be trained in self-assessment so that they can understand the main purposes of their learning and thereby grasp what they need to do to achieve."

Notice where these recommended practices fall on the table in Figure 1—they are the cells labeled "assessment *for* learning." Formative assessment *is* a powerful tool in the hands of both teachers and students, and the closer it is to everyday instruction, the stronger it is. Classroom assessment, sensitive to what teachers and students are doing daily, is most capable of providing the basis for understandable and accurate feedback about the learning, while there is still time to act on it. And it has the greatest capacity to develop students' ability to monitor and adjust their own learning.

Formative assessment in teachers' hands

Many formative assessment strategies address the teacher's information needs, helping to answer questions critical to good instruction:

- Who is and who is not understanding the lesson?
- What are this student's strengths and needs?
- What misconceptions do I need to address?
- What feedback should I give students?
- What adjustments should I make to instruction?
- How should I group students?
- What differentiation do I need to prepare?

There is no doubt that, acting on good information during the course of instruction, teachers can increase what and how well students learn. Indeed, some of the significant achievement gains attributable to formative assessment are due to enhanced questioning and dialogue techniques.

Many strong programs and practices help teachers obtain, interpret, and act on student achievement information. However, if the discussion of formative assessment considers only teachers' use of assessment information, one very important player is sitting on the sidelines, and it's not the principal or the superintendent—we have benched the student.

Formative assessment in students' hands: Assessment for learning

Black and Wiliam's[10] research review showcases the student as decisionmaker. Many other prominent education experts have also described the benefits of student involvement in the assessment process. In an often-cited article describing how formative assessment improves achievement, D. Royce Sadler[11] concludes that it hinges on developing students' capacity to monitor the quality of their own work during production:

> The indispensable conditions for improvement are that the *student* comes to hold a concept of quality roughly similar to that held by the teacher, is able to monitor continuously the quality of what is being produced *during the act of production itself,* and has a repertoire of alternative moves or strategies from which to draw at any given point.

Writing about formative assessment in the science classroom, Atkin, Black, and Coffey[12] translate the conditions Sadler describes into three questions:

1. Where are you trying to go? (identify and communicate the learning and performance goals);
2. Where are you now? (assess, or help the student to self-assess, current levels of understanding); and
3. How can you get there? (help the student with strategies and skills to reach the goal).

Sadler's conditions as represented in these three questions frame what is called "assessment *for* learning"—a collection of formative assessment practices designed to meet students' information needs, maximizing both motivation and achievement by involving students from the start in their own learning.[13] Those practices, summarized below, illustrate how the formative assessment research recommendations play out in the hands of a knowledgeable classroom teacher.

Practices designed to answer the question, "Where am I going?":

- **Provide students with a clear and understandable vision of the learning target.** Motivation and achievement both increase when instruction is guided by clearly defined targets. Activities that help

students answer the question, "What's the learning?" set the stage for all further formative assessment actions.

- **Use examples and models of strong and weak work.** Carefully chosen examples of the range of quality can create and refine students' understanding of the learning goal by helping students answer the questions, "What defines quality work?" and "What are some problems to avoid?"

Practices designed to answer the question, "Where am I now?":

- **Offer regular descriptive feedback.** Effective feedback shows students where they are on their path to attaining the intended learning. It answers for students the questions, "What are my strengths?"; "What do I need to work on?"; and "Where did I go wrong and what can I do about it?"

- **Teach students to self-assess and set goals.** The information provided in effective feedback models the kind of evaluative thinking we want students to be able to do themselves. Teaching students to identify their strengths and weaknesses and to set goals for further learning prepares them to generate their own answers to the questions, "What am I good at?"; "What do I need to work on?"; and "What should I do next?"

Practices designed to answer the question, "How can I close the gap?":

- **Design lessons to focus on one learning target or aspect of quality at a time.** When assessment information identifies a need, teachers can adjust instruction to target that need. They scaffold learning by narrowing the focus of a lesson to help students master a specific learning goal or to address specific misconceptions or problems.

- **Teach students focused revision.** When a concept, skill, or competence proves difficult for students, teachers can structure practice in smaller segments, and give them feedback on just the aspects they are practicing. This allows students to revise their initial work with a focus on a manageable number of learning targets or aspects of quality.

- **Engage students in self-reflection, and let them keep track of and share their learning.** Long-term retention and motivation increase when students look back on their journey, reflecting on their learning and sharing their achievement with others.

The practices described above constitute actions that strengthen students' sense of self-efficacy (belief that effort will lead to improvement), their motivation to try, and, ultimately, their achievement.

Formative assessment and assessment for *learning*

Effective formative assessment, then, is comprised of both teacher and student actions. When teachers assess student learning for formative purposes, the intent is not to generate a final grade for the paper or the grade book. Rather, the assessment event serves as practice for students, developing and refining their mastery of the intended learning goals. Formative assessment that includes assessment *for* learning enhances achievement in two ways:

- teachers can adapt instruction on the basis of evidence, making changes that will benefit learning immediately; and
- students can use evidence of their current progress to actively manage and adjust their own learning.[14]

This is a use of assessment information that differs from the traditional practice of associating *assessment* with *test*, and *test* with *grade*. It is a broader vision of what assessment is and what it is capable of accomplishing. Taken together, these are the practices that research studies indicate will cause significant achievement gains, with the largest gains coming for the lowest achievers.

What Does Formative Assessment Measure?

Visualize a ladder with a state standard resting on top; those students who get to the top have mastered that standard. Formative assessment tells teacher and student where the student is on the ladder leading to the top at any point in time. With this information, they can team up to determine what comes next in that student's learning. The rungs on the ladder

represent the daily targets of instruction, the foundations of knowledge, reasoning, performance skills, and product development capabilities that students must master as they ascend over time to academic competence as reflected in the standard at the top. Where students are in their learning changes day to day, and their current status is best captured through ongoing and accurate classroom assessment.

Implementing Effective Formative Assessment

It is important to keep in mind that formative assessment is a human process that involves teachers and students generating information and acting on it to improve learning. Its effective use hinges on the assessment literacy of educators: classroom teachers must be able to select, modify, or create accurate assessments as needed during the course of instruction, adhering to standards of quality.[15] The authors of this chapter, in collaboration with others, developed a set of five standards, called "Indicators of Sound Classroom Assessment Practice," that describe what teachers need to know and be able to do with respect to classroom assessment. The first three standards ensure accuracy of the assessment information:

- **Clear Purpose:** Teachers must know how to use assessment processes and results to meet the information needs of all users.
- **Clear Targets:** Teachers must be able to establish clear learning targets for students.
- **Sound Design:** Teachers must be able to translate learning targets into assessments that yield accurate results.

The last two standards ensure effective use of the assessment information:

- **Effective Communication:** Teachers must manage assessment results well and communicate them effectively to all stakeholders.
- **Student Involvement:** Teachers must actively engage students in generating, interpreting, and acting on their own assessment information.

Figure 2: Indicators of Sound Classroom Assessment Practice—What Teachers Need to Know and Be Able to Do	
1) Clear Purpose Use assessment processes and results to meet information needs of all intended users.	• Understand who the users of classroom information are and how to meet their information needs. • Understand the relationship between assessment and student motivation and craft assessment experiences to maximize student engagement. • Use classroom assessment processes and results formatively, to plan next steps in learning. • Use classroom assessment results summatively to communicate students' levels of achievement at a point in time. • Know how to balance summative and formative uses of assessment information.
2) Clear Targets Establish clear and valued student learning targets.	• Establish clear learning targets for students; know how to turn broad content standards into classroom-level targets. • Understand the types of learning targets students are to achieve. • Create a comprehensive plan over time for assessing learning targets formatively and summatively.
3) Sound Design Translate learning targets into assessments that yield accurate results.	• Know what the four assessment methods are and when to use each. • Select, modify, or design assessments that serve different purposes. • Select, modify, or design assessments that reflect intended learning targets. • Write assessment questions of all types well. • Sample learning appropriately. • Avoid sources of mismeasurement that bias results.
4) Effective Communication Manage assessment results well and communicate them effectively.	• Record, summarize, and translate assessment information into a grade accurately. • Select the best reporting option for the context. • Interpret and use standardized test results correctly. • Communicate assessment information to students effectively. • Communicate assessment information to parents, colleagues, and other stakeholders effectively.
5) Student Involvement Engage students in generating, interpreting, and acting on their own assessment information.	• Make learning targets clear to students. • Involve students in assessing and setting goals for their own next steps. • Involve students in tracking, reflecting on, and communicating about their learning.

Source: Adapted with permission from R. J. Stiggins, J. Arter, J. Chappuis, and S. Chappuis, *Classroom Assessment* for *Student Learning: Doing It Right—Using It Well* (Portland, OR: ETS Assessment Training Institute, 2004), p. 27.

Knowledge of both assessment accuracy and effective use are necessary conditions to implementing formative assessment; if the assessment itself yields inaccurate information, no judgments or actions based on its results are likely to improve learning. Yet, up to this time, as a nation we have not invested in developing classroom assessment competencies. Few teachers are prepared to meet these standards of classroom assessment practice, because they have not been given the opportunity to do so. As a result, student progress is in jeopardy of daily mismeasurement, thus compromising instructional decisions students, teachers, and parents make on a regular basis—students' understanding of their learning capabilities, teachers' diagnoses of learning needs, and communication to parents and others about student progress.

The educational leader's role in the use of formative assessment

Leadership at the school and district level is crucial to the implementation of sound assessment practices. Building leaders' essential role is comprised of four key actions:

- monitoring assessment quality, including assessment *for* learning practices;
- facilitating department-wide and building-wide collaboration;
- contributing to development of supportive school policies; and
- ensuring professional development to strengthen classroom assessment expertise.

All those who supervise teachers can use the Indicators of Sound Classroom Assessment Practice formally through classroom observations and follow-up conversations and informally through discussions to monitor teachers' knowledge and use of high-quality assessment practices. If they are to provide meaningful feedback to teachers on these subjects, principals and other supervisors must be able to differentiate between sound and unsound practices, and must be committed to deepening their own learning if they are not masters of the indicators themselves. Principals can hold regularly scheduled faculty discussions of formative assessment actions to center collaborative action on using assessment in ways that impact learning beyond final report card grades and test data analysis.

In addition, school leaders should be able to develop and implement school policies that contribute to sound assessment practice and also help achieve a balanced assessment system in the school and district. These include communication policies and practices regarding grading, reporting student progress, and communicating about the variety of school assessments and their relationship to improving curriculum and instruction.

Leaders at the district level are responsible for creating balanced assessment systems, including provisions for the effective use of formative assessment. By developing comprehensive assessment plans that address the information needs of all users of assessment results, district leaders can lay the groundwork for quality at all levels. Along with aligning their local assessment system with the state assessment system, they can ensure that sound classroom assessment practices are considered integral to teaching well in their districts. They can also make the professional development needed to assess well at all levels of schooling a priority district-wide.

Leaders should also be able to plan for, pace, and facilitate or monitor the professional development teachers need to become knowledgeable about accurate and effective assessment practices. That responsibility leads directly into the challenges ahead.

Challenges to implementation

The most significant challenge is that of ensuring that educators are prepared to assess accurately and to use assessment to support learning. Unfortunately, few states explicitly include competence in assessment as a requirement to be licensed to teach. Teacher licensing examinations do not yet verify competence in classroom assessment. Building- and district-level leaders lack the assessment competencies needed to build balanced, instructionally helpful local assessment systems. Assessment literacy training remains minimal in the majority of pre-service teacher and educational administration programs.

Second, the universal lack of pre-service training is exacerbated by similar weaknesses in support for working educators: a focus on assessment competencies is not prominent among in-service professional development offerings. Increasingly, teachers do receive training on the interpretation and

use of data, particularly related to their district- and state-level assessments. But the attention needed at the classroom level is overshadowed by the need to succeed on accountability tests. However, some states (Vermont, Delaware, West Virginia, Ohio, Illinois, South Carolina, and Kentucky) and districts (Clark County SD, Naperville SD, Olentangy SD, and Poway SD) have initiated professional development in-service programs aimed at helping teachers improve classroom assessment.

As a result of inadequate pre-service and in-service training for teachers and leaders, the United States currently relies on a national faculty still largely untrained in the principles of sound classroom assessment. The result can be and often is unsound school- and district-level assessment policies, inappropriate evaluations of teachers' assessment practices, a lack of resources for teachers to learn to assess productively, and poor advice to noneducation policymakers, such as school board members and legislators. What is perhaps most unfortunate about this is that it need not be this way: there are known strategies for raising the level of assessment literacy for K–12 teachers and prospective teachers—effective professional development resources and programs now exist that can help close that gap.

The third challenge is the tendency to bypass professional development in the area of classroom assessment altogether—to teacher-proof the assessment-related parts of instruction—by importing "formative" assessment instruments from outside the classroom. When the classroom teacher's assessment responsibilities are circumvented, albeit unknowingly, it severely limits the student achievement gains attainable through developing teachers' assessment literacy and knowledge of how to make assessment and instruction work together.

The fourth challenge lies in current national assessment policies characterized by the limiting belief that assessment's sole purpose is to measure student performance, coupled with an assumption that improving the quality of annual summative tests will serve to improve schools. Sixty years and billions of dollars later, district, state, and national tests have not produced the magnitude of school improvement expected, especially for low-achieving students.

This is because of the flaw in the theory. Annual summative tests may serve important and valued accountability purposes, but they are limited in their ability to inform instructional decisions. The power to impact the truly crucial decisions that affect teaching and learning resides with the other 99.9 percent of the assessments that occur in a student's academic career—those conducted by teachers in classrooms on a daily basis.

Improving the assessment practices in our classrooms and schools is hard work, particularly while large-scale accountability testing dominates assessment discussions. Until local, state, and federal policymakers better understand sound assessment practices and the limitations of large-scale assessment, and allocate resources to balance large-scale accountability testing with the effective use of high-quality interim and classroom assessments, the nation's students will not achieve at high levels.

Recommended Actions for Policymakers

The challenges faced in bringing better assessment practices to the classroom are ultimately an issue of balance: the nation's education stakeholders need help in understanding and creating comprehensive, *balanced* assessment systems, as described in the opening of this chapter.

Federal policymakers can help overcome the challenges described, thus assisting schools to move toward the effective use of formative assessment. This includes bringing national visibility to the importance of formative assessment practices in improving teaching and learning, and acknowledging that the use of such assessments is an expectation teachers and principals have not been adequately prepared to meet. Specific federal policy action could include

- supporting the implementation of balanced assessment systems and the improvement of both quality and effective use of classroom assessment;
- encouraging state efforts to improve pre-service policies that improve future teachers' and school leaders' assessment literacy; this issue is directly related to teacher quality, and if acted upon would make a contribution that so far has been largely absent;

- requiring the use of federal professional development dollars to include activities designed to improve educators' formative assessment practices;
- calling for the allocation of other resources to prepare teachers and school leaders to use assessment in support of learning for all students; and
- supporting educational research that continues to inform what formative assessment practices contribute most to raising student achievement.

Education policy and practice at any level that leads to a steady diet of ready-made external tests will not bring about the gains in student achievement promised by formative assessment practices. Such external tests cannot substitute for the daily formative assessment practices that only assessment-literate educators are able to conduct. The greatest value in formative assessment lies in teachers and students making use of results to improve real-time teaching and learning at every turn.

(Portions of this paper are adapted from S. Chappuis and J. Chappuis, "The Best Value in Formative Assessment," *Educational Leadership* 65, no. 4 [2007]: 14–19; and J. Chappuis, *Seven Strategies of Assessment for Learning* [Portland, OR: ETS Assessment Training Institute, 2009].)

The views expressed in this chapter are those of the authors and do not necessarily represent those of the Alliance for Excellent Education.

About the Authors

Jan Chappuis has been an elementary and secondary classroom teacher, a curriculum and classroom assessment specialist, and a professional development specialist. She has worked for the past eight years with the ETS Assessment Training Institute in Portland, Oregon. She is the author of *Seven Strategies of Assessment for Learning* (2009) and *Learning Team Facilitator Handbook* (2007), and the coauthor of *Creating and Recognizing Quality Rubrics* (2006), *Classroom Assessment for Student Learning: Doing It Right—Using It Well* (2004), *Assessment FOR Learning: An Action Guide for School Leaders* (2005), and *Understanding School Assessment—A Parent and Community Guide to Helping Students Learn* (2002).

Stephen Chappuis's career as a teacher, counselor, and building and district administrator in public school districts spans twenty-eight years. His experience includes being a junior high principal, a senior high principal, and assistant superintendent for curriculum and instruction. In the latter role he implemented a standards-based instructional program that included comprehensive assessment plans and policies with professional development for teachers in classroom assessment. As executive director of the ETS Assessment Training Institute, he works with school leaders to develop assessment literacy and balanced local assessment systems. He has written for *Education Week, Educational Leadership,* and *School Administrator*, and is the coauthor of *Classroom Assessment for Student Learning: Doing It Right—Using It Well* (2004), *Assessment FOR Learning: An Action Guide for School Leaders* (2005), and *Understanding School Assessment—A Parent and Community Guide to Helping Students Learn* (2002).

Richard Stiggins founded the Assessment Training Institute (ATI) in 1992 to provide professional development in classroom assessment for educators. Throughout his career, he has brought his expertise in educational measurement to bear on supporting teachers and administrators in their preparation to meet the task demands of accurate classroom assessment used in service of student learning. He is the author of numerous articles and books and the coauthor of *Classroom Assessment for Student Learning: Doing It Right—Using It Well* (2004) and *Assessment FOR Learning: An Action Guide for School Leaders* (2005).

---

[1] P. Black and D. Wiliam, "Inside the Black Box: Raising Standards Through Classroom Assessment," *Phi Delta Kappan* 80 (October 1998): 140.

[2] W. Harlen and M. James, "Assessment and Learning: Differences and Relationships Between Formative and Summative Assessment," *Educational Assessment: Principles, Policy and Practice* 4, no. 3 (1997): 365–79, p. 369.

[3] D. R. Sadler, "Formative Assessment: Revisiting the Territory," *Assessment in Education* 5, no. 1 (1998): 77–84, p. 77.

[4] L. A. Shepard, "Formative Assessment: Caveat Emptor," in *The Future of Assessment: Shaping Teaching and Learning*, ed. C. Dwyer, 279–303 (New York: Lawrence Erlbaum Associates, 2008), p. 281.

[5] "Mission and History of the Formative Assessment for Students and Teachers SCASS," www.CCSSO.org (accesssed March 9, 2009).

[6] S. Chappuis, "Is Formative Assessment Losing Its Meaning?" *Education Week* 24, no. 44 (2005).

[7] D. R. Sadler, "Formative Assessment and the Design of Instructional Systems," *Instructional Science* 18, no. 2 (1989): 119–44.

[8] P. Black and D. Wiliam, "Assessment and Classroom Learning," *Educational Assessment: Principles, Policy and Practice* 5, no. 1 (1998): 7–74; Black and Wiliam, "Inside the Black Box."

[9] Black and Wiliam, "Inside the Black Box."

[10] Black and Wiliam, "Assessment and Classroom Learning."

[11] Sadler, "Formative Assessment and the Design of Instructional Systems," p. 121, italics in original.

[12] J. M. Atkin, P. Black, and J. Coffey, *Classroom Assessment and the National Science Standards* (Washington, DC: National Academy Press, 2001), questions on p. 14.

[13] R. J. Stiggins, J. Arter, J. Chappuis, and S. Chappuis, *Classroom Assessment for Student Learning: Doing It Right—Using It Well* (Portland, OR: ETS Assessment Training Institute, 2004).

[14] Ibid.

[15] Ibid.

CHAPTER 4

The Role of Interim Assessments in a Comprehensive Assessment System

Judy Wurtzel*
Aspen Institute

Scott Marion, Marianne Perie, and Brian Gong
National Center for the Improvement of Educational Assessment

The standards-based reform movement has resulted in the widespread use of summative assessments designed to measure students' performance at specific points in time. Recognizing that these end-of-year tests are not intended to and do not provide useful information to regularly inform and track student learning during the year, educators are looking for additional assessments to fill that need. Many vendors are now selling what they call "benchmark," "diagnostic," "formative," and/or "predictive" assessments with promises of improving student performance. These assessment systems often lay claim to the research documenting the powerful effect of formative assessment on student learning. However, the research in this area, including the seminal Black and Wiliam (1998) meta-analysis, evaluated formative assessments of a very different character than essentially all current commercially available interim assessment programs.

This chapter differentiates between true classroom formative assessment and the interim assessments currently in the marketplace and provides a framework for considering the appropriate role of interim assessments. It looks at six issues: (1) distinguishing among assessment types; (2) key questions for educational leaders; (3) determining the purpose for the interim assessment; (4) characteristics of an effective interim assessment system; (5) current commercially available interim assessment systems; and (6) implications for district, state, and federal decisionmakers. Our goals are to help district leaders make better decisions about the purchase and use of interim assessment systems and to help state and federal policymakers consider what role they might play in supporting effective interim assessment practices.

Issue 1: Distinguishing Among Assessment Types

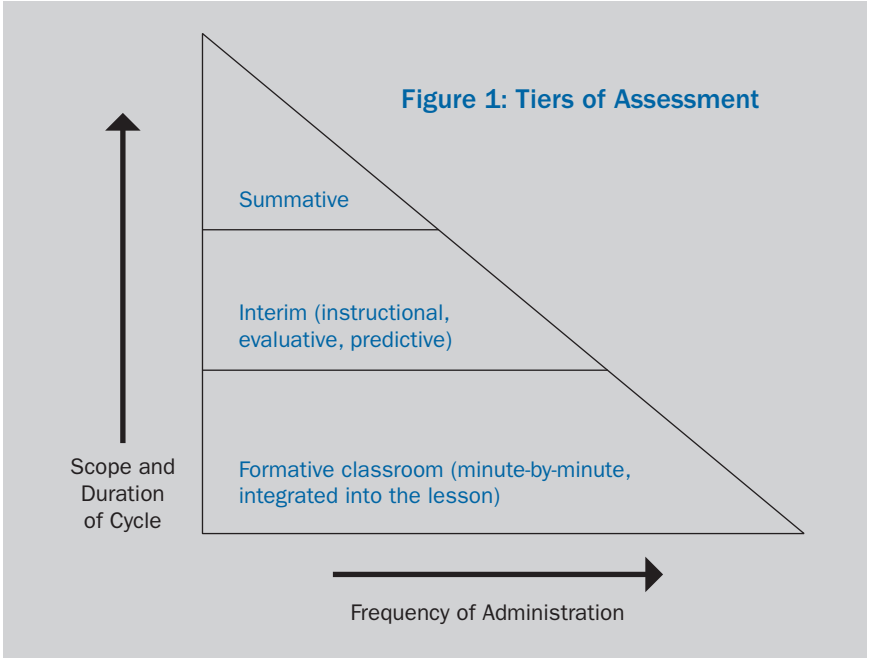
Our schema recognizes three assessment types—summative, interim, and formative—and distinguishes among them based on their intended purposes, audience, and use of the information, rather than simply on when the assessment is given.

Summative assessments are generally given once, at the end of some unit of time (such as the semester or school year), to evaluate students' performance against a defined set of learning targets (e.g., content standards). Because summative assessments are given at the end of a period of instruction, they are not particularly useful for educators to use in adjusting instruction or interventions to address individual student needs. These assessments typically are administered statewide (but can be national or district) and are usually used as part of an accountability program or to otherwise inform policy. State tests mandated under the No Child Left Behind Act (NCLB) are of this type.

Formative assessment is a process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve students' achievement of intended instructional outcomes.[1] It is done by the teacher in the classroom for the explicit purpose of diagnosing where students are in their learning, where gaps in knowledge and understanding exist, and how to help teachers and students improve student learning. Formative assessment is embedded within the learning activity

and linked directly to the current unit of instruction. The assessments/activities generally are small-scale (a few seconds or a few minutes, and certainly less than a class period) and short-cycle (they are often called "minute-by-minute" assessments, or formative instruction). Tasks or prompts may vary among students depending on the teacher's judgment about the need for specific information about a student. Providing corrective feedback and modifying instruction are essential aspects of a classroom formative assessment.

Interim assessment is the suggested term for the assessment that falls between formative and summative assessments, including the medium-scale, medium-cycle assessments currently in wide use. Interim assessments evaluate students' knowledge and skills relative to a specific set of academic goals, typically within a limited time frame, and are designed to inform decisions both in the classroom and beyond the classroom (such as at the school or district level). They may be given at the classroom level to provide information for the teacher, but unlike formative assessments, the results of interim assessments can be meaningfully aggregated and reported at a broader level. As such, the timing of the assessment administration is likely



The triangle in Figure 1 illustrates the distinctions between the three types of assessment.

to be controlled by the school or district rather than by the teacher. Many of the assessments labeled "benchmark," "formative," "diagnostic," or "predictive" fall within our definition of interim assessments.

It is important to begin with these definitions, because even assessment experts have been hobbled by the lack of clear definitions of assessment types. There is concern among some experts that there has been some co-opting of the formative assessment label and research by those purveying nothing of the sort. This imprecision has led to a blurring of the perceived differences between "formative assessment" and "interim assessment." Districts putting interim assessments in place may be getting important and actionable data, but they are rarely getting the power of true formative assessment practices.

Issue 2: Key Questions for Educational Leaders

When deciding whether to include interim assessments in a district's overall assessment system, it is important to be clear about the intended purpose and use of interim assessments and how the particular assessments will work in the teaching-learning cycle. As a start, it will be helpful to address the following questions:

1. What do we want to learn from this assessment?
2. Who will use the information gathered from this assessment?
3. What action steps will be taken as a result of this assessment?
4. What professional development or support structures should be in place to ensure that the action steps are taken and are successful?
5. How will student learning improve as a result of using this interim assessment system, and will it improve more than if the assessment system were not used?

The answers to these questions will reveal a theory of action about how assessments will lead to improved student learning and drive many of the design decisions. Importantly, these questions and the associated answers serve as the beginning of a validity argument to support (or refute) the particular assessment system. While this chapter focuses primarily on the first question, all five, especially the last, are essential to consider.

In addition, it is important to reflect on how an intended assessment works in the context of other assessments in use. Large school districts often have a plethora of assessments put in place for a variety of reasons over the course of many years. It is suggested that districts conduct an assessment audit that examines what assessments exist, their intended purposes, the results produced, and the utility of the data. Based on the audit results, districts may be able to eliminate less useful assessments, reduce assessment burden, avoid distracting educators and the public with non-useful results, and reclaim instructional time.

Issue 3: Determining the Purpose for the Interim Assessment

A critical task for policymakers is to answer the first question posed above—"What do I want to learn from this assessment, and why do I want to learn it?"—and then find or develop a set of assessments that best fits that purpose. Despite claims to the contrary, single assessments rarely serve multiple purposes well. They tend to work best when the limited number of purposes have been prioritized explicitly. Interim assessments can be thought of in terms of three general classes of purposes—instructional, evaluative, and predictive—with many specific purposes within each.

A. Instructional purposes: Interim assessments designed to serve instructional purposes should provide results that enable educators to adapt instruction and curriculum to better meet student needs. Within this general category of instructional purposes, policymakers and assessment leaders must go further and prioritize specific instructional purposes to better guide assessment design and/or selection. For example, interim assessments might be used to enrich the curriculum, determine students' strength and weakness in a particular domain, or provide feedback to students for motivational and metacognitive reasons.

When the purpose is to enrich the curriculum, assessments should be designed to have students explore concepts in greater depth or provide tasks that stretch students and teachers to do things at deeper cognitive levels than they might otherwise. The assessment itself contributes to enriching the instruction. When the purpose is to illuminate the strengths and weaknesses of individuals or groups of students, an

assessment system often contains a bank of items aligned with the state curriculum that teachers can use to create a test to evaluate student learning on the concepts taught to date. Results are reported immediately, and data are disaggregated by content standard (or some other unit of learning), allowing teachers to identify strengths and weaknesses in the students' learning. Ideally, to provide actionable information the assessment is fully aligned with the specific classroom or at least school curriculum and provides more in-depth analyses of student misconceptions along with instructional tools and strategies for improving instruction. When the specific instructional purposes are motivating and providing feedback to students, tasks should engage students and encourage them to wrestle with challenging subject matter knowledge. Quick feedback afforded by computer-based testing programs and rich tasks that make student thinking and productions explicit, such as exhibitions and projects, can both achieve these aims. Unfortunately, many purveyors of computer-based interim assessment systems only provide selected response formats (e.g., multiple choice), thereby delivering on only the fast turnaround part of the promise.

B. Evaluative purposes: The primary goal of interim assessments designed for evaluative purposes is to provide information to help the teacher, school administrator, curriculum supervisor, or district policymaker learn about curricular or instructional choices and take specific action to improve the program. These actions are intended to influence subsequent teaching and thereby, presumably, improving the learning. This can be thought of as a program evaluation designed to change curriculum and instruction over the years. An example would be assessments given at various points throughout the year to provide more details about student performance on instructionally relevant subdomains (e.g., adding simple fractions)—not with the intention of intervening but for evaluating the effectiveness of a program or strategy. Another set of important evaluative purposes is to enforce some minimal quality through standardization of curriculum and pacing guides, centralizing coordination for highly mobile urban student populations and high teacher turnover, or as a lever to overcome differences in learning expectations and grading standards.

C. Predictive purposes: Predictive assessments are designed to determine each student's and groups of students' likelihood of meeting some criterion score on the end-of-year tests or other future outcome. While such predictions are of great interest, they obviously must be coupled with further analysis and action. For example, districts might use predictive assessments as a screener to identify students who are not on track to score proficient on the end-of-year test so that they can be given further probes to determine areas of weakness and be provided with remedial instruction, extra support, and/or tutoring. This scenario highlights the value of having formative, interim, and summative assessment types aligned in a comprehensive system.

D. Multiple purposes: Given constrained resources, it is no wonder that educational leaders are tempted to use a single assessment for as many purposes as possible. Unfortunately, one of the truisms in educational measurement is that when an assessment is designed to fulfill too many purposes it rarely fulfills any purpose well.

This does not mean that certain interim assessment systems cannot fulfill more than one purpose. If the system is intended to provide rich information about individual students' strengths and weaknesses tied to a particular set of curricular goals, then these results can likely be aggregated to the subgroup, school, and/or district level to provide evaluative and predictive information. On the other hand, if the primary goal is to gather predictive or early-warning information, it is unlikely that the assessment will contain rich enough information for full instructional or even evaluative purposes. Therefore, if users want to fulfill multiple purposes, they must design a system to fulfill the finest-grain purposes first and then aggregate the results to more general levels in the educational system. However, users still need to be sure that multiple purposes are not contradictory, such as might be the case when an assessment is used for both instructional and accountability purposes.

Issue 4: Characteristics of an Effective Interim Assessment System to Be Used for Instructional Purposes

Once educational leaders are clear about purposes for the interim assessment system, they still face many additional considerations and decisions about the system's design and implementation. While there is little research evidence about the characteristics of effective interim assessments, our work with states and districts suggests some commonsense guidance. This chapter focuses on characteristics of interim assessments for instructional purposes because most districts appear to want to use assessments for this purpose, most vendors say their assessments can meet that purpose, and there are more concerns about claims for instructional purposes then for evaluative and predictive purposes.

There is no one-size-fits-all assessment, only a best design for a desired use and the existing constraints and resources. Given that, the general characteristics of any interim assessment to be used for instructional purposes should include

- provision for qualitative insights about understandings and misconceptions and not just a numeric score;
- immediate implications for what to do besides reteaching every missed item;
- rich representation of the content standards students are expected to master;
- high-quality test items, including rich open-ended tasks, that are directly linked to the content standards and specific teaching units;
- a good fit within the curriculum (preferably a curriculum aligned to a similar conception of student learning as the formative assessment strategies) so that the test is an extension of the learning rather than a timeout from learning;
- a good fit with curriculum pacing so that students are not tested on content not yet taught;
- clear reporting that provides actionable guidance on how to use the results;
- validation of the uses of the information provided by the assessment;

- administration features (speed, availability of normative information, customization, timing flexibility, adaptability) that match the assessment purposes; and
- professional development for teachers.

While each item on this checklist could be discussed in depth, this chapter focuses on reporting results, inclusion of data in accountability systems, and item type, because of their importance to policymakers.

One strategy for defining the desired characteristics is to focus on reporting. Score reports make test results actionable. Designing the reporting system at the beginning clarifies all the information desired from the assessment. Score reports should be vetted with those who need to use the information—teachers, in most cases, but also school leaders.

Another key issue is item type. Interim assessments can include a wider range of item types than what is typically the case. In particular, extended performance tasks can serve instructional purposes more readily than other interim assessment item types. They enrich the curriculum, increase student motivation by engaging them in meaningful interactions with rich subject matter, and provide opportunities for teachers to learn about student thinking. As long as the results can be aggregated and used at a level beyond the classroom (which can be done through systematic observations, rubrics, and other scoring methods), an assessment with these types of tasks falls under our definition of interim.

Issue 5: Current Commercially Available Interim Assessment Systems

Once decisionmakers have determined purposes and key design characteristics, they can then determine whether to choose among commercially available interim assessment systems or develop their own. Test companies offer interim assessment products, often labeled "formative" or "benchmark," for a wide variety of purposes. The best current commercially available systems can

- provide an item bank reportedly linked to state content standards;

- assess students on a flexible time schedule wherever a computer and, if necessary, an Internet connection are available;
- provide immediate or very rapid results;
- highlight content standards in which more items were answered incorrectly; and
- link scores on these assessments to the scores on end-of-year assessments to predict results on end-of-year assessments.

Many of the better commercially available interim assessment products can address questions such as:

- Is this student on track to score "proficient" on the end-of-year NCLB tests?
- Is the student improving over time?
- What proportion of students are at risk of scoring below proficient on the end-of-year NCLB tests?
- On which content standards are the students performing relatively well (or poorly) (for a student, classroom, school, district, state)?
- How does this student's performance compare to the performance of other students in the class?

Unfortunately, most commercially available interim assessment systems currently do not

- address multiple purposes (i.e., instructional, evaluative, or predictive) well;
- provide rich detail about the curriculum assessed;
- help teachers understand the nature of a student's misconception(s);
- report detailed information on the student's depth of knowledge on a particular topic;
- further a student's understanding through the type of assessment task; and
- give teachers the information on how to implement an instructional remedy.

Furthermore, these systems typically do not answer the following questions:

- Why did a student answer an item incorrectly?
- What are some possible strategies for improving performance in this content area?
- What did the student learn from this assessment?
- What type of thinking process is this student using to complete this task?

Given this analysis, there is continued concern about the weakness of available interim assessment systems for instructional purposes. Nonetheless, interim assessments can play a productive role in this and other areas. In particular, in terms of predictive and evaluative purposes, interim assessments can help districts determine whether all schools have similarly high standards, whether highly mobile students have exposure to a common curriculum, and which students are "off track" so they can intervene.

Issue 6: Implications for District, State, and Federal Decisionmakers

Districts clearly set the policy and practice context for effective use of interim assessments. While interim assessments are typically not used for policy purposes at the state level, and not at all at the federal level, there is also a considerable role for states—and a limited role for the federal government—in support of effective interim assessment use. Productive roles for leaders at the district, state, and federal levels are suggested below.

Leadership: Establishing and maintaining the vision

Leaders at all levels must be clear and coherent as they articulate their vision for learning, instruction, assessment, and school quality and then work to ensure consistency among the various initiatives and policies. This broad vision should support the meaningfulness of formative and interim assessments. For example, leaders who dwell on the large-scale summative results as the only important measure of school quality can undermine the use of other assessment approaches. Leaders should conduct assessment audits and, when designing assessment systems, be thoughtful about their intended purposes, the results produced, and the usefulness of the data.

These steps will help eliminate less useful assessments, reduce the assessment burden, and reclaim instructional time.

Standards, assessment, and curriculum design decisions

In a standards-based environment, interim assessment practices are driven by decisions about standards, curriculum, and state assessments. Leaders can insist on high-quality college- and career-readiness standards that are developed according to the most up-to-date learning theories, which support the assessment of specific learning targets in ways that best facilitate learning and instruction. They can ensure that state assessments focus on a limited number of meaningful outcomes and use rich item formats; this signals to the field the types of learning outcomes and tasks that are valued, and thus supports better interim and formative assessment practices. They can support the development and use of challenging curriculum that includes embedded formative and interim assessments to equip teachers with the tools to translate the ambitious learning goals articulated in good standards into meaningful, rich learning experiences for students. While most of these decisions are made by state leaders, there is an important federal role in supporting state efforts through funding support, accountability policies (see below), and the R&D investments needed to develop these next-generation standards, assessments, and curriculum.

Accountability policies

While there is no hard evidence on the best approach, our sense is that the results of interim assessments should be made public within a district (among teachers, administrators, and parents) but should not be used for school or district accountability purposes. This is particularly true if assessments are to be used for instructional purposes and the goal is for teachers to use assessment results as the basis for conversations among themselves and with their students about the nature of students' work and the changes in their own practice that are needed to improve this work. For such conversations and analyses to take place, teachers must believe in—and not fear—the assessment results. However, state policies, such as those found in Wyoming and Rhode Island, where results of local interim assessments (including portfolios of student work) are used for graduation certification can direct attention and efforts toward improving local

assessment practices. (It should be noted that in both of these states there has been a significant amount of state support to help local districts build and use sound assessment systems.)

Funding and other resources

Money matters! It is cheaper to score multiple-choice items than constructed-response items or performance tasks, and it often costs less to buy a computer-based testing system than to invest in professional development for all teachers. Even within the reality of constrained budgets, saving a few dollars on an assessment system might actually "cost" more in terms of opportunities for learning. State and federal governments can help underwrite the costs of effective assessment systems and associated professional development. Beyond new money, leaders can reallocate existing resources to support formative and other local assessment initiatives in ways that make clear their importance.

Professional development policies

A consistent research finding is that the effectiveness of any test used for instructional purposes is dependent on how the teacher uses the information to give feedback to the students. Teachers need to learn how to administer the assessment, learn from the results, and adjust instruction accordingly. While districts must take the lead in this area, states can play a valuable role in funding, designing, and/or providing such professional development. The federal government can support this as well. Moreover, the state should work with teacher credentialing and teacher education institutions to make this a more salient part of teacher pre-service training.

Quality control

The state can be invaluable in the area of quality control, and the federal government can aid those efforts through funding and research. First, it can vet potential vendors and provide information on the characteristics of the assessments available, quality of the items, and the degree of alignment with state curriculum. The state department of education may simply provide the information to districts or choose to allow state funds to be spent only on interim assessment systems that meet specific qualifications. Secondly,

the state can provide access to research various interim assessments and best practices in their use. Thirdly, the state department of education can network districts using similar assessment systems.

Evaluation

Given the lack of research supporting the use of interim assessment and the many questions about the power and validity of different types of interim assessments, it is suggested that decisionmakers at all levels deliberately and continuously evaluate the effectiveness of interim assessment strategies within and across districts and adjust accordingly. Evaluations should include teacher surveys or focus groups to determine how the data were used and if there is evidence that the information gained from interim assessments improved student learning. Leaders should critically consider whether investments in improving daily classroom assessment practices are a more effective strategy for improving student learning than purchasing interim assessments.

Conclusion

Our hope is that policymakers will take at least six points from their reading of this chapter.

First, interim assessments, as defined in this chapter, are distinct from formative assessments. While a definition may seem trivial, it is clear that the many terms currently used to describe interim assessments (benchmark, periodic, predictive, formative) have impeded clear discussions and decisions about whether and how interim assessments should be used.

Second, the research supporting the efficacy of assessment to improve teaching and learning is based on formative assessment—minute-by-minute classroom assessment. While interim assessment has considerable intuitive appeal, there simply is no research base to support the claim that interim assessments improve student learning.

Third, there are useful and valid purposes for interim assessments within a comprehensive assessment system. However, in deciding whether implementing an interim assessment system is an appropriate strategy

and, more specifically, what interim assessment design is appropriate, policymakers must go through the analysis laid out above of the purpose and expected use of interim assessment.

Fourth, policymakers should evaluate commercially available assessments cautiously. In particular, if policymakers desire interim assessments to serve instructional purposes, they should ask whether they meet the suggested criteria of effective assessments.

Fifth, policymakers should seek to eliminate the "zone of wishful thinking" in the design and implementation of interim assessment systems. Policymakers often hope that data will automatically lead to improved practice. However, experience shows that data must be accompanied by the reporting systems, professional development, support structures, and management practices that will impact teacher and student beliefs and behaviors. Each of these elements should be considered at the initial phases of designing or selecting and implementing an interim assessment system.

Finally, as with any assessments, policymakers should regularly ask whether the benefits of interim assessments outweigh the costs in terms of instructional time, teacher time, and fiscal resources. Further, they should be considered in light of the possibility of providing professional development to implement true formative assessment practices. One reason school districts invest in interim assessment systems rather than promoting formative classroom assessment may be that they lack the capacity to implement formative assessment well at scale. As Black and Wiliam have noted, "The improvement of formative assessment cannot be a simple matter. There is no quick fix that can alter existing practice by promising rapid rewards. On the contrary, if the substantial rewards promised by the research evidence are to be secured, each teacher must find his or her own ways of incorporating the lessons and ideas set out above into his or her own patterns of classroom work and into the cultural norms and expectations of a particular school community. This process is a relatively slow one and takes place through sustained programs of professional development and support."[2]

Leaders interested in improving formative assessment practices should support interim assessment systems designed with explicit attention to

increasing teachers' ability to do formative classroom assessment. The choice of item types, the format of reports and data analysis, and the structure and content of professional development can be carried out in ways that help teachers learn how to embed assessment within a learning activity, provide immediate corrective feedback, and modify instruction to meet students' needs. Over the long term, the focus of assessment efforts can move from interim assessment to the formative assessment practices that research suggests have the most payoff for student learning.

(This chapter was adapted from *The Role of Interim Assessments in a Comprehensive Assessment System: A Policy Brief*, by Marianne Perie, Scott Marion, Brian Gong of the National Center for the Improvement of Education Assessment, and Judy Wurtzel of the Aspen Institute.)[3]

The views expressed in this chapter are those of the authors and do not necessarily represent those of the Alliance for Excellent Education.

About the Authors

Judy Wurtzel* is codirector of the Aspen Institute Program on Education and Society. The program helps local, state, and national education leaders share knowledge about how school systems can improve the education and life chances of poor and minority students, and works with them to create programs and policies to accomplish these goals. Current initiatives include the Aspen Senior Congressional Staff Network, the Aspen Urban Superintendents' Network and complementary Urban Literacy and Mathematics Leadership Networks, and a project on rethinking human capital in urban school districts. From 1999 to 2005, Ms. Wurtzel was executive director of the Learning First Alliance, a permanent partnership of twelve leading educational associations dedicated to improving public education. During the Clinton administration, she was a senior adviser to the deputy secretary of the U.S. Department of Education, working on a range of elementary and secondary education issues.

Scott Marion is the vice president of the National Center for the Improvement in Educational Assessment, where his current projects include evaluating the technical quality of state alternate assessment systems, exploring the instructional usefulness of interim assessment approaches, and helping states design valid accountability systems. Dr. Marion serves on the U.S. Department of Education's National Technical Advisory Committee and on a national research committee investigating the issues

and challenges associated with incorporating value-added measures in educational accountability systems. Dr. Marion received his PhD in measurement and evaluation from the University of Colorado at Boulder, an MS in science education from the University of Maine, and a BA in biology from the State University of New York. Prior to joining the Center for Assessment, Dr. Marion was most recently the director of assessment and accountability for the Wyoming Department of Education and was responsible for overseeing the Wyoming Comprehensive Assessment System and designing the technical and policy structures to implement a multiple-measures, locally created graduation assessment system. Dr. Marion regularly publishes and presents the results of his work in peer-reviewed journals and at several national conferences. For more information about the Center for Assessment and Dr. Marion, please see www.nciea.org.

Marianne Perie is a senior associate with the National Center for the Improvement of Educational Assessment. She received her PhD in educational research, measurement, and evaluation from the University of Virginia. Prior to joining the center, she worked on district, state, and international assessments as well as the National Assessment of Educational Progress (NAEP) as an employee of first the American Institutes for Research and then the Educational Testing Service. Her primary interests are standard setting, reporting, accountability, and validity studies. She has conducted standard-setting studies in more than sixteen states, districts, and foreign countries. She taught a course in standard setting as part of the federally funded Graduate Certificate Program and coauthored a revision of the 1982 publication *Passing Scores*, published in 2008 as *Cutscores*. Dr. Perie is currently working with several states on exploring a validity argument on alternate assessments for students with significant cognitive disabilities and enhancing their technical documentation. For other publications, see www.nciea.org and click on "publications."

Brian Gong is the executive director and cofounder, with Rich Hill, of the National Center for the Improvement of Educational Assessment. Dr. Gong's experience as the associate commissioner for curriculum, assessment, and accountability in the Kentucky Department of Education, in addition to his previous work in research and development at Educational Testing Service, allows him to bring a multifaceted approach to solving the complex educational problems. Dr. Gong was the author of an important Council of Chief State School Officers publication on the design of educational accountability systems, and continues to advise several of the council's projects. Dr. Gong received his PhD in curriculum and instruction from Stanford University. Additionally, Dr. Gong currently serves as an advisor to the U.S. Department of Education as it considers how to incorporate student growth-based models into state accountability systems for NCLB.

---

*On May 19, 2009, Judy Wurtzel was named deputy assistant secretary for planning, policy, and evaluation at the U.S. Department of Education.

[1] L. Shepard, "Benchmark Assessments Be Formative?: Distinguishing Formative Assessment from Formative Program Evaluation," paper presented at the CCSSO Large Scale Assessment Conference, June 2006, San Francisco, CA.

[2] P. Black and D. Wiliam, "Assessment and Classroom Learning," *Educational Assessment: Principles, Policy and Practice* 5, no. 1 (1998): 7–74; P. Black and D. Wiliam, "Inside the Black Box: Raising Standards Through Classroom Assessment," *Phi Delta Kappan* 80 (October 1998).

[3] M. Perie, S. F. Marion, B. Gong, and J. Wurtzel, *The Role of Interim Assessments in a Comprehensive Assessment System: A Policy Brief* (Washington, DC: Achieve, Inc., the Aspen Institute, and the National Center for the Improvement of Educational Assessment Inc., 2007).

CHAPTER 5

International Assessments of Student Learning Outcomes

Andreas Schleicher
Organisation for Economic Co-operation and Development

Introduction

Parents, students, and educators who teach and run education systems seek good information on how well their education systems prepare students for life. Most countries now monitor students' learning and the functioning of schools in order to provide answers to this question: among the thirty Organisation for Economic Co-operation and Development (OECD) countries and six other countries with comparable data, twenty-two countries undertake student examinations and/or assessments and seventeen require schools to be evaluated (either self-evaluations and/or inspections by an external body) at regular intervals. For student performance measures, student assessments (evaluations without direct consequences for the individual student) are used in seventeen countries, whereas national examinations (with direct consequences for the individual student) are used in ten OECD countries.

Comparative international assessments can extend and enrich the national picture by providing a larger context within which to interpret national

performance. They have gained prominence over recent years because the benchmarks for public policy in education are no longer solely national goals or standards, but increasingly the performance of the most successful education systems internationally.[1] International assessments can provide countries with information that allows them to identify areas of relative strengths and weaknesses and monitor the pace of progress of their education system. They can also stimulate countries to raise aspirations by showing what is possible in education in terms of the quality, equity, and efficiency of educational services provided elsewhere, and they can foster better understanding of how different education systems address similar problems.

Following a brief introduction to the history of international assessments, this chapter sets out the potential that international assessments offer for educational policy and practice as well as some of the methodological challenges they face in providing valid, comparable, and reliable evidence.

History of International Assessments

While efforts to compare education systems internationally can be traced back to the early nineteenth century,[2] the discourse on international comparisons of learning outcomes started to emerge during the 1950s and '60s. In 1958, an expert group led by William Douglas Wall and including prominent researchers such as Benjamin Bloom, Robert Thorndike, Arthur Wellesley Forshay, Arnold Anderson, Gaston Mialaret, and Torsten Husen met under the auspices of UNESCO's International Institute of Education in Hamburg, Germany, to launch a feasibility study to compare student performance internationally. The feasibility study involved twelve thousand thirteen-year-olds in twelve countries, and its results were published in 1962.[3] The International Association for the Evaluation of Educational Achievement (IEA) emerged out of this collaboration, and later conducted a series of international assessments. The most prominent regular survey carried out by the IEA is the quarterly Trends in Mathematics and Science Study (TIMSS), which assesses fourth- and eighth-grade students' acquisition of math and science skills, and the Progress in Reading Literacy Study (PIRLS), which is given five times a year and measures reading literacy achievement of fourth-grade students.

The U.S. Education Testing Service conducted the International Assessment of Educational Progress (IAEP)[4] in 1998 and a follow-up study in 1991.[5] The latest generation of international assessments has been developed by the OECD as part of the Program for International Student Assessment (PISA). PISA surveys have been given every three years since 2000 in key content areas such as reading, mathematics, and science, but they also cover cross-curricular domains such as problem solving and a range of noncognitive outcomes. PISA is currently the most rigorous and also the most comprehensive international assessment, not least in terms of its coverage of subject areas and its geographic coverage, with the latest survey in 2009 testing more than 400,000 students in over seventy countries that together comprise close to 90 percent of the world economy. To implement the assessment, each country draws a random sample of between 3,500 and 50,000 fifteen-year-olds enrolled in school. Each participating student spends two hours carrying out pencil-and-paper tasks, solving electronically delivered problems, and answering multiple-choice questions. Students also answer a questionnaire focused on their personal background, their learning habits, and their engagement with and motivation at school. Principals complete a questionnaire about their school that includes demographic characteristics and an assessment of the quality of the school's learning environment.

Research Frameworks of International Assessments

The international assessments of the OECD and IEA seek to contextualize measures of student learning outcomes with background information collected from students, principals, and sometimes teachers and parents in order to interpret the observed variation in learning outcomes between students, classrooms, schools, and education systems.

To facilitate this, the tests operate with research frameworks that typically address **three research areas** (learning outcomes, policies shaping education outcomes, and factors that constrain policies and outcomes) with data at up to **four levels of the education system** (individual learners, classrooms or instructional settings, educational institutions and providers of educational services, and the education system as a whole). (See Table 1.)

These international assessments can then be used to address a variety of research issues from different perspectives relating, for example, to the quality of educational outcomes, to issues of equality of educational outcomes and equity in educational opportunities, or to the adequacy, effectiveness, and efficiency of resource management.

Table 1: Research Frameworks for International Assessments

		Research areas		
		Education and learning outputs and outcomes	Policy levers and contexts shaping educational outcomes	Constraints that contextualize policies and outcomes
Levels of education system	Individual participants in education and learning	The quality and distribution of individual educational outcomes	Individual attitudes, engagement, and behavior	Background characteristics of the individual learners
	Instructional settings	The quality of instructional delivery	Curriculum, pedagogy and learning practices, and classroom climate	Student learning conditions and teacher working conditions
	Providers of educational services	The output of educational institutions and institutional performance	School environment and organization	Characteristics of the service providers and their communities
	The education system as a whole	The overall performance of the education system	System-wide institutional settings, resource allocations, and policies	The national educational, social, economic, and demographic contexts

The Potential of International Assessments for Policy and Practice

The design and conduct of international assessments was originally motivated by research objectives. More recently, governments have begun to attribute growing importance to international assessments and have invested considerable resources into their development and implementation. This interest derives from several considerations:

- There is increasing recognition in many countries that the yardstick for educational success is no longer improvement by national standards but the performance of the best-performing education systems internationally. By **revealing what is possible in education** in terms of the performance levels demonstrated in the countries that perform strongest in international comparisons, international assessments can enhance the quality of existing policies but also create a debate about the paradigms and beliefs underlying policies. While international assessments alone cannot identify cause-and-effect relationships between inputs, processes, and educational outcomes, they can shed light on key features in which education systems show similarities and differences, and make those key features visible for educators, policymakers, and the general public. This, in turn, can generate powerful hypotheses for further analysis and research.

- In some countries, international assessments are also used to **set policy targets** in terms of measurable goals achieved by other systems, and seek to identify policy levers and establish trajectories as well as delivery chains for reform. In a number of countries, international assessments are also used to contextualize national standards and assessments.

- International assessments can assist with gauging the pace of educational progress, through assessing to what extent **achievement gains** observed nationally are in line with achievement gains observed elsewhere.

- Finally, international assessments can **support the politics** of educational reform, which is a major issue in education, where

any payoff to reform almost inevitably accrues to successive governments, if not generations.

These issues are examined more closely in the remainder of this section.

Revealing what is possible in education and identifying factors that contribute to educational success

International assessments seem to impact more on countries whose performance is comparatively low. Although it is sometimes argued that weighing the pig does not make it fatter, diagnosing underweight can be an important first step toward treatment. Also, as the level of public awareness was raised by international comparisons, it has in some countries created an important political momentum and engaged educational stakeholders, including teacher and/or employer organizations, in support of policy reform.[6]

Equally important, international assessments have had a significant impact in some countries that did not do poorly in absolute terms but found themselves confronted with results that differed from how educational performance was generally perceived in that country. (See, for example, the profile on Germany's experiences in the box on the opposite page.)

Showing that strong educational performance and improvement are possible seems to be one of the most important aspects of international assessments. Whether in Asia (like in Japan, Korea, or Singapore), in Europe (like in Finland or the Netherlands), or in North America (like in Canada), many countries display strong overall performance in PISA, and, equally important, some of these countries also show that poor performance in school does not automatically follow from a disadvantaged socioeconomic background. In addition, some countries show that success can become a consistent and predictable educational outcome. In Finland, for example, the country with the strongest overall results in PISA, the performance variation between schools amounted in 2006 to only 5 percent of students' overall performance variation on PISA. So parents can rely on high and consistent performance standards in whatever school they choose to enroll their children. Considerable research has been invested in the features of these education systems. In some countries, governments have used

knowledge provided by PISA as a starting point for a peer review to study policies and practices in countries operating under similar context that achieve better results.[7] Such peer reviews, each resulting in a set of specific policy recommendations for educational improvement, are also being carried out by the OECD, the results of which have been published so far for Denmark and Scotland.[8]

Profile: Germany

In Germany, equity in learning opportunities across schools was historically often taken for granted, as significant efforts were devoted to ensuring that schools were adequately and equitably resourced. The results from the PISA 2000 assessment, however, revealed large socioeconomic disparities in educational outcomes between schools. Further analyses separated equity-related issues between those that relate to the socioeconomic heterogeneity within schools and those that relate to socioeconomic segregation through the school system. These results taken together suggested that German students from more privileged social backgrounds were being directed into the more prestigious academic schools, which yielded superior educational outcomes, while students from less privileged social backgrounds were being directed into less prestigious vocational schools, which yielded poorer educational outcomes, even where their performance on the PISA assessment was similar.

This raised the specter that the German education system was reinforcing rather than moderating socioeconomic background factors. Such conclusions, and the ensuing vivid public debate, inspired a wide range of equity-related reform efforts in Germany, some of which have been transformational in nature. These include

- giving early childhood education, which had hitherto been considered largely an aspect of social welfare, an educational orientation;
- establishing national educational standards for schools in a country in which regional and local autonomy had long been the overriding paradigm;
- introducing full-day schooling in a system where half-day schooling had been the norm for centuries; and
- enhancing the support for disadvantaged students, such as students with a migration background.

For many educators and experts in Germany, the socioeconomic disparities that PISA revealed were unsurprising. However, it had often been taken for granted and outside the scope of public policy that disadvantaged children would fare less well in school. The fact that PISA revealed that the impact that socioeconomic background has on students and school performance varied considerably across countries, and that other countries appeared to moderate socioeconomic disparities much more effectively, showed that improvement was possible—and provided the momentum for policy change.

As a result, the benchmarks for public policy in education are no longer national goals or standards alone, but increasingly the performance and achievement gains of the most successful education systems measured internationally. International assessments have at times raised awareness, leading to a public debate about education in which citizens have recognized that their country's educational performance will not just need to match average performance, but will have to do better if their children want to justify above-average wages.

Putting national targets into a broader perspective

International assessments can also play an important role in putting national performance targets into perspective. Educators are often faced with a dilemma: if, at the national level, the percentage of students achieving good exam scores in school increases, some will claim that the school system has improved. Others will claim that standards must have been lowered, and behind the suspicion that better results reflect lowered standards is often a belief that overall performance in education cannot be raised. International assessments allow those perceptions to be related to a wider reference framework by allowing schools and education systems to compare themselves with schools and education systems in other countries. Some countries have actively embraced this perspective and systematically related national performance to international assessments. Australia and Germany, for example, have embedded national items into the PISA assessments in order to relate what is considered important nationally to what is valued in other countries. Conversely, Japan has embedded PISA-type questions into its national assessment. By their very nature, international assessments assess aspects of students' skills and knowledge that are not *completely* covered by *all* national curricula, simply because curricula vary across countries. So they require national experts and authorities to examine what are the dimensions covered and uncovered in their schools, then to decide whether the uncovered ones should or should not be taught. When a country discovers that its students are unable to do things that students in other countries can do, the crucial question is, "Do *our* students need these skills too, to be able to survive in our modern society?" If the answer is yes, there is an opportunity to review and improve the standards, assessments, and curriculum.

Assessing the pace of change in educational improvement

A third important aspect is that international comparisons provide a frame of reference to assess the pace of change in educational development. While a national framework allows progress to be assessed in absolute terms, an internationally comparative perspective allows an assessment of whether that progress matches the pace of change observed elsewhere. Indeed, while all education systems in the OECD area have seen quantitative growth over past decades, international comparisons reveal that the pace of change in educational output has varied markedly.

For example, among fifty-five- to sixty-four-year-olds, the United States is well ahead of all other OECD countries in terms of the proportion of individuals with both school and university qualifications. However, international comparisons show that this lead is largely a result of the "first-mover advantage" that the United States gained after World War II by massively increasing school enrollments. This gain has eroded over the last few decades as more and more countries have reached and surpassed qualification levels in the United States in younger cohorts. While many countries are now close to ensuring that virtually all young adults leave schools with at least a high school qualification—which the OECD benchmarks highlight as the baseline qualification for reasonable earnings and employment prospects—the United States has stood still on this measure, and among OECD countries only New Zealand, Spain, Turkey, and Mexico now have lower secondary school completion rates than the United States.[9]

In contrast, two generations ago, South Korea had the economic output of Afghanistan today and was ranked twenty-fourth in terms of schooling performance among today's OECD countries. Today it is the top performer in the proportion of successful school leavers, with 96 percent of an age cohort obtaining a high school degree. While progress from a national perspective matters, in this global framework the internationally comparative perspective is having a growing impact not just on public policy, but on institutional behavior as well. The results of international assessments of student performance are beginning to demonstrate similar influence.

A tool for changing the politics of education reform

International assessments can also affect the politics of education reform. For example, in the 2007 Mexican national survey of parents, 77 percent of parents interviewed reported that the quality of educational services provided by their children's school was good or very good. But in OECD's PISA 2006 assessment, roughly half of the Mexican fifteen-year-olds who were enrolled in school performed at or below the lowest level of proficiency established by PISA.[10] There may be many reasons for this kind of discrepancy between perceived educational quality and performance on international assessments—it may be due in part, for instance, to the fact that the educational services that Mexican children receive are significantly better than what their parents experienced. However, the point here is that justifying the investment of public resources in areas for which there seems no public demand poses difficult challenges. One recent response by the Mexican presidential office was to include a "PISA performance target" in the new Mexican education reform plan. This performance target—based on the outcome of international assessments, and set to be achieved by 2012—will serve to highlight the gap between national performance and international standards, and monitor how educational improvement feeds into closing this gap. It is associated with a reform trajectory and delivery chain of support systems, incentive structures, and improved access to professional development to assist school leaders and teachers in meeting the target. Such reforms draw on the experience of other countries. Brazil has taken a similar route, providing each secondary school with information on the level of progress that is needed to perform at the OECD average performance level on PISA in 2021.

Japan is one of the best-performing education systems on the various international assessments. However, PISA results revealed that while Japanese students tended to do very well on tasks that require reproducing subject matter content, they did much less well on open-ended constructed tasks requiring them to demonstrate their capacity to extrapolate from what they know and apply their knowledge in novel settings. Conveying that situation to parents and a general public used to certain types of tests providing the gateway to further education poses a challenge for reform too. The policy response in Japan has been to incorporate "PISA-type" open-constructed tasks into the national assessment, with the aim that

skills that are considered important internationally will become valued in the national education system. Similarly, Korea has recently incorporated advanced PISA-type literacy tasks in its university entrance examinations, in order to enhance excellence in the capacity of its students to access, manage, integrate, and evaluate written material. In both countries, these changes represent transformational change that would have been much harder to imagine without the challenges revealed by PISA.

Design Issues and Challenges for International Assessments

The design of international assessments of learning outcomes needs to fulfill different and sometimes competing demands.

- International assessments need to ensure that their **outcomes are valid** across cultural, national, and linguistic boundaries, and that the target populations from which the samples in the participating countries are drawn are comparable.

- International assessments need **to offer added value** to what can be accomplished through national assessment and analysis.

- While international assessments need to be as comparable as possible, they also need to **be country specific** so they can adequately capture historical, systemic, and cultural variation among countries.

- The measures need to be as simple as possible to be widely understood, while remaining as complex as necessary to **reflect multifaceted educational realities**.

- While there is a general desire to keep any set of performance measures as small as possible, the picture should not be reduced to a small common denominator that no longer represents the variability of approaches and policy issues across countries, since this variability provides the foundation for countries to learn from each other's experiences.

Important issues that arise in meeting these demands are examined in the remainder of this section in more detail.

Cross-country validity and comparability in the assessment instruments

International assessments necessarily are limited in their scope for several reasons. First, there is no overarching agreement, across countries, on what students in a particular grade or at a particular age should know and be able to do—often referred to as "competencies." Second, any single assessment can only measure a selection of such competencies. Lastly, there are various methodological constraints that limit the kinds of competencies that currently can be measured through large-scale assessment.

International assessments have made considerable progress toward assessing knowledge and skills in content areas such as mathematics, reading, science, and problem solving. However, they are still limited in the coverage of important cognitive outcomes, in particular the assessment of creative competencies. Similarly, achieving high degrees of objectivity in the assessments, which favor multiple-choice tasks that can be scored without human judgment, tends to detract from the assessment of the higher-order competencies and the production of knowledge, which require open-ended assessment tasks. At times, in order to make the assessments affordable to lower-income countries, international assessments have also sacrificed validity gains over efficiency gains, by giving undue weight to assessment tasks that can be easily administered and scored. Even less progress has been made to assess interpersonal dimensions of competencies that are often recognized as of increasing importance, such as the capacity of students to relate well to others or to manage and resolve conflicts. Last but not least, international assessments provide only very crude self-reported measures of intrapersonal dimensions of competencies.

Establishing the assessment domains

Even in established content areas, internationally comparative measurement poses major challenges. Countries vary widely in their intended, implemented, and achieved curricula. Inevitably, international assessments need to strike a balance between narrowing the focus to what is common across the different curricula of school systems, on the one hand, and capturing a wide enough range of competencies to reflect the content domains to be assessed adequately, on the other. Leaning toward the former—as has been the tendency for the assessments of the IEA—ensures

that what is being tested internationally reflects what is being taught in all countries. This is an important aspect of fairness, but there is a risk that the assessment reflects just the lowest common denominator of national curricula. It also lacks important aspects of curricula that are not taught in all of the countries, as well as the content validity that is required to faithfully represent the relevant subject area. Leaning toward the latter—as is the case for the assessments of the OECD, with their focus on the capacity of students not merely to reproduce what they have learned but to extrapolate from what they have learned and apply their knowledge and skills in novel settings—enhances content validity but risks that students are being confronted with assessment material they may not have been taught in their national context.

In whatever way the various international assessments have struck these balances, they have tried to build them through a carefully designed interactive process between the agencies developing the assessment instruments, various international expert groups working under the auspices of the respective organizations, and national experts charged with the development and implementation of the surveys in their countries. Often, a panel of international experts, in close consultation with participating countries, has led the identification of the range of knowledge and skills in the respective assessment domains that have been considered to be crucial for students' capacity to fully participate in and contribute to a successful modern society. A description of the assessment domains—the assessment framework—was then used by participating countries and other test development professionals as they contributed assessment materials.

For example, in the development of PISA, this involved

- the development of a working definition for the assessment area and a description of the assumptions that underlay that definition;
- an evaluation of how to organize the set of tasks constructed in order to report to policymakers and researchers on performance in each assessment area among fifteen-year-old students in participating countries;
- the identification of a set of key characteristics to be taken into account when assessment tasks were constructed for international use;

- the operationalization of the set of key characteristics to be used in test construction, with definitions based on existing literature and the experience of other large-scale assessments;
- the validation of the variables, and assessment of the contribution that each made to the understanding of task difficulty in participating countries; and
- the preparation of an interpretative scheme for the results.

The PISA assessment is defined through three interrelated dimensions: the knowledge or structure of knowledge that students need to acquire (e.g., familiarity with scientific concepts); competencies that students need to apply (e.g., carrying out a particular scientific process); and the contexts in which students encounter scientific problems and relevant knowledge and skills are applied (e.g., making decisions in relation to personal life, understanding world affairs). (See Table 2.)

Once the assessment framework is established and agreed upon (which tends to be the most challenging aspect of an international assessment), assessment items are developed to reflect the intentions of the frameworks, and they need to be carefully piloted before final assessment instruments can be established. To some extent, the question of to what extent the tasks in international assessments are comparable across countries can be answered empirically. Analyses to this end were first undertaken for the IEA Trends in Mathematics and Science Study.[11] The authors compared the percentage of correct answers in each country according to the international assessment as a whole with the percentage correct in each country on the items said by the country to address its curriculum in mathematics. Singapore, for example, had 144 out of 162 items that were said to be covered by the Singaporean curriculum. The percentage of items correct on the whole test and on the items covered in the curriculum was seventy-nine in both cases.

Singapore also scored between 79 and 81 percent correct on the items that other countries considered covered in their own curricula. These ranged from seventy-six items in Greece to 162 items in the United States. For most countries, the results were similarly consistent, suggesting that the composition of the tests had no major impact on the relative standing of countries in the international comparisons. Such analyses have also been conducted for PISA, and have yielded similar results.

Table 2: Defining an Assessment Domain—An Example from PISA	
	Science
Definition and its distinctive features	The extent to which an individual • possesses scientific knowledge and uses that knowledge to identify questions, acquire new knowledge, explain scientific phenomena, and draw evidence-based conclusions about science-related issues; • understands the characteristic features of science as a form of human knowledge and inquiry; • shows awareness of how science and technology shape our material, intellectual, and cultural environments; and • engages in science-related issues and with the ideas of science, as a reflective citizen. Scientific literacy requires an understanding of scientific concepts, as well as the ability to apply a scientific perspective and to think scientifically about evidence.
Knowledge domain	Knowledge *of* science, such as: • "Physical systems" • "Living systems" • "Earth and space systems" • "Technology systems" Knowledge *about* science, such as: • "Scientific inquiry" • "Scientific explanations"
Competencies involved	Type of scientific task or process: • Identifying scientific issues • Explaining scientific phenomena • Using scientific evidence
Context and situation	The area of application of science, focusing on uses in relation to personal, social, and global settings such as • "Health" • "Natural resources" • "Environment" • "Hazard" • "Frontiers of science and technology"

Reflecting national, cultural, and linguistic variety

International assessments pay close attention to reflecting the national, cultural, and linguistic variety among participating countries. OECD's PISA assessments employ the most sophisticated and rigorous process to this end. The agency charged with the development of the instruments uses professional test item development teams in several different countries. In addition to the items developed by these teams, assessment material is contributed by participating countries and is carefully evaluated and matched against the framework. Furthermore, each item included in the assessment pool is rated by each country: (1) for potential cultural, gender, or other bias; (2) for relevance to the students to be assessed in school and nonschool contexts; and (3) for familiarity and level of interest.

Selecting assessment nature and form

Also important is the nature and form of the assessment, as reflected in the task and item types. While, as noted before, multiple-choice tasks are the most cost-effective way to assess knowledge and skills, and have therefore dominated earlier international assessments, they have important limitations in assessing more complex skills, particularly ones that require students not just to recall but to produce knowledge. Moreover, since the nature of assessment tasks, and in particular student familiarity with multiple-choice tasks, varies considerably across countries, heavy reliance on any single item type such as multiple-choice tasks can be an important source of response bias. The PISA assessments have tried to address this through employing a broad range of assessment tasks, with about 40 percent of the questions requiring students to construct their own responses. Other tasks require students to either provide a brief answer (short-response questions) or construct a longer response (open-constructed-response questions), allowing for the possibility of divergent individual responses and an assessment of students' justification of their viewpoints. Partial credit can be given for partly correct or less complex answers, with answers judged by trained specialists (or "coders") using detailed scoring guides. Open-ended assessment tasks, however, raise other challenges, in particular the need to ensure inter-rater reliability in the results. For PISA, there are a number of checks in place to ensure reliability. First, samples of the assessment booklets are coded independently by four coders and examined

by the international contractor. Second, an inter-coder reliability study and a homogeneity analysis are currently being implemented to examine the consistency of this coding process in more detail within each country, and to estimate the magnitude of variance associated with the use of coders. Third, an international coding review is now examining how consistently the response-coding standards are being applied across all participating countries, with the goal of estimating potential bias (either leniency or harshness) in the coding standards applied in participating countries. Lastly, in order to measure the intended broad range of content while meeting the limits of individual assessment time, PISA, like most modern international assessments, is now using multiple test forms within a country's test population.

Ensuring external validity

Ensuring that international assessments are comparable across countries is one thing, but the more important challenges relate to their external validity, which involves verifying that the assessments measure what they set out to measure. An important question is whether the knowledge and skills that are being assessed are predictive for the future success of students. In the case of PISA, the Canadian Youth in Transition Survey (YITS), a longitudinal survey that investigates patterns of and influences on major educational, training, and work transitions in young people's lives, provided a way to examine this empirically. In 2000, 29,330 fifteen-year-old students in Canada participated in both YITS and PISA. Four years later, the educational outcomes of the same students, then aged nineteen, were assessed, and the association of these outcomes with PISA reading performance at age fifteen was investigated.[12] The results showed that students who had mastered PISA performance Level 2 on the PISA reading test at age fifteen were twice as likely to participate in postsecondary education at age nineteen as those who performed at Level 1 or below, even after accounting for school engagement, gender, mother tongue, place of residence, parental education, and family income. The odds increased to eightfold for those students who had mastered PISA Level 4 and to sixteenfold for those who had mastered PISA Level 5. A similar study undertaken in Denmark led to similar results, in that the percentage of youth who had completed post-compulsory, general, or vocational upper-secondary education by the age of nineteen increased significantly with

their reading ability as assessed by PISA at age fifteen (see http://www.sfi.dk/sw19649.asp). Last but not least, the International Adult Literacy Study allowed reading and numeracy skills (defined in similar ways to those measured by PISA) to be related to earnings and employment outcomes in the adult population, and the analyses showed that such indicators were generally a better predictor for individual earnings and employment status than the level of formal qualification individuals had attained.[13]

Comparability of the target populations

Even if the assessment instruments are valid and reliable, meaningful comparisons can only be made if the target populations being assessed are also comparable. International assessments therefore need to use great care when defining comparable target populations, ensuring that they are exhaustively covered with minimal and well-defined population exclusions, and ensuring that the sampled students do participate in the assessment.

As regards defining target populations, important trade-offs need to be made between international comparability and relating the target populations to national institutional structures. Differences between countries in the nature and extent of pre-primary education and care, the age of entry to formal schooling, and the institutional structure of educational systems do not allow the establishment of internationally comparable grade levels. Consequently, international comparisons of educational performance typically define populations with reference to a target age group. International assessments of the IEA have defined these target groups on the basis of the grade level that provides maximum coverage of a particular age cohort (such as the grade in which most thirteen-year-olds are enrolled). The advantage of this is that a grade level can be easily interpreted within the national institutional structure and provides a cost-effective way toward assessment, with minimal disruption of the school day. However, a disadvantage is that slight variations in the age distribution of students across grade levels often lead to the selection of different target grades in different countries, or between education systems within countries, raising serious questions about the comparability of results across, and at times within, countries. In addition, because not all students of the desired age are usually represented in grade-based samples, there may be a more serious potential bias in the results if the unrepresented students

are typically enrolled in the next higher grade in some countries and the next lower grade in others. This excludes students with potentially higher levels of performance in the former countries and students with potentially lower levels of performance in the latter. To address these problems, the assessments of the OECD use an age-based definition for their target populations. For example, PISA assesses students who were between fifteen years and three (complete) months and sixteen years and two (complete) months at the beginning of the assessment period and who were enrolled in an educational institution, regardless of the grade level or type of institution in which they were enrolled and whether they were in full-time or part-time education. The disadvantages of this age-based approach is that it is costly, that the assessment process becomes more disruptive, and that it is more difficult to relate the results of individual students to teachers and classrooms.

The accuracy of any survey results also depends on the quality of the information on which national samples are based as well as on the sampling procedures. For the latest international assessments, advanced quality standards, procedures, instruments, and verification mechanisms have been developed that ensure that national samples yielded comparable data and that the results could be compared with confidence.

Even the best international samples will only translate into comparable results if the sampled schools are willing to take part in the assessment. While most countries participating in PISA now achieve high response rates, some countries, most notably the United States, have faced major challenges in securing school participation. At times, schools do not perceive sufficient benefit from an assessment that only yields national outcomes. Some countries have started to link PISA more closely to participating schools, either through providing them with school-level outcomes from the assessment or the related questionnaires, or through the provision of better information on the objectives and nature of these assessments. Incentives or feedback that have been deployed or are being considered by countries include

- better explanation of the context and usefulness of PISA at the start of the process to help engage teachers and schools;

- preparing a briefing pack to prepare teachers, schools, pupils, and parents to overcome pupils' initial anxieties and stimulate better communication within schools;
- setting up an international buddies scheme with schools doing PISA in other countries, in particular for sharing ideas for using the results to improve education;
- giving out student certificates on the day, perhaps as part of a small awards ceremony;
- encouraging PISA to be seen as a whole-school issue and to ensure corresponding dissemination;
- preparing electronic versions of feedback—perhaps in PowerPoint format to allow easier dissemination among staff;
- sharing good practice on what schools did with the feedback on the PISA website; and
- making the student questionnaire accessible so that schools can use it for benchmarking whenever they want and with a wider range of students.

Comparability in survey implementation

Well-designed international assessment needs to be well implemented to yield reliable results. The process begins with ensuring consistent quality and linguistic equivalence of the assessment instruments across countries. PISA, which provides the most advanced available procedures to this end, seeks to achieve this through providing countries with equivalent source versions of the assessment instruments in English and French and requiring countries (other than those assessing students in English and French) to prepare and consolidate two independent translations using both source versions. Precise translation and adaptation guidelines are supplied, including instructions for the selection and training of the translators. For each country, the translation and format of the assessment instruments (including test materials, marking guides, questionnaires, and manuals) is verified by expert translators appointed by the agency charged with the development of the assessment instruments (whose mother tongue was the language of instruction in the country concerned and who were knowledgeable about education systems) before they are used.

The assessments then need to be implemented through standardized procedures. Comprehensive manuals typically explain the implementation of the survey, including precise instructions for the work of school coordinators and scripts for test administrators for use during the assessment sessions. Proposed adaptations to survey procedures, or proposed modifications to the assessment session script, are reviewed internationally before they are employed at a national level. In the case of PISA, specially designated quality monitors visited all national centers to review data-collection procedures and school quality. Monitors from the international agency visited a sample of fifteen schools during the assessment. Marking procedures are designed to ensure consistent and accurate application of the internationally agreed-upon marking guides.

Recommendations and Conclusion

In a globalized world, the benchmarks for public policy in education are no longer national goals or standards alone, but increasingly the performance of the most successful education systems internationally. International assessments can be powerful instruments for educational research, policy, and practice by allowing education systems to look at themselves in the light of intended, implemented, and achieved policies elsewhere. They can show what is possible in education in terms of quality, equity, and efficiency in educational services, and they can foster better understanding of how different education systems address similar problems. Most importantly, by providing an opportunity for policymakers and practitioners to look beyond the experiences evident in their own systems and thus to reflect on some of the paradigms and beliefs underlying these, they hold out the promise to facilitate educational improvement. As this chapter has shown, designing and implementing valid and reliable international assessments poses major challenges, including defining the criteria for success in ways that are comparable across countries while remaining meaningful at national levels, establishing comparable target populations, and carrying the surveys out under strictly standardized conditions. However, more recently, international assessments such as PISA have made significant strides toward this end.

Some contend that international benchmarking encourages an undesirable process of degrading cultural and educational diversity among institutions and education systems, but the opposite can be argued as well: in the dark,

all institutions and education systems look the same, and it is comparative benchmarking that can shed light on differences on which reform efforts can capitalize. Who took notice of how Finland, Canada, or Japan ran their education systems before PISA revealed their success in terms of the quality, equity, and coherence of learning outcomes?

Of course, international assessments have their pitfalls, too. Policymakers tend to use them selectively, often in support of existing policies rather than as instruments to challenge them and to explore alternatives. Moreover, highlighting specific features of educational performance may detract attention from other features that are equally important, thus potentially influencing individual, institutional, or systemic behavior in ineffective or even undesirable ways. This can be like the drunk driver who looks for his car key under a street lantern and, when questioned whether he lost it there, responds that he didn't—but that it was the only place where he could see. This risk of undesirable consequences of inadequately defined performance benchmarks is very real, as teachers and policymakers are led to focus their work on the issues that performance benchmarks value and put into the spotlight of the public debate.

While the development of international assessments is fraught with difficulties and their comparability remains open to challenges, cultural differences among individuals, institutions, and systems should not suffice as a justification to reject their use, given that the success of individuals and nations increasingly depends on their global competitiveness. The world is indifferent to tradition and past reputations, unforgiving of frailty, and ignorant of custom or practice. Success will go to those individuals, institutions, and countries that are swift to adapt, slow to complain, and open to change. The task for governments will be to ensure that their citizens, institutions, and education systems rise to this challenge, and international benchmarks can provide useful instruments to this end.

There are specific actions U.S. leaders can take to use international assessments as a tool in evaluating progress and establishing effective policies toward the goal of graduating every student from high school prepared for college and the demands of the twenty-first-century global economy. These include

- measuring state-level education performance globally by examining student achievement and attainment in an international context to ensure that, over time, students are receiving the education they need to compete in the twenty-first-century economy (this can be achieved through participating at both the national and state levels in international studies like PISA that serve to collect data about student performance as well as related policies and practices);
- creating an ongoing public awareness and interest in the importance of international education comparisons by communicating results widely, encouraging discussion about findings, and partnering with key stakeholders;
- embedding international indicators into policy goals and decisionmaking processes; and
- employing the PISA framework as a tool in evaluating and improving U.S. standards and assessments.

The views expressed in this chapter are those of the author and do not necessarily represent those of the Alliance for Excellent Education.

About the Author

Andreas Schleicher is head of the Indicators and Analysis Division (Directorate for Education) at the Organisation for Economic Co-operation and Development (OECD). He also holds an honorary professorship at the University of Heidelberg in Germany. As division head at the OECD, his responsibilities include directing the Program for International Student Assessment (PISA) and the Indicators of Education Systems program (INES) and steering the development of new projects such as the OECD Teaching and Learning International Survey (TALIS) and the OECD Programme for the International Assessment of Adult Competencies (PIAAC). At the OECD, Mr. Schleicher has also held the posts of deputy head of the Statistics and Indicators Division in the former Directorate for Education, Employment, Labour and Social Affairs (1997–2002) and project manager in the OECD Centre for Educational Research and Innovation (CERI) (1994–1996).

Before joining the OECD, he served as director for analysis at the International Association for Educational Achievement (IEA) within the Institute for Educational Research in the Netherlands (1993–1994) and international coordinator for the IEA Reading Literacy Study, at the University of Hamburg, Germany (1989–1992).

In 2003, Mr. Schleicher was awarded the Theodor Heuss Prize, named after the first president of the Federal Republic of Germany, for "exemplary democratic engagement" in association with the public debate on PISA. In 2002, he was awarded the *educación y libertad en el ámbito educativo* prize by the Spanish national association of private schools. Mr. Schleicher earned a bachelor's degree in physics and a master's degree in mathematics from Deakin University in Australia, where his master's thesis received the Bruce Choppin Award.

---

[1] D. Hopkins, D. Pennock, and J. Ritzen, *Evaluation of the Policy Impact of PISA* (Paris: OECD, 2008).

[2] M. A. Jullien, *Esquisse et vues préliminaires d'un ouvrage sur l'éducation compare* (Paris: L. Colas, 1817).

[3] A. W. Foshay, R. L. Thorndike, F. Hotyat, D. A. Pidgeon, and D. A. Walker, *Educational Achievement of Thirteen-Year-Olds in Twelve Countries* (Hamburg: UNESCO Institute for Education, 1962).

[4] A. E. LaPointe, N. A. Mead, and G. W. Phillips, *A World of Differences: An International Assessment of Mathematics and Science* (Princeton, NJ: Educational Testing Service, 1989).

[5] A. E. LaPointe, N. A. Mead, and J. M. Askew, *The International Assessment of Educational Progress Report* (Princeton, NJ: Educational Testing Service, 1992).

[6] Hopkins, Pennock, and Ritzen, *Evaluation of the Policy Impact of PISA*.

[7] H. Döbert, E. Klieme, and W. Sroka, *Vertiefender Vergleich der Schulsysteme ausgewählter PISA-Teilnehmerstaaten* (Frankfurt: Deutsches Institut für pädagogische Forschung, 2004).

[8] OECD, *Reviews of National Policies for Education—Denmark: Lessons from PISA 2000* (Paris: OECD, 2004); OECD, *Reviews of National Policies for Education: Quality and Equity of Schooling in Scotland* (Paris: OECD, 2007).

[9] OECD, *Education at a Glance—OECD Indicators 2007* (Paris: OECD, 2008).

[10] IFIE-ALDUCIN, *Mexican National Survey to Parents Regarding the Quality of Basic Education* (Mexico City: IFIE-ALDUCIN, 2007); OECD, *Reviews of National Policies for Education: Quality and Equity of Schooling in Scotland*.

[11] A. E. Beaton, I. V. S. Mullis, M. O. Martin, E. J. Gonzales, D. L. Kelly, and T. A. Smith, *Mathematics Achievement in the Middle School Years* (Chestnut Hill, MA: Boston College Center for the Study of Testing, Evaluation, and Educational Policy, 1996).

[12] T. Knighton and P. Bussiere, *Educational Outcomes at Age 19 Associated with Reading Ability at Age 15* (Ottawa: Statistics Canada, 2006).

[13] OECD and Statistics Canada, *Literacy Skills for the Information Age* (Ottawa and Paris: Authors, 2000).

CHAPTER 6

Measuring Student Achievement Growth at the High School Level

Joseph Martineau
Michigan Department of Education

As education policies at the local, state, and federal levels increasingly include accountability for student achievement, and as the stakes attached to that accountability have risen, interest in various accountability models has grown substantially.

Most accountability models currently in use—including those initially implemented by states to comply with the requirements of the No Child Left Behind Act of 2001 (NCLB)—focus on an absolute level of student achievement. These "status-based" models primarily hold schools, districts, and states accountable for meeting a state-set percentage of students performing at some minimum achievement standard on state-administered assessments.

For example, consider the following scenario of two high schools in a state that uses a status-based accountability model for NCLB. High School A serves a high-challenge student population: only 30 percent of entering freshman scored "proficient" or above on the eighth-grade mathematics

exam. High School B serves a lower-challenge student population: 65 percent of entering freshman scored proficient or above on the eighth-grade mathematics exam. To make Adequate Yearly Progress (AYP) and avoid NCLB-mandated sanctions, both schools need to meet the state-set goal of 70 percent of students scoring proficient or above on the mathematics exam administered to tenth graders. Such status-based accountability models have come under criticism for a number of reasons, including the following:

- Regardless of student background characteristics, risk factors, and incoming education levels, both schools in our hypothetical scenario need to reach the same status target. That means that, over the same period of time, High School A is expected to achieve significantly more progress than High School B.

- There is significant pressure to meet the status goal. As a result, educators in both high schools must focus efforts on meeting the proficiency target, but the impact is greatest in High School A because there is more ground to make up. This has two unintended consequences:

    - First, in both schools, classroom teachers may narrow instruction to focus on the content and test-taking skills that will help students score proficient on the exam, at the expense of other rich content, but the incentive to do so is stronger in High School A.

    - Second, classroom teachers may pay disproportionate attention to the students who are just below the proficient level and can most easily be supported to score proficient and help the school meet its proficiency level. Meanwhile, little attention may be paid to the equally important progress of other students on the performance spectrum, including both those who are furthest behind and those who are already likely to score proficient or above. Again, the incentive is stronger in High School A.

As states have implemented NCLB, criticism of these status-based models has increased, along with calls for a shift to accountability models. This has led to widespread interest in the implementation of "growth models" that value *progress or growth* in addition to *absolute performance*, and that measure—and, therefore, provide incentive to improve—the *progress of all students* along the performance spectrum, not only those s*tudents who perform just below the proficient* bar. Advocates for such accountability models see them as mechanisms for measuring and supporting the goal of improving outcomes for all students over time.

As the policy community looks to the possible inclusion of growth models in the reauthorization of NCLB, there are several issues that need to be better understood. This chapter explains the technical underpinnings of growth models, describes the various types of growth models, states challenges inherent to measuring "growth" at the high school level, and explores implications for policymakers interested in moving toward the widespread use of growth models.

Accountability Models Reflect Expectations

The shift from status-based models to growth-based models would include a significant shift in the balance of expectations within school systems. The expectations implicit in both status- and growth-based models are described below.

Expectations implicit in status-based models

In accountability models based purely on status (such as those currently used by most states to comply with NCLB), the following expectations are implicit:

- All students will be expected to achieve the *same minimum level of achievement* at the same moment in time, regardless of previous level of academic achievement or socioeconomic factors.

- Educators will be held accountable for achieving *different levels of effectiveness* in eliciting student progress depending on the previous level of academic achievement of the children they serve.

Expectations implicit in pure growth-based models

In accountability models based purely on student growth, the following expectations are implicit:

- All educators will be held accountable for the *same level of effectiveness* in eliciting student progress or growth, regardless of the previous level of academic achievement of the children they serve.

- Students will be expected to achieve the *same minimum amount of progress*, regardless of the previous level of academic achievement or socioeconomic factors. In a pure growth model, expecting the same amount of growth from every student regardless of incoming achievement assumes no expectation for closing achievement gaps and no common expectation of ultimate achievement.

These pure models represent the extremes on the accountability spectrum. They also represent the tension between conflicting policy goals: expecting common achievement for all students versus setting common expectations for educators' relative performance. The challenge for policymakers is to implement accountability models that balance the tensions between these goals.

In 2005, the U.S. Department of Education announced a pilot program to allow some states to use an approved growth model for NCLB accountability purposes that counts students "on trajectory toward proficiency within X years."[1] In establishing the parameters for the pilot, the department defined an appropriate balancing of expectations in this way:

- Educators will be expected to elicit more growth (or learning) in their students whose incoming achievement is below grade level in such a way that on-grade-level competency will be achieved within three years (instead of one).

- Students achieving below grade level will be expected to demonstrate slightly more growth (or learning) than their peers who are achieving at or above grade level.

- Students achieving below grade level will *not* be expected to perform at the same minimum competency level as their peers until three years into the future.

This balance can be further described as delayed but completely common expectations for students, and more similar but not completely common expectations for educators.

One intended major benefit of this prescribed balance is that the achievement or growth of *all* students can count positively toward the accountability determinations of educators, schools, and systems: students already proficient count positively, of course, and any students not yet proficient count positively as long as they are progressing toward proficiency. This is the significant difference from the status models, in which accountability determinations benefit only from extraordinary effort with students already very close to the proficiency target. However, one considerable flaw in this prescribed balance is the unreasonable expectation that educators will effectively move *all* students from below proficient to proficient or above within three years. It can be reasonably argued that students farther from acceptable competency should be given more time to rise to the proficient level, and the amount of time allowed should be based upon an aggressive but commonly observable level of consistent improvement. This approach would make it more plausible that all students can count positively in accountability determinations, by demonstrating that reasonably large numbers of students were able to show the targeted level of growth.

This suggested alternate balance can be described as delayed but eventually common expectations for all students, and similar enough expectations of educators that the expectations are reasonably attainable.

What Are Growth Models?

There are several different types of growth models that can be used to measure growth or progress in student learning, each with different technical requirements.

Types of achievement scales

The foundation for a growth model is the scale on which achievement is measured. The characteristics of the scale define the type of numerical operations that can be performed to calculate growth or progress. There are three important characteristics of scales that have an impact on the types of growth models that are possible to implement: numerical level, span of measurement, and measurement frequency.

Numerical level of achievement scales

Scales on which growth might be measured have typically been described in three broad numerical categories:[2]

1. **Ratio scales**, in which (a) a true zero exists, (b) the difference between numbers equally distant numerically represent comparable differences in value, and (c) rank order is preserved.

   Example: A salary scale, in which (a) $0 indicates no income, (b) the difference between $40,000 and $50,000 represents the same amount of money as the difference between $1,040,000 and $1,050,000, and (c) higher numbers always represent greater salaries.

   Application to growth in student learning: If a ratio scale is used to calculate student growth or progress in achievement, it is possible to measure that, for example, a student has doubled his previous achievement. However, achievement scales do not in practice have meaningful zero points—what would it mean to have zero mathematics or reading achievement? Therefore, it is unreasonable to expect that such inferences could be made legitimately from a growth model of student achievement.

2. **Interval scales**, in which (a) a true zero does *not* exist, (b) the difference between equally distant numbers represents comparable differences in value, and (c) rank order is preserved.

Example: The Fahrenheit temperature scale, in which (a) 0°F does *not* represent absence of temperature, (b) the difference between 40°F and 50°F represents the same amount of additional heat as the difference between 140°F and 150°F, and (c) higher numbers always represent hotter temperatures.

Application to growth in student learning: If an interval scale is used to calculate student growth in achievement, it is possible to measure that a student has progressed twice as much as a peer, because differences can be compared directly. Many achievement test producers and psychometric scholars claim that an interval achievement scale can be produced, and that therefore differences in growth or progress can be directly compared. However, many other test producers and scholars dispute this claim, indicating that this is only true if the psychometric model used to produce the scales is a true mathematical representation of the relationship between student achievement and answers they give to test questions.* Therefore, it may or may not be reasonable to expect that inferences directly comparing the growth or progress of one student to the growth or progress of another could be made legitimately from a growth model of student achievement.

3. **Ordinal scales**, in which (a) a true zero does *not* exist, (b) the difference between equally distant numbers does not represent comparable differences in value, and (c) rank order is preserved.

   Example: Placement in a running event, in which (a) zero does not indicate absence of placement, (b) the difference between rankings 1 and 2 may be minor, but the difference between rankings 3 and 4 may be major, and (c) a higher number always means a longer running time and worse placement.

---

* The psychometric models typically represent the probability that a student of a certain achievement level will answer a test item correctly. It is clearly reasonable to assume that higher-achieving students generally have a higher probability of answering a test item correctly. However, the exact form of the relationship between student achievement and probability of answering a test item correctly is a matter of debate. Current psychometric models attempt to mirror reality by changing the form of the relationship. The model that best conforms to how things actually happen inside students' heads will produce scores with the best interval-level measurement properties. However, it is impossible to know what the actual form of the relationship should be, because it is unobservable. Therefore, psychometric claims to produce an interval-level scale are unprovable and subjective.

Application to growth in student learning: If the increasing community of skeptics is correct that the existence of interval-level scales cannot be verified, then the scales must be treated as ordinal scales to avoid significant skewing of analyses based on the scales. Achievement scales can be reasonably described as the interval scale at a minimum. Because ordinal scales can be used to compare ranks, it is therefore reasonable to expect that inferences comparing ranks can be legitimately drawn from a growth model of student achievement. If ranks can be compared from year to year, then a growth model is possible but is limited in usefulness by the number of ranks. There are typically four ranks in each grade level, representing far below grade level, below grade level, on grade level, and above grade level in assessments used for accountability under NCLB. To expand the usefulness of growth models, a fourth type of scale is needed:

4. **Ordered interval**, which lies between ordinal and interval and thus is called "ordered interval" here, and in which (a) a true zero *does not* exist, (b) the difference between equally distant numbers represents approximately comparable differences in value for numbers close to each other, and (3) rank order is preserved.

   Example: Scale score from an achievement test, in which (a) zero does not mean no achievement, (b) the difference between 100 and 110 is likely approximately equal in value to the difference between 110 and 120, but its comparability to the difference between 310 and 320 is questionable, and (c) larger numbers always represent greater achievement.

   Application to growth in student learning: With an ordered interval scale, it is possible to identify certain points on a scale that represent on-grade-level achievement of third graders (say, 300), fourth graders (say, 400), and so on. With such anchor points on the scale, movement along the scale from 300 to 400 represents movement from minimum expected competency in third grade to minimum expected competency in fourth grade, or the target amount of growth from one year to the next to maintain minimum expected

competency. Any movement to measurably above 400 would indicate "more than one year's growth for one year of instruction," and any movement to measurably below 400 would indicate "less than one year's growth for one year of instruction." However, this only works for students whose proficiency in the first year is at minimum expected competency on the scale.

To expand the usefulness of growth measures, an ordered interval scale can also be used to define additional equivalency anchors. For example, the point of minimum expected competency might be labeled "proficient," but there might also be another point on the scale above which one would not expect students performing at grade level to score. This point might be labeled "advanced" and identified for third graders (say, 350), fourth graders (say, 450), and so on. For example, with these additional anchor points on the scale, movement along the scale from advanced in third grade to advanced in fourth grade represents moving from above grade level one year to an equivalent point the next. In this case, the amount of growth can be defined as the target amount of growth from one year to the next to maintain above-grade-level achievement, or one year of growth for one year of instruction for such students.

Any number of such equivalency anchors can be defined on an ordered interval scale to expand the usefulness of a growth model in measuring the growth of a student starting out at any point to determine whether a student demonstrated more, equal to, or less than one year's growth for one year of instruction. For example, one might use the three anchor points on typical state assessments used for NCLB compliance that define four levels, and subdivide the four ranges of the scale based upon the approximate equivalence of differences (or approximately interval-level measurement) in nearby parts of the scale. Such subdivision provides a larger number of ranks that can then be used to measure student growth more precisely than is possible with the relatively wide rankings used for NCLB. For example, rather than having only proficient (300/400) and advanced (350/450) cut score equivalency points for grades three and four, additional equivalency anchors could be added. In this example, one might define equivalency points for grades three and four that identify when a student has reached a middle portion of the proficiency category, and others that identify when a student has reached the high end of the proficiency category.

Span of achievement scales

A second critical characteristic of achievement scales is the span of the scales. There are two types of scale span: multigrade or vertical scales, and separate grade scales. Where possible, it is preferable to create multigrade scales to avoid complications that arise from determining the progress of students when they are measured on different scales from one year to the next.

1. **Multigrade or vertical scales:** In a multigrade or vertical scale, one must place the skills learned in third grade on the same scale as the skills learned in fourth grade, fifth grade, sixth grade, and so on. In some subjects this seems to be a more reasonable assertion than in others. When the nature of the skills changes substantially from grade to grade, it is difficult to claim in good faith that a single scale has been developed.[3] For example, in science, grade-to-grade differences in high school subject matter (e.g., earth science, biology, physical science, chemistry) are so stark that it is difficult to claim they can all be placed on a single scale, even though scientific reasoning runs through all four types of content.

   In mathematics it might be more reasonable, but one might have to make a break in the scale between the grades where the focus of instruction shifts from basic numeracy to algebraic operations. The introduction of geometry, statistics, trigonometry, and calculus could also be argued to require different scales.

   In reading/language arts it might be even more reasonable, but problems still remain. The focus of instruction for younger children might be decoding, sight word recognition, basic comprehension, spelling, and minimally coherent text production. As students move upward in grades, the focus of instruction may shift to fluency, advanced comprehension, rhetorical and grammatical structures, style, voice, and literary criticism. As similar as these skills may be, it is unclear whether they belong on the same scale for younger children as for older children.

   Therefore, it may or may not be reasonable to expect that inferences from growth models based on multigrade or vertical scales can

result in valid conclusions about school effectiveness. A more valid approach may be to create separate scales for each grade and address the complexities brought about by measuring student achievement on different scales at different occasions.

2. **Separate grade-level scales:** In separate grade level scales (or nonvertical scales), the only way to measure student progress in achievement is to define anchor points (and bands) of equivalence on the separate scales (as described in the section above on ordered interval scales). This is because in using separate scales, there is no claim made that the scales are measuring the same thing—only that there are points on both scales that can be considered equivalent in evaluating whether students have demonstrated enough achievement. Even with modest changes in content across grades (e.g., in high school science, the scientific method as related to physics versus the scientific method as related to biology), it is possible to identify, for example, that in one year the student achieved far above the acceptable level and in the next year not nearly so far above the acceptable level. With significant changes in content across years (e.g., physics content versus biology content), defining equivalency points on separate scales becomes a tenuous exercise and, therefore, so does the measurement of growth.

Measurement frequency

A final important characteristic of achievement scales is the frequency with which student achievement is measured. Current models are mostly based on annual measurement. However, such infrequent testing is not necessary, and, in fact, measuring different skills at each measurement occasion creates some of the problems discussed above.

The farther apart the measurement occasions are, the more likely it is that the skills being measured will change substantially and qualitatively. If measurement frequency were increased from the typical once-yearly administration, the same scale could be used consistently to measure student growth within a school year or course. For example, the same test*

* "Same test" here means a test that measures the same content and has been placed on the same scale, but does not necessarily contain the same test questions or present them in the same order.

could be used at the beginning and the end of each course, or even more frequently. This is particularly applicable to growth-based accountability models in high school where content tends to change significantly from course to course and grade to grade. The more similar the content across measurement occasions is, the easier it is to measure student growth. While increasing the number of measurement occasions for accountability purposes may be a difficult prospect, it would provide the optimal conditions for measuring growth in high school, where content differs significantly not just from grade to grade, but from course to course.

Types of growth models

Of the many different models that have been termed "growth" models, some measure growth and some measure something else entirely. Each of these models—and the associated measurement scale characteristics required—are described below.

Gain score models simply subtract previous achievement scores from current achievement scores to estimate the amount of growth made by individual students. That individual growth is then aggregated (for a school, district, or state) to estimate the amount of growth observed, on average, in the performance of students served by that school, district, or state. Statistical tests are provided that can differentiate the statistical significance of the growth observed in one group of students from that observed in another. Gain score models require measurement on the same scale at each testing occasion and measurement on an interval or ratio level. This provides a powerful growth model if the assumptions are reasonable. However, such models make suspect assumptions about scale characteristics, and the reliability of the outcomes is limited because gain scores are less reliable than the measurement at either occasion.[4]

Regression growth trajectory models estimate a growth trajectory (or growth rate) for each student, based on each student's performance on three or more previous measurements. The models may range from relatively simple regression equations to very complex statistical models.[5] Observed growth rates for individual students are then aggregated for a school, district, or state to estimate the average growth rate observed for students taught in that school or system. Statistical tests are provided that

can differentiate the statistical significance of differences in growth rates from one group of students to another. Regression growth trajectory models require measurement on the same scale at each testing occasion for at least three testing occasions, and require measurement on an interval or ratio scale. Of the different types of growth models, this is the most powerful (if assumptions are met), but makes the strongest demands of measurement scales.

Ordered transition models follow students' rankings from testing occasion to testing occasion (e.g., from proficient last year to advanced this year). The type of transition (which might be classified as some variation on positive, neutral, or negative) made by each individual student is aggregated for an educator, school, or district to describe the types of transition made by students taught in that school or by that educator. These models require measurement data from two or more testing occasions. The models may also range from relatively descriptive models to complex statistical models. Statistical tests may be provided that can differentiate and compare typical transition type across educators, schools, or districts. Ordered transition models do not require either the use of interval-level measurement or measurement on the same scale at each measurement occasion. Ordered transition models require only ordinal or ordered interval measurement.[6]

Prediction deviation models use data about past achievement and/or student and community background characteristics to predict future student achievement and identify the degree to which students underperformed or outperformed their personalized predicted achievement or their personalized predicted growth. These are models that compare expected versus observed achievement and/or growth. In prediction deviation models, the important outcome is how far, on average, a school's or educator's students deviated from what was predicted in terms of either achievement or growth. Statistical tests are provided that can differentiate the degree of positive or negative deviation from expectation from school to school or educator to educator. Prediction deviation models are often called value-added models (or VAMs), because the deviation from prediction is often interpreted as the value added to a student's learning by an individual teacher, school, or district. Prediction deviation models are typically based on any of the previously described types of models, and have the same technical measurement

requirements. A major drawback of this type of model is that any variation in student growth or achievement that cannot be explained by the model is automatically attributed to the school or educator. In other words, the value assigned to individual schools or educators includes the true effects of schools and/or educators lumped together with any sampling, specification, and measurement error in the model.[7]

Prediction deviation models are not growth models, although they are sometimes described as such. They do not qualify as growth models because the outcomes of interest are deviations from expected growth rather than actual student growth.

The "growth" model pilots approved by the U.S. Department of Education are not pure growth models, but **hybrid status/growth models** (also called **on track to proficiency models**). Consistent with the principles laid out by the secretary of education (described above), these models track the progress of not-yet-proficient students toward proficiency within a specified period of time.[8] Such hybrid models may be based on any of the models previously described, where, rather than simply describing the amount of growth students make in a school or class, the result is whether the student is on track to become proficient within the next X years. (X may be three, four, or five, depending upon the model.) Some of the approved growth models are applied to high schools, as summarized later in this chapter.

Types of Inferences Made from Growth Models

There are two basic types of inferences made from growth models: descriptive; and attributive, causal, or value-added. In a descriptive model, the purpose of the interpretation is simply to describe what has been observed for a given school or system. In an attributive, causal, or value-added model, the purpose of the interpretation is to claim that what has been observed for a given school or system can be attributed solely to the school or system.

There are certain requirements for an attributive, causal, or value-added interpretation to be valid (or at least reasonably approximated). Several researchers, including the author of this chapter,[9] identify the minimum requirements to be that

- students are randomly assigned to schools and/or educators;
- any missing data must be missing randomly (e.g., students in each school and demographic group are just as likely to miss measurement occasions);
- the technical measurement scale requirements are met; and
- the model contains all appropriate components to isolate the effects of schools/educators.

In reality, it is almost certain that the first requirement is unmet, because students are not deliberately randomly assigned to schools and it is highly unlikely that students are randomly sorted into schools by circumstance. It is almost certain that the second requirement is unmet, because mobility and absenteeism tend to be associated with certain areas or groups of students. The fulfillment of the third requirement can only be logically (rather than empirically) validated, for the reasons described in the section above on numerical level of achievement scales. Finally, the fulfillment of the fourth requirement can only be logically (rather than empirically) evaluated, because one cannot definitively demonstrate that all important factors have been taken into consideration.

Because of these problems with attributive, causal, or value-added interpretations, it is more valid to interpret the results of growth models in a descriptive manner. In practical terms, what this means for accountability models is that schools or systems whose students demonstrate less growth than desired may be called "schools/systems in need of improvement" not because they are poor schools/systems, but because the students they are serving may be in need of schools/systems of extraordinary effectiveness. This is a fine, *but critical*, line demarcating descriptive from attributive (value-added, causal) interpretations.

Summary of Growth Model Requirements

A summary of the minimum technical measurement characteristics required by the different types of growth models is provided in Table 1, followed by qualitative ratings of the statistical and measurement defensibility and validity of each type of model.

Table 1: Summary of Growth Model Requirements					
	Minimum technical measurement requirements				
Type of growth model	Achievement scale level	Scale span, when measuring		Same scale on all occasions	Defensibility
		Once per year or less	More than once per year		
Gain score	Interval	Multigrade/ vertical	Single grade	Yes	Low to moderate
Regression growth trajectory	Interval	Multigrade/ vertical	Single grade	Yes	Very low to low
Ordered transition	Ordinal	Single grade	Single grade	No	Moderate to high
Prediction deviation (not a true growth model)*	Same as base model	Same as base model	Same as base model	Same as base model	Same as base model
Hybrid status/ growth (on track to proficiency)*	Same as base model	Same as base model	Same as base model	Same as base model	Same as base model

* Both prediction deviation and hybrid status/growth models can be based on gain score, regression growth trajectory, or ordered transition models.

Growth Model Challenges Unique to High School

High school is the endgame

There is a legitimate argument (as well as a legitimate counterargument) that implementing a growth model in high school may not be appropriate. The argument is that when low-achieving students are in high school, they have very little time to rise to an acceptable level of competency before universal public education has been completed—that high school graduation is the endgame of universal public education, and high schools must be held accountable for eliciting minimally acceptable competency by the time their students graduate.

The counterargument is that high school educators should not be held completely accountable for early education quality. They should be required instead to demonstrate the elicitation of extraordinary student growth

in achievement for low-achieving students—but a reasonably observable ceiling should be placed on that expectation. Some critics believe that accountability for student outcomes should rest more heavily on districts, not individual schools, because for high schools with very low-achieving incoming students it is arguably the district (not the high school) that failed to prepare students adequately for high school education.

The nature of high school subject matter

Because subject matter is much more differentiated in high school than in the early grades (e.g., algebra, geometry, trigonometry, and calculus compared to basic numeracy), it is much more difficult to measure growth. One might ask, "When you say growth in science, do you mean growth in biology, chemistry, physics, earth science, or science reasoning?" The specificity of measurement in high school must also be better differentiated than in lower grades to match the differentiation of content expectations in high school. For example, in elementary science, it may be reasonable to measure growth across years, because the emphasized content across multiple years may be based on the concepts of science in general. In high school, however, where very different and very specific discipline-based content (e.g., physics, biology, chemistry) may be measured from year to year and from course to course, that method would not be appropriate.

The frequency of measurement in high school

In most states, as required by NCLB, student achievement is measured only once in high school for each subject. In most states, this means there is at least a two- or three-year lapse in testing, between measurement in grade eight and measurement in grade ten or eleven. Without multiple years of testing in high school, it is difficult to measure growth. This exacerbates the problems of changes in the nature of skills gauged from measurement occasion to measurement occasion.

Current Use of Growth Models at the High School Level

All of the "growth" models described previously are used for accountability by states approved in the Department of Education pilot. Because of the requirement that students must be "on trajectory to proficiency," all of the

Table 2: Summary of Growth Models in Place for NCLB

Measurement characteristics			Type of Model							
			Gain score		Regression growth trajectory		Ordered transition		Prediction deviation (VAM)	
Frequency	Level	Scale type[4]	HS yes	HS no[1]	HS yes	HS no[1]	HS yes	HS no[1]	HS yes	HS no[1]
Once yearly	Ordinal	Single grade						IA[3]		
	Ordered inverval	Single grade					DE MN[2]	MI		
	Interval	Single grade	AK		TX	PA TN			OH[2] Typical VAM	Typical VAM
		Multigrade	FL	AZ, AR, MO, NC	CO				Typical VAM	Typical VAM
More than once yearly	Ordinal	Single grade								
	Ordered interval	Single grade								
	Interval	Single grade								
		Multigrade								

[1] Where states have not applied the growth model to AYP it is because of the lack of adjacent grade-level tests in high schools. Generally, where states apply the growth model to AYP, those states also have adjacent grade-level measurement in high schools. They are included in this chart in part to demonstrate that because of the challenges to measuring growth in high school, many states have opted not to include high school in their growth models.

[2] Ohio and Minnesota apply their growth model to high school AYP even though there is not adjacent grade testing in high schools.

[3] Iowa optionally applies its growth model to high school AYP where schools opt to provide grade-ten tests. In this case, growth is followed from tenth to eleventh grade. Where schools opt not to provide grade-ten tests, those schools are excluded from the growth model.

[4] Some states indicate that they have multigrade and/or interval-level scales, but their growth models do not require such scales. The measurement requirements of the growth models are listed here instead.

department-approved growth models are hybrid growth/status models. The four types of pure growth models—not including prediction deviation—are all represented in the approved growth models, as shown in Table 2 (derived from review of all approved growth model applications, which can be seen at www.ed.gov/admins/lead/account/growthmodel/index.html). Only gain score, regression growth trajectory, and ordered transition models have been

implemented for high school accountability under the growth model pilots. The table identifies the frequency of measurement, the numerical level of measurement scales, and the scale span required by each state's growth model, as well as the type of growth model and whether the state's growth model is applied to high school achievement.

Note that in all cases, the growth models are based upon annual measurement. Note also that only the Ohio growth model resides in the space inhabited by typical value-added models (prediction deviation models). Of the fifteen states with approved growth models, four (Delaware, Iowa, Michigan, and Minnesota) have minimal scale demands: they require only ordinal or ordered interval measurement on scales that do not span multiple grades. Five more states (Alaska, Ohio, Pennsylvania, Tennessee, and Texas) have additional scale demands: they require scales at the interval level, but do not require scales that span multiple grades. Six states (Arizona, Arkansas, Colorado, Florida, Missouri, and North Carolina) go one step further: they require interval-level scales that span multiple grades. Just two of those six states (Colorado and Florida) require a scale that spans grades three through ten rather than just grades three through eight.

There are two other important possible characteristics of a strong growth model for high school. At a minimum, measurement should occur at adjacent grade levels in high school to make the growth model interpretable in terms of individual grades and courses. At a maximum, in order to truly separate out the impact of individual courses on student learning, measurement should occur before and after instruction to measure directly the impact of instruction. This could be at the beginning and end of each course, or as often as at the beginning or end of each unit. Some states have arrived at the minimum by measuring in all high school grades, but none have maximized the usefulness of a high school growth model by implementing pre- and post-instruction measurement. There is one significant caution for a growth model based on pre-/post-measurement—disincentive to encourage students to perform poorly on the pre-test and at maximum competency on the post-test would have to be developed.

Components of an Acceptably Valid High School Growth Model

Measurement components

From a measurement perspective, students should be tested at the very least in adjacent grades so that each individual student's achievement can be tracked from grade to grade. Once-yearly measurement is the minimal measurement requirement needed for a valid high school growth model. If measurement is done any less frequently, it will be impossible to disentangle the growth a student attained in one year's class from the growth that student attained in the next year's class.

For a high school growth model to be useful in making judgments about high school instruction, measurement occasions should occur at the beginning and end of each high school course, to determine how much growth occurred for each student in each class. Because of the specialization in disciplines within a subject (e.g., science, mathematics, language arts) in high school instruction, such a measurement system would allow the growth model to disentangle the growth in mathematics achievement that occurred in, say, an algebra course from growth in mathematics achievement that occurred in a geometry course taken in the same year. This would also enable schools to target professional development efforts in the classrooms where students are making the least progress.

Statistical model

With once-yearly measurement (the minimum measurement requirement), a statistical model would need to take into account qualitative shifts in the types of skills measured in each year by either using a statistical model that does not require interval-level measurement or measuring generic subject matter content rather than subject matter content specific to differentiated disciplines. The disadvantage of the second option is that the information gleaned from the growth model is likely to be less useful in evaluating the impact of specific courses on student growth. The more differentiated the content is from grade to grade, the more difficult it is to validly measure growth from year to year.

With pre- and post-course measurement, a statistical model based on interval-level measurement may be defensible, and would provide strong capacity to differentiate between the student growth or progress observed in one school or class over growth observed in another school or class in which the same course is offered.

Policy Action for Moving Forward with Growth Models for High Schools

The challenges are significant in developing both the necessary assessments and the political will to support a defensible growth model for high schools in every state. Therefore, an aggressive and reasonable target for the large-scale implementation of growth models at the high school level might be eight years. This estimate includes

- approximately two years for the development and passage of federal legislation laying the groundwork for appropriate high school growth models;
- approximately one year for content standards to be developed in accordance with legislation;
- two years for the development of appropriate assessment systems to support appropriately valid measurement of growth in high school;
- two years for the first cohort of students to be measured on the new assessments at least twice, and for pilot growth model analyses to be performed before operational use of high school growth model results for accountability; and
- a final-year application to a second cohort of students, upon whose score data the operational high school growth models would be based.

Requirements for a minimally defensible growth model

Implementing a minimally acceptable and informative growth model will require additional assessments to fill the gap between the last grade of measurement in middle school and the grade level where student achievement is measured in high school that exists in most states. An expansion of testing to these grades would be unpopular, and would require the political will to extend the requirements of NCLB.

Implementing additional tests will necessitate significant additional funding to support the development and implementation of assessments; additional testing-contractor capacity to support development and implementation; and increased capacity of psychometric, statistical, program management, test development, and IT staff for both test contractors and states.

Requirements for a more defensible growth model

Implementing a much more defensible and informative growth model would require either the development of high school course expectations common within each state but different across the country or the development of common high school course expectations across the entire country. There are some states where there are common course expectations, but this is not a universal condition across the fifty states.

The latter is a highly sensitive political issue, as the development of common standards and course expectations is seen by some as a major states' rights issue. However, such a testing system would be much more useful than once-yearly testing in terms of the information that would result from growth models. Such a system would provide disentangled information about student achievement growth occurring in individual courses. In addition, disincentive to encourage students to perform poorly on the pre-test and at maximum competency on the post-test would have to be developed.

The resources needed to implement such new testing programs are similar to those mentioned above, but on a larger scale. Because each course would require a pre- and post-test, the number of tests to be developed in each subject would be twice the number of courses in each subject rather than one test in each grade level where testing is not currently performed. This increase in required resources would even more significantly strain budgets, contractor staff, and state staff.

In order to overcome these obstacles, political will would need to be gathered to pass federal legislation providing additional funding for increased measurement and increased state staffing. Such legislation would also minimally need to require annual testing in every state to fill in the gap between middle school and high school measurement occasions. Finally, in

order to provide a high school growth model that is more than minimally defensible, the legislation would need to require pre- and post-testing in each required high school course, as well as standardization of required high school courses at least within each state, and possibly across the entire nation.

Conclusion

Putting a valid growth model in place for high schools is a worthwhile and important endeavor, in part because measuring student growth is closer to the educational mandate to facilitate student learning than simply measuring students' level of achievement. While growth models do not resolve the tension between setting common expectations for educators and setting common expectations for students, they are capable of balancing that tension in such a way that achievement gaps can be closed without needing educators to perform at unobservable and unreasonable levels.

The challenges to implementing a minimally valid growth model for high schools are significant, and the challenges to implementing an optimally valid growth model for high schools are even more so. In spite of this, however, the emphasis on student learning implicit in growth models, and the usefulness of the information available from growth models for policymakers, educators, and stakeholders, is of sufficient value that the challenges should be taken on.

The views expressed in this chapter are those of the author and do not necessarily represent those of the Alliance for Excellent Education.

About the Author

Joseph Martineau is currently the director of the Office of Educational Assessment and Accountability in the Michigan Department of Education. He received a bachelor's degree in linguistics and master's degree in instructional psychology and technology from Brigham Young University, and a PhD in measurement and quantitative methods in education from Michigan State University. Before assuming his current position, his career included positions as an instructional designer, educational programmer, university instructor, research consultant, psychometrician for the state of Michigan,

and manager of Michigan's K–12 general education assessments. Most importantly, Dr. Martineau and his wife have school-age children directly affected by his work in assessment and accountability.

---

[1] M. Spellings, "Letter to Chief State School Officers Regarding the Opportunity to Participate in a Growth Model Pilot," http://www.ed.gov/policy/gen/guid/secletter/080818.html (accessed on February 21, 2009).

[2] S. S. Stevens, "Mathematics, Measurement and Psychophysics," in *Handbook of Experimental Psychology*, ed. S. S. Stevens, 1–49 (New York: Wiley, 1951).

[3] J. A. Martineau, "The Effects of Construct Shift on Growth and Accountability Models," unpub. diss., Michigan State University, East Lansing, 2004; J. A. Martineau, "Distorting Value Added: The Use of Longitudinal, Vertically Scaled Student Achievement Data for Growth-Based Value-Added Accountability," *Journal of Educational and Behavioral Statistics* 31, no. 1 (2006): 35–62; M. D. Reckase, "Controlling the Psychometric Snake: Or, How I Learned to Love Multidimensionality," paper presented at the Annual Meeting of the American Psychological Association, August 1989, New Orleans, LA; M. D. Reckase, "Real World Is More Complicated Than We Would Like," *Journal of Educational and Behavioral Statistics* 29, no. 1 (2004): 117–20; W. H. Schmidt, R. T. Houang, and C. C. McKnight, "Value-Added Research: Right Idea but Wrong Solution?" in *Value Added Models in Education: Theory and Applications*, ed. R. Lissitz, 145–64 (Maple Grove, MN: JAM Press, 2005).

[4] J. B. Willett, "Some Results on Reliability for the Longitudinal Measurement of Change: Implications for the Design of Studies of Individual Growth," *Educational and Psychological Measurement* 49, no. 3 (1989): 587–602.

[5] For comparisons, see D. F. McCaffrey, J. R. Lockwood, D. M. Koretz, T. A. Louis, and L. S. Hamilton, "Models for Value-Added Modeling of Teacher Effects," *Journal of Educational and Behavioral Statistics* 19, no. 1 (2004): 67–101.

[6] For examples, see D. W. Betebenner, "Performance Standards in Measures of Educational Effectiveness," paper presented at the 25th Annual Conference on Large-Scale Assessment of the Council of Chief State School Officers, June 2005, San Antonio, TX; R. Hill, "Measuring Student Growth Through Value Tables," paper presented at the 25th Annual Conference on Large-Scale Assessment of the Council of Chief State School Officers, June 2005, San Antonio, TX; J. A. Martineau and D. W. Betebenner, "A Hybrid Value Table/Transition Table Model for Measuring Student Progress," paper presented at the 26th Annual National Conference on Large-Scale Assessment of the Council of Chief State School Officers, 2006, San Francisco, CA.

[7] D. F. McCaffrey, J. R. Lockwood, L. T. Mariano, and C. Setodji, "Challenges for Value-Added Assessment of Teacher Effects," in Lissitz, ed., *Value Added Models in Education*, 111–41; S. L. Rigney and J. A. Martineau, "NCLB and Growth Models: In Conflict or in Concert?" in ibid., 47–81.

[8] Spellings, "Letter to Chief State School Officers."

[9] C. Glymour, "A Review of Recent Work on the Foundations of Cause Inference," in *Causality in Crisis? Statistical Methods and the Search for Causal Knowledge in the Social Sciences*, ed. V. McKim and S. Turner (Notre Dame: University of Notre Dame Press, 2007); McCaffrey et al., "Challenges for Value-Added Assessment"; J. A. Martineau, "Distorting Value Added."

CHAPTER 7

Assessing High School English Language Learners

Jamal Abedi
University of California at Davis

As the number of English language learners (ELLs) in the American school system grows, issues regarding their education need to be given more attention. ELL enrollment has grown 57 percent since 1995, while the rate for all students has been at less than 4 percent. Currently, there are 5.1 million ELL students, forming more than 10 percent of the country's student population.[1] Because of the rapid growth of this group, we need to accurately determine which ELL students require English language services, and then work to support all of their academic needs.

English language learners often differ with respect to their sociocultural background, parents' level of education, ethnic background, family characteristics, and level of native and English language fluency.[2] But they all share a common need for help with English language proficiency.

The federally mandated inclusion of ELLs in state assessment systems necessitates an examination of how assessments of both English language

proficiency and content knowledge affect these students' academic lives. It is imperative to identify and examine factors that affect the academic performance of ELL students in order to provide ways to effectively deal with the assessment and instructional issues facing this population.

While the issues concerning instruction and assessment for ELL students are ultimately inseparable, this chapter focuses on the assessment issues that arise for these students and offers federal policy recommendations that will help ensure successful academic careers for all ELLs.

The Role of Assessment in ELL Students' Academic Career

Traditionally, students are taught and then tested to assess what they have learned. For ELL students, however, assessment comes before instruction begins. This is due to the fact that ELL students' level of English language proficiency (ELP) must first be evaluated so ELLs can be properly placed into appropriate instructional settings when an English-only instructional environment is the preferred choice. ELL students who are in English-only instructional classes and required to take assessments in English without having mastered the necessary level of English proficiency are at risk of failure.

For ELL students, performance on ELP assessments is the main criterion for classification into the ELL category and for reclassification from ELL to RFEP (reclassified fluent English proficient). Improper classification may lead to inappropriate and inadequate instruction and may also affect accountability, such as in the reporting of Adequate Yearly Progress[3] for ELLs. Misleading results of invalid ELP assessment and inaccurate classification may lead to the disproportionate placement of ELL students in special education classrooms, which can negatively affect both their academic career as a whole and the time it takes them to graduate.[4] Furthermore, ELLs, like their peers, are subject to content-area assessments that have student-level implications, including grades, promotion, and graduation, as well as system-level implications, including AYP determinations. Thus, it is clear that assessment outcomes profoundly impact ELLs' academic performance. Proper attention must therefore be given to the ways that tests are developed, field tested, and reported for these students.

Assessment issues for ELLs are more complicated and important at the high school level than in the previous grades. Not only are assessment materials heavily impacted by linguistic factors because of the more complex language used at this level, but high-stakes decisions about students' academic performance are also made more frequently during these years. As research on the classification of ELL students reveals, there is increased pressure on schools to reclassify ELL students out of the ELL category at higher grade levels.[5] If students are reclassified prematurely, they may not have enough language proficiency to meaningfully participate in content-area assessments. Many states, for instance, require high school students to take the state high school exit examination in order to graduate. As with many other state assessments, these tests often suffer from cultural and linguistic biases. ELL students may be unable to pass such exams, not because of a lack of content knowledge, but because of a lack of a thorough understanding of the exit examination language.

Assessment of English Language Proficiency

To make sure that ELL students are ready to take state content-area assessments in English, their level of academic English proficiency must be gauged. If they are not at a level where they can meaningfully participate in assessments conducted in English, their content-area assessment outcomes may not be valid. ELP assessments include cut scores for determining the level of English proficiency. While there are some differences between the reporting policies of ELP scores among states, Level 1 usually refers to no or very low proficiency in English and Level 5 represents high proficiency. ELL students are typically reclassified from LEP to fluent English proficient (FEP) at ELP Level 4 or above.[6]

Issues regarding ELP assessment, including those that were developed and used prior to the implementation of the No Child Left Behind Act of 2001 (NCLB) and those developed in accordance with federal regulations accompanying NCLB, are explored below.

Status of ELP assessments prior to NCLB

There were many English language proficiency assessments available for public use prior to the implementation of NCLB. These assessments

were based on one or more of at least three different schools of thought, and thus provided differing measures of proficiency.[7] Reviews of the pre-NCLB ELP assessments found major variations in the content, structure, test administration procedures, theoretical bases, and issues related to the validity and reliability of the tests. Researchers found that the assessments also differed in their approaches to defining language proficiency, the types of tasks and specific item content, the grade-level ranges, and the specific time limits.[8]

Similarly, a review of the content and psychometric characteristics of some of the most commonly used English language proficiency tests prior to NCLB found major differences between these tests with respect to their purpose, age and language group, administration, cost, items, scoring, test design, theoretical foundation, and reliability and validity.[9] Such discrepancies are cause for concern, as their outcomes may not be comparable and may negatively impact the authenticity of the English language proficiency assessments.[10]

Impact of NCLB on ELP assessments

Title III of NCLB requires states receiving Title I funding to annually assess ELL students' level of English language proficiency using reliable and valid measures in four areas: reading, writing, listening, and speaking. These assessments must be aligned with the state's ELP content standards and should measure academic English.

At least four consortia of states developed four batteries of ELP assessments based on the NCLB Title III guidelines.[11] These assessments included all four modalities required by NCLB Title III and were aligned with the states' ELP content standards. They were developed for four or more grade clusters (typically K–2, 3–5, 6–8, and 9–12) and included common sets of items across adjacent grade clusters. The test developers conducted extensive pilot and field testing on large and representative samples of students. The total tests and the content and psychometric properties of the individual items were then carefully examined, and changes were made where needed.

There has clearly been improvement in the content and psychometric properties of the post-NCLB English language proficiency assessments as

compared to those created prior to NCLB implementation. However, there are still significant problems that need to be resolved before the assessments can be safely used, including the following:

- the lack of a commonly acceptable definition of English language proficiency;
- issues concerning standard settings for the newly developed ELP assessments;
- issues concerning dimensionality of scores obtained from the four modalities and reporting these scores;
- comparability of the new assessments with the pre-NCLB assessments in establishing the baseline; and
- the lack of an objective definition of the concept of academic English.[12]

Assessment of Content-Area Knowledge

NCLB requires English language learners to be assessed in the content areas in which all other students are required to be tested: reading, language arts, math, and science. Results from these standardized academic achievement tests are then used in high-stakes assessment and accountability decisions for high school students, such as classification/reclassification,[13] promotion, and graduation.

There are a number of challenges related to assessing ELLs in the content areas. Many critics believe that the tests used for these purposes are not appropriately designed for such use with ELLs,[14] and feel that there should be standardized achievement tests specifically designed to assess these students' content knowledge.[15]

Requiring English proficiency for participation in content-area assessments

Many state education officials and ELL assessment experts believe that ELL students should take content-area assessments in English only when they are proven proficient enough in English. In many states, students are considered proficient in English if they score at proficiency Level 4 or higher, but this is not consistent across the country.[16] Defining an appropriate level

of proficiency is of paramount importance for high school students, particularly in states that administer high school exit examinations. Since these exams are mostly presented in English, ELL students with a lower level of English proficiency may not be able to fully demonstrate their content knowledge.

Impact of linguistic factors on the assessment of ELL students

The main issue with many of the achievement test items that are developed for native speakers of English is that there may be cultural and linguistic biases that affect the validity and reliability of the assessments. Research on the assessment of ELL students clearly indicates that unnecessary linguistic complexity of test items is a source of measurement error, and that construct-irrelevant variance may threaten the validity of standardized achievement tests for ELLs.[17] Researchers have found that linguistically complex items largely contribute to the measurement error for ELL students, which may cause lower reliability and result in the misinterpretation and misunderstanding of test questions.[18]

Research has also demonstrated that unnecessary linguistic complexity of test items contributes to the performance gap between ELL and non-ELL students.[19] The higher the level of linguistic complexity, the larger the performance gap between ELL and non-ELL students. In this way, language factors play an important role in the assessment of ELL students, particularly at higher grade levels.

There is a substantial performance gap between ELLs and their native English speaker peers in all content areas, and this gap widens as the level of language demand in assessments increases.[20] There are many linguistic features that make the comprehension of assessment materials difficult for English language learners, including unfamiliar vocabulary, complicated grammatical structures, and styles of discourse that include extra material, abstractions, and passive voice.[21] Reports of studies on the impact of language factors on the assessment of ELLs have included a comprehensive review of linguistic features that affect performance outcomes of ELL students, along with citations to relevant research.[22]

To make content-based assessments more accessible to ELL students, the concept of a linguistic modification approach has been suggested.[23] The main theme underlying this method is to reduce or eliminate unnecessary linguistic complexities in order to make assessments more reliable, more valid, and more accessible for ELLs. Under this approach, the linguistic features that make assessments more complex are identified and then revised. For example, ELL students have been shown to have difficulty with unfamiliar vocabulary, passive voice, conditional clauses, long and complex phrases, relative clauses, and long nominals. In the linguistic modification, unfamiliar or infrequent words are changed to familiar words, passive verbs are changed to active verbs, conditional clauses are replaced with separate sentences or the order of conditional and main clauses are changed, complex question phrases are changed to simple question words, relative clauses are either removed or recast, and long nominals are shortened. Below is an example of an original test item and a proposed linguistically modified version of that item. As can be seen from this example, multiple sources of linguistic complexities were involved, and multiple modifications were performed.

Original test item:
The census showed that three hundred fifty-six thousand, ninety-seven people lived in Middletown. Written as a number, that is
A. 350,697
B. 356,097
C. 356,907
D. 356,970

Modified test item:
Janet's video game score was three hundred fifty six thousand, ninety-seven. Written as a number that is
A. 350,697
B. 356,097
C. 356,907
D. 356,970

The main issue in the implementation of the linguistic modification approach is how to decide which linguistic features are necessary and

relevant to the construct being measured and which are unnecessary and irrelevant. Researchers highly recommend that a team of specialists, including math content experts, linguists, and test item developers, decide what language in the test items is considered unnecessary and needs to be modified.[24]

The use of accommodation in the assessment of ELL students

It is recommended that accommodations be used to provide fair content-area assessment for ELL students. Literature on the accommodations for ELL students shows that many different types of accommodations are already provided to ELL students, but some of them may not be relevant or effective. For example, some of these accommodations may alter the construct being measured; therefore their validity might be questionable, particularly for high school students.[25]

In an analysis of seventy-three accommodations used for ELL students across the nation,[26] it was shown that only eleven of them (15 percent) were deemed appropriate for ELL students. Below are examples of accommodations that are used for ELL students that may not be as relevant:

- enlarged answer sheets;
- multiple breaks throughout the testing period;
- the administration of tests to individual students;
- the administration of tests in small groups; and
- the administration of tests in locations with minimal distraction.

Researchers studying accommodations used in administering the National Assessment for Educational Progress found that the use of accommodations such as one-on-one testing, small-group testing, extended time, and oral reading of directions did not help to improve the performance of ELL students or reduce the performance gap between ELLs and native English speakers in content-based assessments.[27]

To reduce the performance gap, ELLs need help with the language of assessment and instruction. As indicated earlier, assessment and instructional materials that are developed mainly for native speakers of English may contain linguistic structures that are too complex for non-

native speakers, so accommodations that help ELLs cope with language issues would be the most effective way to increase their assessment scores. But none of the accommodations mentioned above directly address ELL language needs. Research suggests that providing language-based accommodations such as customized dictionaries or glossaries or a linguistically modified version of the assessment helps ELL students present a more valid picture of what they know and can do.[28]

Methodological Issues Related to the Assessment of ELL Students

The fundamental principle underlying any assessment for all students is the validity of the assessments. If tests are not valid, then their outcomes can be misleading and can negatively impact students' academic life. There are many different ways to determine assessment validity for students, including content, criterion, consequential, and construct validity approaches.[29] Below is a brief discussion of the concerns with regard to each of those approaches.

Content validity: For an assessment to have content validity, it must include a representative sample of the universe of all possible test questions based on the relevant content standards.[30] Assessments may not be valid for ELL students in terms of content, since they may include language content that is unrelated to the focal content being measured.

Criterion-related validity: To be valid in terms of a criterion-related approach, the assessment outcome must be highly correlated with a valid criterion or criteria.[31] Examining the criterion-related validity for ELL students may be technically unfit, since it is difficult to find measures that are free of linguistic and cultural biases.

Consequential validity: Assessments can be consequentially valid if the consequences of a particular use or interpretation of assessment results are important in arriving at an overall evaluative judgment of the validity of the assessment for that use or interpretation.[32] Content-based assessments that are developed for native English speakers are particularly problematic for ELL students in terms of consequential validity. These assessments are used for many different purposes in the ELLs' academic career, and some of these—such as those used for classification,[33] accountability,[34] and

graduation—may not be relevant.[35] Due to technical issues involved in these assessments for ELLs, the outcomes may be misleading.

Construct validity: Construct validation is the most fundamental and important approach in the validity of assessments, particularly with regard to academic assessments. Assessments have construct validity if they measure the intended construct and nothing else. Literature on the assessment of ELL students demonstrates that unnecessary linguistic complexity of content-based assessment is a source of construct-irrelevant variance and negatively impacts the validity of ELL assessments.[36] Language factors in content-based assessments add a new dimension and reduce internal consistency between test items within a test.

Conclusion

Assessment plays a vital role in the academic career of ELL high school students. It shapes their classification evaluations, is used for accountability purposes, and helps educators make decisions about promotion and graduation. It therefore constitutes the foundation of ELL students' educational career. Even a minor problem in test development, field testing, and scoring could cause serious consequences in ELL students' academic lives.

As the content and linguistic structure of assessments become more complex at the high school level, so do the choice and effectiveness of ELL accommodation strategies. Thus, a sound decision on which accommodations should be used in the assessment of ELL students requires a correspondingly complex set of criteria. Unfortunately, there are very few research-supported accommodations available for high school students, and even those that are supported by research are not frequently used by schools.[37] It is therefore imperative to carefully examine accommodation needs for these students and to provide accommodations that properly address those needs.

Federal Policy Recommendations

Federal legislation such as the No Child Left Behind Act and its predecessor, the Improving America's Schools Act of 1994, addresses the need to advance

the quality of teaching and learning for *every* child, including English language learners. These laws mandate inclusion of ELL students in large-scale state and national assessments. However, mandating inclusion of these students alone may not produce desirable outcome unless attention is paid to the quality of ELL instruction and assessments. The first and most important step in providing reliable and valid assessments for ELL students is to understand the complex nature of the ELL assessment system and to identify variables that impact such assessments. There are many issues that have a profound interactive impact on the assessment of this particular subgroup of students. These issues affect ELL classification, accountability, promotion, and graduation. Federal policymakers should consider the following recommendations:

1. Encourage state assessment divisions and test publishers to clearly examine factors that hinder the accessibility of assessments for ELL students. Examples of these factors include unnecessary linguistic complexity and cultural biases. A linguistic and cultural bias review process should be included in the test development process to address these issues.

2. Ensure that tests used for high-stakes purposes, including classification, promotion, and graduation, are carefully reviewed for any validity and accessibility concerns.

3. Support the use of multiple criteria for high-stakes decisions; a single criterion may not provide the preponderance of evidence that is needed for such decisions.

4. Support states and test publishers to provide assessments that are sensitive to ELL students' language and cultural needs. Only such assessments should be used for high-stakes decisionmaking.

5. Include formative assessment in the curriculum for all students, particularly for ELLs. The outcome of formative assessment would be extremely helpful in improving the academic performance of ELL students.

6. Support states in providing testing accommodations that are relevant for these students and are effective and valid in making assessments fair and accessible, while recognizing that many current accommodations may not be relevant for these students, or may even invalidate the assessment outcomes.

7. Encourage research on the effectiveness and validity of accommodations currently used by states.

The views expressed in this chapter are those of the author and do not necessarily represent those of the Alliance for Excellent Education.

About the Author

Jamal Abedi is a professor at the School of Education of the University of California at Davis and a research partner at the National Center for Research on Evaluation, Standards, and Student Testing (CRESST). His research interests include psychometrics and test and scale developments. Among his recent work are studies on the validity of assessments and accommodations and the opportunity to learn for English language learners (ELLs). He is the recipient of the 2003 Professional Service Award in recognition of "Outstanding Contribution Relating Research to Practice," from the American Educational Research Association, and the 2008 Lifetime Achievement Award, from the California Educational Research Association. Dr. Abedi's educational background is in psychometrics and research methodology. He holds a master's degree and a PhD in psychometrics from Vanderbilt University.

---

[1] M. E. Flannery, "A New Look at America's English Language Learners," *NEA Today*, http://www.nea.org/home/29160.htm (accessed January 5, 2009).

[2] J. Abedi and P. Gandara, "Performance of English Language Learners as a Subgroup in Large-Scale Assessment: Interaction of Research and Policy," *Educational Measurement: Issues and Practice* 25, no. 4 (2006): 36–46.

[3] No Child Left Behind Act of 2001, HR 1, 107th Cong., 1st sess.

[4] J. Abedi, "Classification System for English Language Learners: Issues and Recommendations," *Educational Measurement: Issues and Practices* 27, no. 3 (2008): 17–22.

[5] Abedi, "Classification System for English Language Learners."

[6] J. Abedi, ed., *English Language Proficiency Assessment in the Nation: Current Status and Future Practice* (Davis, CA: University of California Press, 2007).

[7] G. Valdes and R. A. Figueroa, *Bilingual and Testing: A Special Case of Bias* (Norwood, NJ: Ablex, 1994).

[8] A. M. Zehler et al., *An Examination of Assessment of Limited English Proficient Students* (Arlington, VA: Development Associates, Special Issues Analysis Center, 1994).

[9] A. Del Vecchio and M. Guerrero, *Handbook of English Language Proficiency Tests* (Washington, DC: National Clearinghouse for Bilingual Education, 1995).

[10] For a more detailed description of this issue, see J. Abedi, "Measuring Students' Level of English Proficiency: Educational Significance and Assessment Requirements," *Educational Assessment* 13 (2008): 193–214.

[11] See Abedi, *English Language Proficiency Assessment in the Nation.*

[12] For a more detailed discussion of issues concerning the newly developed ELP assessments, see Abedi, "Measuring Students' Level of English Proficiency."

[13] A. L. Kindler, *Survey of the States' Limited English Proficient Students and Available Educational Programs and Services, 2000–2001 Summary Report* (Washington, DC: National Clearinghouse for English Language Acquisition and Language Instruction Educational Programs, 2002).

[14] K. S. Mahoney and J. MacSwan, "Reexamining Identification and Reclassification of English Language Learners: A Critical Discussion of Select State Practices," *Bilingual Research Journal* 29 (2005): 31–42.

[15] E. H. Stefanakis, *Whose Judgment Counts?: Assessing Bilingual Children, K–3* (Portsmouth, NH: Heinemann, 1998).

[16] Abedi, "Measuring Students' Level of English Proficiency."

[17] J. Abedi, "Language Issues in Item-Development," in *Handbook of Test Development*, ed. S. Downing and T. Haladyna, 377–98 (Mahwah, NJ: Lawrence Erlbaum Associates, 2006); R. A. Figueroa, "Psychological Testing of Linguistic-Minority Students: Knowledge Gaps and Regulations," *Exceptional Children* 56, no. 2 (1989): 145–53; R. A. Figueroa, "Best Practices in the Assessment of Bilingual Children," in *Best Practices in School Psychology*, ed. A. Thomas and J. Grimes, 93–106 (Washington, DC: National Association of School Psychologists, 1990); T. M. Haladyna and S. M. Downing, "Construct-Irrelevant Variance in High-Stakes Testing," *Educational Measurement: Issues and Practice* 23, no. 1 (2004): 17–27; S. Messick, "The Interplay of Evidence and Consequences in the Validation of Performance Assessments," *Educational Researcher* 23, no. 2 (2004): 13–23; G. Solano-Flores and E. Trumbull, "Examining Language in Context: The Need for New Research and Practice Paradigms in the Testing of English-Language Learners," *Educational Researcher* 32, no. 2 (2003): 3–13.

[18] Solano-Flores and Trumbull, "Examining Language in Context."

[19] Abedi, "Language Issues in Item-Development"; J. Abedi and C. Lord, "The Language Factor in Mathematics Tests," *Applied Measurement in Education* 14, no. 3 (2001): 219–34; Solano-Flores and Trumbull, "Examining Language in Context"; Valdes and Figueroa, *Bilingual and Testing*.

[20] Abedi, "Language Issues in Item-Development"; Solano-Flores and Trumbull, "Examining Language in Context."

[21] Abedi, "Language Issues in Item-Development"; J. Abedi, C. Lord, and J. Plummer, *Language Background as a Variable in NAEP Mathematics Performance* (Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing, 1997).

[22] Readers who are interested in a more detailed discussion of linguistic features are referred to Abedi, "Language Issues in Item-Development," and Abedi, Lord, and Plummer, *Language Background*.

[23] See, for example, Abedi, "Language Issues in Item-Development"; Abedi, Lord, and Plummer, *Language Background*.

[24] Abedi, Lord, and Plummer, *Language Background*.

[25] J. Abedi, C. Hofstetter, and C. Lord, "Assessment Accommodations for English Language Learners: Implications for Policy-Based Empirical Research," *Review of Educational Research* 74, no. 1 (2004): 1–28; C. Rivera et al., "Study 1: An Analysis of State Assessment Policies Regarding the Accommodation of English Language Learners," in *State Assessment Policy and Practices for English Language Learners*, ed. C. Rivera and E. Collum, 1–174 (Mahwah, NJ: Lawrence Erlbaum Associates, 2006); S. G. Sireci, S. Li, and S. Scarpati, *The Effects of Test Accommodation on Test Performance: A Review of the Literature* (Amherst, MA: School of Education, University of Massachusetts, Amherst, 2003).

[26] C. Rivera, "State Assessment Policies for English Language Learners," presented at the Council of Chief State School Officers' 33rd Annual Conference on Large-Scale Assessment, June 2003, San Antonio, TX.

[27] J. Abedi and F. Hejri, "Accommodations for Students with Limited English Proficiency in the National Assessment of Educational Progress," *Applied Measurement in Education* 17, no. 4 (2004): 371–92.

[28] Abedi, Hofstetter, and Lord, "Assessment Accommodations for English Language Learners"; V. L. Kiplinger, C. A. Haug, and J. Abedi, "Measuring Math—Not Reading—on a Math Assessment: A Language Accommodations Study of English Language Learners and Other Special Populations," paper presented at the annual meeting of the American Educational Research Association, April 2000, New Orleans, LA; N. A. Maihoff, "Using Delaware Data in Making Decisions Regarding the Education of LEP Students," paper presented at the Council of Chief State School Officers 32nd Annual National Conference on Large-Scale Assessment, June 2002, Palm Desert, CA; Sireci, Li, and Scarpati, *The Effects of Test Accommodation on Test Performance.*

[29] M. J. Allen and W. M. Yen, *Introduction to Measurement Theory* (Monterey, CA: Brooks/Cole, 1979); R. L. Linn and N. E. Gronlund, *Measurement and Assessment in Teaching*, 7th ed. (Saddle River, NJ: Merrill, 1995); R. M. Thorndike, *Measurement and Evaluation in Psychology and Education* (Saddle River, NJ: Merrill, 2005).

[30] Thorndike, *Measurement and Evaluation in Psychology and Education*.

[31] Ibid.

[32] Linn and Gronlund, *Measurement and Assessment in Teaching*.

[33] Abedi, "Classification System for English Language Learners."

[34] J. Abedi, "The No Child Left Behind Act and English Language Learners: Assessment and Accountability Issues," *Educational Researcher* 33, no. 1 (2004): 4–14.

[35] S. J. Cech, "Graduation Hurdles Prove High for ELLs," *Education Week* 28, no. 17 (2009); P. A. Garcia and M. Gopel, "The Relationship to Achievement on the California High School Exit Exam for Language Minority Students," *Journal of Research and Practice* 1, no. 1 (2003): 126–40; J. Wang, D. Niemi, and H. Wang, *Predictive Validity of an English Language Arts Performance Assessment* (Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing, 2007).

[36] Abedi, "Language Issues in Item-Development"; Haladyna and Downing, "Construct-Irrelevant Variance in High-Stakes Testing"; S. Messick, "The Interplay of Evidence and Consequences in the Validation of Performance Assessments," *Educational Researcher* 23, no. 2 (2004): 13–23.

[37] Abedi, Hofstetter, and Lord, "Assessment Accommodations for English Language Learners"; Rivera et al., "Study 1"; Sireci, Li, and Scarpati, *The Effects of Test Accommodation on Test Performance*; Thorndike, *Measurement and Evaluation in Psychology and Education.*

CHAPTER 8

Students with Disabilities: Expecta… Academic Achievement, and the Criti… Role of Inclusive Standards-Based Assessments in Improving Outcomes

Rachel Quenemoen
National Center on Educational Outcomes, University of Minnesota

Prologue: Story from San Diego—The Dilemma

Excerpt from E. Alpert, "Deterred from Diplomas for Better or Worse," Voiceofsandiego.org, October 2, 2008:

> Just after starting high school, Lance Rogers was told he wouldn't earn an ordinary diploma. He struggled with attention deficit hyperactivity disorder and other disabilities, and had trouble focusing in big classes at Point Loma High School.
>
> Instead he took special education classes that were smaller and easier, but wouldn't help him earn a degree. His mother Ruth Rogers hoped he would flourish there, even if he was "non-diploma bound." It is a label given to thousands of San Diego Unified students with disabilities who focus on skills that will help them live independently instead of prepping for college or beyond, studying shopping lists and sales tax instead of calculus or Cervantes.
> But Lance Rogers grew depressed and bored in those classes. He

> can't remember what he learned—only that he was often asked to draw pictures or maps—and ultimately ditched school.
>
> "I was downhearted," said Lance Rogers, now 16 years old. "I didn't do my work, because what was the point of doing it? I didn't get any credit. So I didn't go to school."
>
> Yet when the Rogers family moved to Texas, their son thrived in a school with a mixture of small classes and counseling. His grades rose from Ds to Bs. And when the family returned to San Diego, teachers at another school said Lance Rogers was perfectly capable of earning a diploma.
>
> "I was blown away," Ruth Rogers said. "I was shocked that he was in the classroom, doing what he's supposed to be doing." …
>
> No educator means to shortchange children with disabilities, but an overburdened and underfunded system causes mistakes when diagnosing and placing children in classes, said parent Joyce Clark, chairwoman of a San Diego Unified committee on special education. Clark said some children are funneled into easier classes instead of making ordinary classes accessible through technology or other aids.
>
> "Teachers are wonderful but they get weary of trying to address all the needs they are asked to do," Clark said. "And somehow [some students] just fall through the cracks."

This story illustrates a dilemma that creates false choices for students, parents, and teachers. This dilemma stems from common misconceptions held by the public, policymakers, school leaders, and even teachers of how specific learning needs related to identified disabilities affect a student's ability to learn and to earn a regular diploma. It results in persistently low expectations for the achievement of students with disabilities, unwillingness of schools to be held accountable for their progress (or lack of it), and low levels of achievement and postschool success for many of these students. These erroneous but pervasive misconceptions of high school students with

disabilities are deeply affected by policies related to standards, assessments, and accountability systems, as described in this chapter. To improve expectations, teaching, learning, and outcomes—for all students, including students with disabilities—it is critical that policies at all levels help leverage implementation of an inclusive assessment system that supports these goals.

This chapter describes issues concerning assessing students in a standards-based accountability system and related federal policies. It also describes ways to evaluate assessments that are inclusive of all students in the accountability system. It concludes with recommendations for policymakers.

Misconceptions and Low Expectations

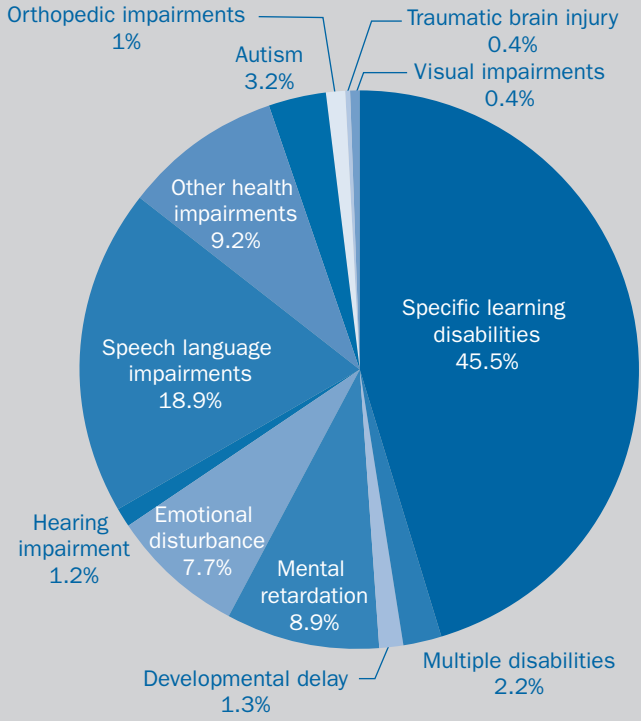
The learning characteristics of students with disabilities vary greatly. Understanding who these students are, and how their disabilities may affect their learning, is foundational to understanding how assessments can yield data to hold schools accountable for the results. See Figure 1 for a summary of categorical distribution of students with disabilities.

The impact of the spectrum of disabilities on students' ability to learn is described by disabilities expert Martha Thurlow, director of the National Center on Educational Outcomes (NCEO), as follows:

> Most students with disabilities (75 percent altogether) have learning disabilities, speech/language impairments, and emotional/behavioral disabilities. These students, along with those who have physical, visual, hearing, and other health impairments (another 4–5 percent), are all students *without intellectual impairments*. When given appropriate accommodations, services, supports, and specialized instruction, these students (totaling over 80 percent of students with disabilities) can learn the grade-level content in the general education curriculum, and thus achieve proficiency on the grade-level content standards. In addition, research suggests that many of the *small* percent of students with disabilities *who have intellectual impairments* (i.e., generally includes students in categories of mental retardation, developmental delay, some with multiple disabilities, some with autism), totaling less than 2 percent of the entirety of the student population, *or less than 20*

percent of all students with disabilities, can *also* achieve proficiency when they receive high quality instruction in the grade-level content, appropriate services and supports, and appropriate accommodations.

Figure 1: Categorical Distribution of Students with Disabilities



Note: Percentages in this figure are based on a total number of 6.5 million students receiving special education services (www.IDEAdata.org, based on 2005 data).

Source: IDEA Part B Child Count, 2005, "Students Ages 6 Through 21 Served Under IDEA," by disability category (Tables 1–3), www.IDEAdata.org (accessed September 2008).

This last point—that students with intellectual impairments can also achieve proficiency—is supported by the research of Dr. Kevin McGrew, a coauthor of the Woodcock-Johnson III, one of the most widely used instruments for assessing both cognitive abilities and achievement in children and adolescents. Using student testing data from the Woodcock-

Johnson III development processes, McGrew studied whether measured intelligence quotients, or IQs, can predict eventual academic achievement. He found that "it is not possible to predict which children will be in the upper half of the achievement distribution based on any given level of general intelligence. For most children with cognitive disabilities (those with below average IQ scores), it is not possible to predict individual levels of expected achievement with the degree of accuracy that would be required to deny a child the right to high standards/expectations."[1]

Still, it is common to hear educators, members of the public, or policymakers say, "Well, of course these students don't do well on the tests, they have disabilities!" Once those low expectations are entrenched, they play out in very destructive ways. The literature on the effects of teacher expectations on student achievement is compelling, demonstrating conclusively that what we expect in student learning is typically what we get, regardless of student ability. (See McGrew and Evans 2004 for a summary of the literature.) This is alarming, given that so many educators seem to believe that students with disabilities cannot learn the content expected for other students—or, as the earlier excerpt from Voice of San Diego suggests, that they have too many learning needs to warrant the effort required for them to learn it. The next section will address what content is expected for all students, including high school students with disabilities.

Standards-Based Reform, Expectations, and Student and System Accountability

Federal policy has played a significant role in improving how the public education system serves students with disabilities. By the end of the 1970s, federal policy (in the form of PL 94-142) guaranteed that students with disabilities had access to school buildings. Over the course of the 1990s, federal laws funding both special education[2] and education for the disadvantaged[3] as part of the standards-based-reform national agenda not only defined the right of students with disabilities to the same goals and standards as all other students, but also required the full inclusion of every student in assessments designed to provide data on how well all students were being taught. Most recently, the reauthorization of the Elementary and Secondary Education Act—known as the No Child Left Behind (NCLB) of 2001 (PL 107-110)[4]—and the passage of the Individuals with Disabilities

Education Improvement Act (IDEA) of 2004 (PL 108-446)[5] more closely aligned the two major education laws with common accountability for results. Although the focus in IDEA is on individual accountability and the focus in NCLB is on systems accountability, both laws are built on the goal of raising academic achievement through high expectations and high-quality education programs, to improve outcomes for all students, including those with disabilities.

Regardless of these laws, decisions about what every student should know and be able to do in a standards-based system are made at the state and local levels and not at the federal level. Since the late 1990s, policymakers and citizens in every state have grappled with the fundamental question of "What is a well-prepared student?" and each state has answered that by defining content standards (what) and achievement standards (how well) which identify essential skills and knowledge for students to master at each grade level. Cumulatively, these standards define what a high school graduate is expected to know and be able to do.

States receiving federal funding under either IDEA or Title I of NCLB are required to develop such standards, and NCLB requires standards to apply to all students in all public schools. The state-developed grade-level content and achievement standards are the foundation on which states build an assessment system. NCLB requires states to assess all students once annually in grades three through eight and at least once in grades ten through twelve in mathematics and reading, and once annually in each of three grade bands in science. IDEA clarifies that students who receive special education services are to have access to and make progress in the general curriculum based on these same standards, and reinforces the requirement of full inclusion of all students with disabilities in the NCLB-required assessments (as well as in all other assessments administered by the schools). IDEA further specifies that states and districts must provide appropriate options for all students with disabilities to participate in these assessments, including requirements for universal design of assessments, accommodations, and alternate assessments.

These assessments are used for a variety of purposes:

- NCLB requires states to use the assessments to hold schools accountable for the achievement of these standards by all students. A standards-based accountability system that requires the school system to demonstrate that all students can meet state or locally defined standards is meant to ensure that all students are prepared for a successful future. This is called **system stakes**, or **system-level/school-level accountability**. This means that there are consequences when public schools do not ensure that all students have mastered the essential content. Inclusive large-scale assessments used for system accountability are meant to shine a light on whether schools are teaching all students well.

- The use of standards-based assessments to hold students accountable for learning is called **high-stakes testing for students**, or **student accountability**. Typically, student accountability is meant to assure future educators or employers that this student knows and can do what the state has determined to be essential for future success.[6] The stakes for students may include some type of grade promotion or retention; in some states, assessments are used to decide whether or not a student may be granted a regular diploma.

It is possible that student accountability and system accountability working together may have powerful effects on student achievement, but there are unintended consequences when students are held accountable within a system that is not achieving the expectation that all students be successfully taught. Too often in the past, when some students or groups of students were not achieving, the easy answer was to lower individual or group expectations through strategies like those evidenced in the opening story of this chapter—"focus[ing] on skills that will help them live independently instead of prepping for college or beyond, studying shopping lists and sales tax instead of calculus or Cervantes … funneled into easier classes instead of making ordinary classes accessible through technology or other aids." As this experience shows, the disconnect can result in negative outcomes for some students, and raises questions about whether the students who have a right to a free appropriate public education (as required by IDEA) have in fact been provided with that opportunity.

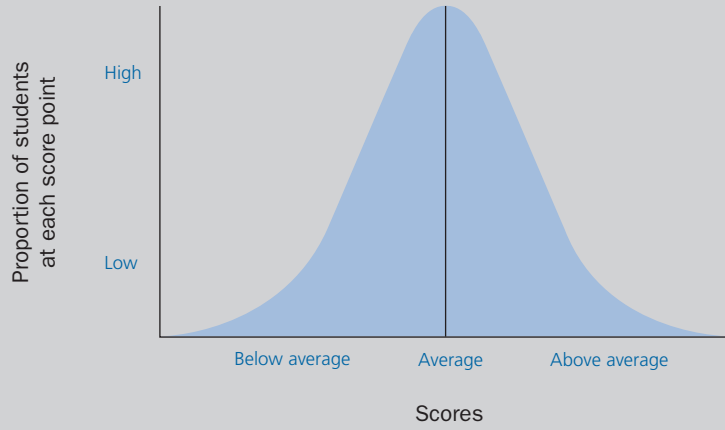
When some students, or groups of students, are not achieving to the levels expected for all students, the *expectations* should not change. Instead, *the services, supports, and specialized instruction* that ensure all students achieve the standards should change; in many cases, this requires increased training and support to the educators who work with these students. Through its standards and assessments requirements, NCLB has initiated a positive shift from the promotion of a separate and less rich or challenging curriculum for students with disabilities to an expectation of grade-level academic achievement as defined for all students. IDEA and NCLB jointly support this shift, but much work needs to be done to overcome decades of low expectations and deeply engrained beliefs among some stakeholders that nothing can be done to improve the achievement of students who have disabilities. As policy choices are made now and in the future, this history and set of beliefs needs to be articulated so that the implications of choices are clearly understood, and the consequences carefully monitored.

Understanding the Purposes of Assessments

Parents, policymakers, and the public can make informed decisions for individual students and about the policies that support improved student achievement by understanding the varying purposes and uses of common types of large-scale assessments. These varying purposes and uses define how students with disabilities should participate in each testing option. These fundamental differences are commonly misunderstood by many stakeholders in discussions of NCLB.

Norm-referenced tests (NRTs): One common use of tests has been to compare students' performance and rank-order them accordingly. NRTs are well suited for this, as these tests provide percentile ranks that tell us the percentage of a norm group—that is, the scores obtained by other students—that a given student performed as well as or better than. A graph called a "normal curve" or a "bell-shaped curve" similar to the one in Figure 2 is sometimes used to show that often most students perform about average and fewer students score much higher or much lower than the average. A student who performs as well as the average student will be equal to or better than 50 percent of the norm group and equal to or lower than 50 percent of the norm group. Hence, the average student is smack in the middle of this "normal" curve.

Figure 2: Norm-Referenced Test Results



NRTs are good for comparing students with respect to very general domains of performance, as well as for sorting out large populations for specific purposes (like army personnel assignments, or admission to college). However, they are not as well suited for evaluating students' performance with respect to curricula such as those defined in state-developed curriculum frameworks. They are also not meant to provide diagnostic information with respect to specific skill areas. Rather, they help to indicate strengths or weaknesses relative to others—that is, to the specific norm group that was available when the test was developed. Tests that are designed to measure what groups of students know compared to well-defined content and achievement standards require a different type of assessment called criterion-referenced testing (see next page).

More than a century of norm-referenced testing designed to distribute students along a normal curve has affected the perceptions of teachers, parents, policymakers, and the public—many are familiar with the NRTs they have themselves taken throughout their own academic career. This has resulted in a popular belief that on any skill taught, half of the students will perform "below average." In this context, there is a temptation to predict or assume *which* students will end up on the bottom. Although data from many states suggest that more than half of the students performing "below

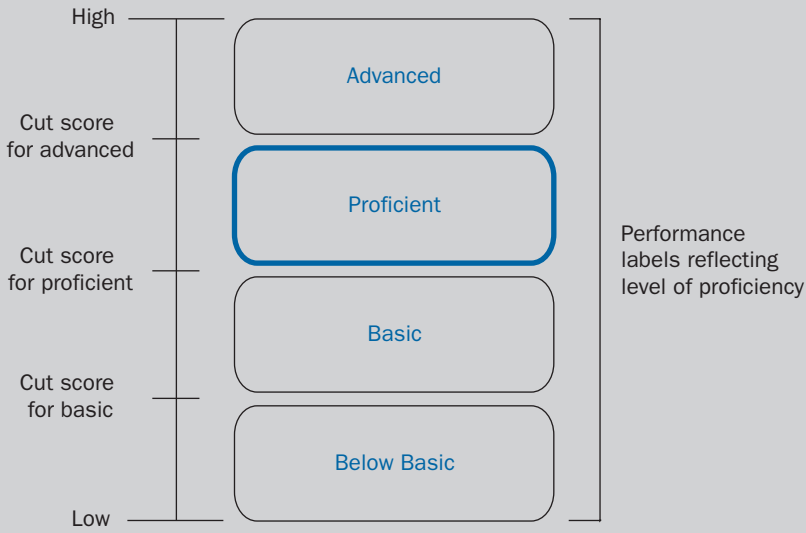
average" do not have disabilities and are disproportionately poor and minority students,[7] there is a negative and pervasive public perception that students with disabilities are always—by definition—the lowest-performing students.

Another drawback to norm-referenced testing is that it cannot indicate whether a student's test performance is satisfactory in and of itself. Scoring "above average" or "below average" means little if the "average" is very low or very high.

Criterion-referenced tests (CRTs): By contrast, criterion-referenced tests are designed to measure and provide information regarding how well students have mastered specific knowledge and skill areas they have been taught. For example, they can be used to answer questions regarding whether a student's performance on a test signifies "proficiency" in a subject area. Such achievement-level classifications are usually set on a test by establishing one or more "cut scores" to reflect standards determined by carefully selected groups of educators. Examples of these achievement standards include the performance standards established by states to meet the requirements of NCLB (e.g., below basic, basic, proficient, or advanced). Students' test results are compared to these cut scores to determine whether their performance is satisfactory. In CRTs, it is irrelevant whether a student's performance is at, above, or below the average of some norm group. Students who have been taught well should be able to demonstrate proficiency on the standards. Figure 3 depicts how CRT results reflect the levels of achievement of knowledge and skills defined by the academic standards.

In the past, when, after receiving the same instruction provided to all students, students with disabilities did not perform well compared to their peers, common practice was to provide a lower or slower curriculum than that taught to their peers. This inevitably resulted in an increasingly large gap in performance, year after year. By advancing the use of criterion-referenced assessments, NCLB promotes a shift in the teaching of students with disabilities. The altered goal contains three parts: to teach all students a rich curriculum based on the same content; to tailor services, supports, and specialized instruction for some students to be sure they master the content; and to test to ensure that they have indeed learned it.

Figure 3: Criterion-Referenced Test Results



Cut scores required for each category of performance are determined through systematic panel review processes by stakeholders with expertise in the content being assessed and experience with the students being tested.

Out-of-level testing: Sometimes, when comparing a low-performing student's test results to a norm group's performance (through NRTs) or to predefined achievement standards (through CRTs), information about the student's learning is limited. For this reason, educators have sought additional ways to measure the performance of some students—in particular, students who achieve at very low levels on the assessments designed for their enrolled grade. One strategy has been to use out-of-level testing, assessments based on below-grade-level skills and knowledge, instead of the skills and knowledge for a student's enrolled grade level. Prior to NCLB's requirements that states use CRTs, out-of-level testing was performed by many local schools, districts, and even states as part of accountability systems, and results were used in developing students' Individualized Education Plans (IEPs). This strategy was often employed in the name of shielding some students from difficult testing situations and protecting their sense of self-worth. However, research suggests that with these lower expectations came lower achievement.[8] A student with

disabilities who is consistently tested out of level may never be able to meet the requirements for high school graduation—a consequence that permanently affects their dreams, aspirations, self-worth, and postsecondary options. Special educators report that, too often, teachers, parents, and students do not think about these long-term consequences.[9]

Although out-of-level testing may be appropriate for some purposes, it is not designed to measure how well students are achieving against common expectations in a standards-based accountability system.

Adaptive testing: A related strategy for assessing students with disabilities has been the use of a type of adaptive testing. These tests typically include a computerized administration of test items that, based on how a student responds to earlier items (and often the relative difficulty of items), automatically selects additional items for each student. The use of such assessments results in individualized tests for each student. Advocates for the use of adaptive tests suggest that the individualized set of items will generate data that better matches what each student actually can do, as opposed to what they cannot do. Given the endless variation of what is tested in each situation, the results of these assessments also reflect varied and often unspecified content expectations, essentially measuring different standards for each student. In other words, putting a bank of test items on a computer that selects items for each student by difficulty does not generate useful information about how well the student has been taught the grade-level content. These adaptive assessment tools meet the requirements—and higher expectations—of standards-based reform only when they are designed to ensure that they are measuring student performance relative to grade-level content and achievement standards for the student's enrolled grade.

Options for Including All Students in Standards-Based Assessments

To effectively measure standards-based learning by all students, assessment systems must include options that ensure all students can demonstrate what they know. Currently, NCLB—both the legislation and related regulations—permit an array of assessment options for students with disabilities. These options are described below and summarized in Table 1.

General assessment, with or without accommodations: The vast majority of students with disabilities are given the general assessment that is taken by their peers without disabilities, with or without accommodations. These tests are developed with design elements—often called "universal design"—that make tests more accessible and result in more accurate scores that reflect actual student knowledge and skills, not extraneous factors. For example, most standards-based academic assessments used for school accountability are *not* intended to measure student characteristics and skills such as visual acuity, hand-eye coordination, or the ability to find isolated facts within a puzzle of distracting information. During the development of assessment items, universal design characteristics are considered in order to remove the effects of these kinds of extraneous and confounding factors. The result is an assessment that measures student abilities in the skills and knowledge intended to be assessed by the test, and not the unrelated effects of their disabilities.

Universal design does not mean that accommodations are no longer necessary. Accommodations are *changes in testing administration or materials* that enable students to participate in assessments in a way that allows their abilities—rather than the effects of their disabilities—to be assessed. Without accommodations, the assessment may not accurately measure a student's knowledge and skills. Even with assessments that are universally designed, some students may still need accommodations related to

- presentation (e.g., repeat directions, read aloud, large print, Braille, etc.);
- equipment and material (e.g., calculator, amplification equipment, manipulatives, etc.);
- response (e.g., mark answers in book, scribe records response, point, etc.);
- setting (e.g., study carrel, student's home, separate room, etc.); or
- timing/scheduling (e.g., extended time, frequent breaks, etc.).

Use of accommodations on a standards-based assessment assumes that careful consideration is given to whether the grade-level content and achievement standards being measured remain constant despite the use of the accommodation. The foundation for the assessment—the academic

content and achievement standards—remains the foundation even when accommodations are thoughtfully provided, thus maintaining high expectations for student achievement.

The collective knowledge base on the effects of accommodations on the content being measured is growing, but there are considerable complexities in the case of the most challenging content and student combinations. While it is widely accepted that most accommodations do not change what the assessment is measuring, more substantive changes in test administration and materials sometimes alter both what is measured and the meaning of the results. These are considered *modifications*. For example, a common focus in the lower grades is to measure a student's ability to decode printed text (as in the third-grade reading assessments). If the tester reads the test passages aloud to a student, this fundamentally changes the nature of the test, and measures reading comprehension, perhaps, but not the decoding of printed material. In this case, a read-aloud strategy would be considered a modification—a change that alters what is being measured—and not an accommodation.

However, for a small number of students, modifications are the only way they can interact with portions of the test. As noted by special education legal advocates Kathleen Boundy and Joanne Karger,

> Many students with specific learning/reading disabilities struggle greatly with decoding text, yet have strong comprehension skills when access to information is provided through alternative modes that include: auditory, tactile, visual, and a combination of auditory and visual modalities. Similarly, individuals who are visually impaired may not be able to decode text and participate in the State assessment without an accommodation that allows them access to the information and questions in the text on which their comprehension is being assessed. The failure to differentiate between decoding skills and the broader comprehension of information and range of literary knowledge that are within the scope of the academic content standards embedded in a language arts curriculum denies the meaningful and effective participation of students with specific learning disabilities, who are otherwise unable to participate.[10]

States are working to develop testing strategies that capture student knowledge without lowering standards or losing the integrity in measurement and the meaning of the test. A few states have made and defended sometimes controversial decisions on these issues, like those described above by Boundy and Karger, but these states also require close monitoring and accountability for schools where these accommodations are selected for students. There is a delicate balance between access and inadvertently lowering standards and, thus, expectations and outcomes.

Alternate assessments: There is a small group of students with disabilities who cannot demonstrate what they know on the regular assessments, even with the use of accommodations. For these students, NCLB permits the development of alternate assessments to measure how well students know the skills defined in their enrolled-grade content standards. Over the past decade, there has been a rapid evolution of our understanding of how students build competence, especially with regard to students who need alternate assessments, but many misconceptions and erroneous assumptions around this issue remain. It is important to build a common understanding of the distinctions so the goals of reform are not derailed by these misunderstandings. Alternate assessments provide a critical role in ensuring that we obtain truly accurate measures of the knowledge and skills of all students with disabilities. NCLB allows for three alternate assessment options, described below and summarized in Table 1.

Alternate assessments based on grade-level achievement standards (AA-GLAS) are meant to assess the *same* content with the same definition of "how well and how much" as is measured by the general assessment. The format of the test is the key variable—for example, a portfolio of student work and a panel of content experts/teachers who review each portfolio instead of a multiple-choice or constructed-response test. There are very few such alternate assessments in place in states now, in part because of the very difficult challenge of showing comparability to the general assessment. This challenge is in part a policy decision that requires high comparability as a control on potentially lowered standards being hidden by such alternate assessments. Still, there are policy options that would avoid lowering standards and also allow for multiple ways to demonstrate grade-level achievement. The goal is to ensure that students who have disabilities that

interfere with good measurement of the skills and knowledge they do in fact have are included in accountability on the equivalent expectation of grade-level achievement standards. The challenges faced in designing the strong accommodations policies described by Boundy and Karger are directly related to these potential policy shifts. By reexamining the opportunities of the AA-GLAS, policymakers should find ways to avoid unnecessary use of less rigorous testing options like the remaining two alternate assessments, discussed below.

Alternate assessments based on modified achievement standards (AA-MAS) are meant to assess the *same* content with a possibility of a slightly less difficult definition of "how well and how much" than is used for the general assessment. It is unclear from NCLB regulatory language and guidance what exactly that means, and not every state is opting to develop an AA-MAS. If a state chooses to develop an AA-MAS, it is required to define the standard of "how well and how much" through a documented and validated standard-setting process, involving stakeholders who know the student and the content and who in theory have the best interests of the students in mind. Given the history of low expectations and deeply held beliefs among some stakeholders that many students with disabilities cannot learn the same content as their peers, these definitions are of concern to many advocates, who fear that lowered expectations will be reinforced by provision of this testing option.

The governing regulations limit the number of IDEA-eligible students who can be determined proficient for purposes of determining whether a school, district, or state has made AYP, based on the modified assessment—up to 2 percent of the general school population, or about 20 percent of special education students. To date, no state has received full approval through the federal peer-review process for its AA-MAS.

To complicate the identification of students who can take the AA-MAS, the data from multiple states indicate that the students who are the lowest-performing 2 percent on regular assessments are a blend of students with and without disabilities, and who predominantly represent student groups who have historically been on the low side of the achievement gap.[11] However, the current regulation allows states to implement the modified

assessment only for students with disabilities. This data, combined with evidence that many of these students have not been taught the challenging curriculum expected for all students, suggest a need for thoughtful and data-based processes to understand what a modified achievement standard should represent.

Over the next several years, states will need to work in partnership with researchers and experts to better understand and identify the appropriate students for participating in the AA-MAS. A key question in this process is whether these students have been provided the services, supports, and specialized instruction they need to succeed. The goal of this work must be to understand how these students can build and demonstrate the skills and knowledge they need to earn a regular diploma and to succeed in adult life. States must focus on the development of modified achievement standards and modified assessments that raise expectations for all students and close persistent achievement gaps.

Some have argued that this testing option has resulted in unexpected positive consequences—specifically, that there is attention placed on students who previously were ignored. Even so, it has the potential to create unintended but negative effects that may perpetuate low expectations, sustain achievement gaps, and limit students' access to graduation with a regular diploma and college and work opportunities.

Alternate assessments based on alternate achievement standards (AA-AAS) are meant to assess the *same* content with less depth, breadth, and complexity than the regular assessment, and with a *different* definition of "how well and how much." Just as is the case with the AA-MAS, states must define these standards using a documented and validated standard-setting process, and there are concerns about whether these standards reflect an appropriate raising of the bar of expectations that will yield increased achievement. There is, however, less controversy among advocates about whether this type of alternate assessment is a necessary option. Instead, there is debate within special education about what content should be assessed—given the flexibility states have in defining the depth, breadth, and complexity of the content to be assessed—and which students should be eligible for this testing option.

The AA-AAS are intended to be used with students with significant cognitive disabilities, typically defined as those with the most severe intellectual disabilities and multiple disabilities—children who represent fewer than 1 percent of all students, or less than 10 percent of all students who have disabilities. When the regulation permitting the AA-AAS was proposed, stakeholders debated which students should be included. Estimates of how many students have the most severe intellectual and multiple disabilities ranged from less than 0.5 percent of the total student population to as high as 3 percent. The lowest percentage (0.5 percent) was supported by data in states that report moderate and severe mental retardation as separate from all students with mental retardation and from Centers for Disease Control data on incidence of correlated disability diagnoses.[12] The higher estimates generally included many students with mild disabilities from all disability categories, many of whom do not have intellectual disabilities but who were performing at low levels. The eventual selection of 1 percent as the cap on the percent of students whose scores can be treated as proficient for purposes of school accountability was a compromise. Certainly, more students can participate in the AA-AAS than 1 percent, but from a policy perspective the cap on how the scores are used in accountability was intended to prevent inappropriate inclusion of many students in a lower achievement expectation than evidence suggests is warranted.

Inclusive assessment systems have been the cornerstone of policies based on an assumption that by including all students in assessments used to determine how well schools have taught all students, educators will be motivated to ensure that students who have not had access to the general curriculum in the past will be taught. Studies from multiple states show that students with significant cognitive disabilities have benefited from a noticeable increase in their access to the general curriculum because of the NCLB requirements for these assessments. The consequential validity studies of the AA-AAS document the benefits of the policy push of including all students in standards-based assessment and accountability systems. There have been reports of dramatic increases in other valued outcomes for these students, such as increased use of assistive technology, which leads to an increased level of independence, an increased implementation of inclusive settings, and an increased interaction with typical peers.[13]

The achievement of these students on grade-level content is very different from their general education classroom peers, but the evidence of their work is compelling. These students are able to learn academic content with reduced complexity, breadth, and depth clearly linked to the same grade-level content as their peers. (The federally produced publication *Learning Opportunities for Your Child Through Alternate Assessments* provides specific examples of what that can look like.)[14]

Researchers and practitioners are working side-by-side to capture the nature of the linkages to the grade-level content, but the evidence of this improvement in student learning is startling, given that schools have not given these students access to this content in the past.[15]

Examining the Effect of Inclusive Testing Practices

Current test results show us that, generally, students with disabilities are not performing as well as typical peers. As states have examined this gap, many have found that the students with and without disabilities who currently score low on tests often have not been taught the tested content.[16] These investigations have served to raise awareness about improving instruction for students with disabilities. The combination of the pressure to test all students and the focus on improving instruction has increased the pressure on schools to learn what works to ensure successful outcomes for all students.

To supplement anecdotal evidence, many states have instituted formal procedures to use assessment and accountability data to identify schools where reforms are yielding very high achievement for students with disabilities. There are also a few well-designed studies focused on what is occurring in schools where test scores are higher for students with disabilities. These studies consistently identify common characteristics among schools where students with disabilities achieve at high levels. As summarized in one study, the schools have

- a pervasive emphasis on the curriculum and alignment with the standards;
- effective systems to support curriculum alignment;
- an emphasis on inclusion and access to the curriculum;

- a culture and practices that support high standards and student achievement;
- a well-disciplined academic and social environment;
- continuous use of student data to inform decisionmaking;
- unified practices supported by targeted professional development;
- access to resources to support key initiatives;
- effective staff recruitment, retention, and deployment;
- flexible leaders and staff who work effectively in a dynamic environment; and
- effective leadership.[17]

In the past decade, NCEO's surveys of states have recorded state staff perceptions of changes occurring in their districts and schools. Survey respondents speak of improvements in the performance of their students, attributing those improvements to clear assessment participation policies, alignment of student Individualized Education Plans (IEPs) with standards, improved professional development, development and provision of accommodation guidelines and training, increased access to standards-based instruction, and improved data collection.[18] Analyses of publicly reported assessment data since 2000–01 show improvements in the transparency of data for students with disabilities, for both participation and performance.[19] For example, according to NCEO's research, the number of states with clear reporting to the public about students with disabilities' participation in testing options increased from only five states in 2000–01 to twenty states in 2004–05. These data also showed large increases in the percentage of students with disabilities who participate in the assessment system across time for most states.

A Vision for a Principled Approach to Accountability Assessments for Students with Disabilities

Building on research and practice, NCEO has identified the principles and characteristics that underlie inclusive assessment and accountability systems.[20] The vision for a principled approach to accountability assessments for students with disabilities includes the following six core principles:

Principle 1. All students are included in ways that hold schools accountable for their learning.

Principle 2. Assessments allow all students to show their knowledge and skills on the same challenging content.
Principle 3. High-quality decisionmaking determines how students participate.
Principle 4. Public reporting includes the assessment results of all students.
Principle 5. Accountability determinations are affected in the same way by all students.
Principle 6. Continuous improvement, monitoring, and training ensure the quality of the overall system.

These principles reflect the belief that all students with disabilities can and should have meaningful access to the same education as their peers without disabilities. For that to occur, they need to be taught well in the same curriculum and with the same expectations as their same-age classmates. Teachers should use systematic standards-based assessments both in the classroom as they are teaching and at the end of instruction to know how well the students were taught. These assessments will give schools the information they need to design instruction and supports so these students achieve in spite of barriers related to disabilities.

Although standards-based assessments are essential tools in the drive to ensure all students achieve at high levels, assessments can simply point out where teaching and opportunities to learn need to be improved. For improvement to occur, teachers need and deserve high-quality training, coaching, and professional support so they can be successful teaching all students. Staff development support to teachers is a necessary but sometimes neglected component of standards-based reform. Parents and the public can join forces to ensure that all teachers have the skills they need to do this important and challenging work. Our schools must be structured to allow students with disabilities to avoid the barriers that their disabilities create when accessing the curriculum and when demonstrating what they know and can do on assessments. Success in doing so is a critical step on the path toward lifelong success.

Policy Conclusions

National and state efforts in standards-based reform should include all students in standards, assessment, and accountability systems and

processes, including all students with disabilities. Any attempt to exempt schools from being accountable for a group of students undermines the entire reform effort. Previous assessment and accountability policies supported by federal funding for the disadvantaged (e.g., allowing the use of NRTs and out-of-level testing) and for students with disabilities (e.g., reliance on individual student accountability provisions connected to IEPs) have contributed to the well-documented achievement gaps. Federal policy related to standards-based reform, including standards, assessments, and accountability systems and processes, must adhere to this principle to diminish the negative effects of applying different expectations to different groups of students.

Standards-based assessment systems should include strategies that permit all students to show what they know and can do on the academic content standards defined for typical peers of the same age and grade level, despite the barriers of disability. This includes a continued emphasis on universal design of assessments, the development and implementation of high-quality accommodations policies, and provision of alternate assessments that allow different ways of demonstrating what a student knows and can do, including expanded options for alternate assessments based on grade-level achievement standards (AA-GLAS).

Any change in academic achievement standards for a group of students, including those already defined in regulatory language, should be reviewed to ensure that these options raise the bar of academic expectations, and thus increase system accountability for the outcomes of students who may participate in the option. Each option should be able to withstand scrutiny by external experts on whether the underlying assumptions are grounded in current research or practice that supports improved academic achievement for students who may participate in these options. Although there is evidence emerging from the use of the AA-AAS to that effect, there does not appear to be a similar body of evidence from the use of the AA-MAS thus far, or in the research base cited in the regulation.

The consequences of participation in any testing option that changes the achievement standards (including those options in place now and those proposed in the future) should be closely monitored. Higher achievement should be evidenced independently of state-set proficiency levels, which

could be artificially elevated by state-defined modified or alternate achievement-standard-setting practices. Careful study of performance-level descriptors and achievement-standard-setting practices can inform conclusions of whether low expectations are reinforced by these options. If so, any option that does not withstand scrutiny as a high expectation standard should be discontinued so that schools, districts, and states are held accountable for educating all students to high standards.

Afterword: Story from San Diego

Excerpt from E. Alpert, "Deterred from Diplomas for Better or Worse," Voiceofsandiego.org, October 2, 2008:

> As graduation rates have grown, brain research has shown the risks of underestimating children with disabilities, even those with severe conditions that prevent them from speaking, said Anne M. Donnellan, director of the Autism Institute at University of San Diego's School of Leadership and Education Sciences. Donnellan has seen a number of nonverbal students such as Peyton Goddard overcome diagnoses of mental retardation and graduate from college.
>
> Hehir [Tom Hehir, Harvard researcher and former assistant secretary of education in the Clinton administration] likewise noted that diagnoses are sometimes wrong and students should be given the benefit of the doubt. For instance, conventional wisdom that students with Down syndrome couldn't learn to read has been shattered as many prove themselves capable of reading and writing as well.
>
> "I'm not interested in predicting what people can do," Donnellan said. "We've made some terrible mistakes with that."
>
> Lance Rogers believes he was a victim of those mistakes. Now a sophomore at the Marcy School, a San Diego Unified center that combines classes and counseling, Rogers said he's taking algebra, chemistry and history to earn the diploma he once was blocked from.
>
> By pursuing a diploma, "I did something they didn't think I was going to accomplish," he said. "They didn't say it like that. But that's what it comes down to."

The views expressed in this chapter are those of the author and do not necessarily represent those of the Alliance for Excellent Education.

About the Author

Rachel Quenemoen is a senior research fellow at the University of Minnesota and the technical assistance team leader for the National Center on Educational Outcomes. She is coprincipal investigator of NCEO's federally funded technical assistance center. Ms. Quenemoen has worked for thirty years as an educational sociologist focused on research-to-practice efforts. She has been a multidistrict cooperative administrator in both general and special education, and for the last fifteen years has worked at the state and national levels on educational change processes and reform efforts related to standards-based reform and students with disabilities, building consensus and capacity among practitioners and policymakers. Her current research and technical assistance priorities include alternate assessment of students with significant disabilities and research focused on the causes of and solutions for the achievement gap between students with disabilities and their typical peers. She is the author of numerous chapters, articles, presentations, and papers related to inclusive assessment of students with disabilities, and has coauthored a book on alternate assessment. She is the proud parent of an adult daughter who has Down syndrome.

---

[1] K. S. McGrew and J. Evans, *Expectations for Students with Cognitive Disabilities: Is the Cup Half Empty or Half Full? Can the Cup Flow Over?* (Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes, 2003).

[2] Individuals with Disabilities Education Act Amendments (IDEA) of 1997, Public Law 105-17, 105th Cong., 1st sess.

[3] Improving America's Schools Act of 1994, Public Law 103-382, 103rd Cong., 2nd sess.

[4] No Child Left Behind Act of 2001, Public Law 107-110, 107th Cong., 1st sess.

[5] Individuals with Disabilities Education Improvement Act (IDEA) of 2004, Public Law 108-446, 108th Cong., 1st sess.

[6] M. Johnson, L. Thurlow, and K. E. Stout, *Revisiting Graduation Requirements and Diploma Options for Youth with Disabilities: A National Study* (Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes, 2007).

[7] M. Perie, *Understanding the AA-MAS: How Does It Fit into a State Assessment and Accountability System?* presentation to CCSSO SCASS meeting, February 4, 2009, http://nciea.org/cgi-bin/pubspage.cgi?sortby=pub_date (accessed March 17, 2009).

[8] J. Minnema, M. Thurlow, and S. H. Warren, *Understanding Out-of-Level Testing in Local Schools: A First Case Study of Policy Implementation and Effects* (Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes, 2004).

[9] Ibid.

[10] K. Boundy and J. Karger, "Special Issues Affecting Inclusion of Students with Dyslexia in Statewide Assessments and Their Implications," *Perspectives on Language and Literacy International Dyslexia Association* 34, no. 4 (2008): 36–40.

[11] Perie, *Understanding the AA-MAS*.

[12] U.S. Department of Education, *Raising Achievement: Alternate Assessments for Students with Disabilities*, http://www.ed.gov/print/policy/elsec/guid/raising/alt-assess-long.html (accessed March 8, 2009).

[13] H. Kleinert, S. Kennedy, and J. Kearns, "Impact of Alternate Assessments: A Statewide Teacher Survey," *Journal of Special Education* 33, no. 2 (1999); M. Turner, L. Baldwin, H. Kleinert, and J. Kearns, "An Examination of the Concurrent Validity of Kentucky's Alternate Assessment System," *Journal of Special Education* 34, no. 2 (2000).

[14] U.S. Department of Education, Office of Special Education and Rehabilitative Services, *Learning Opportunities for Your Child Through Alternate Assessments* (Washington, DC: U.S. Department of Education, 2007).

[15] D. M. Browder, S. L. Gibbs, L. Ahlgrim-Delzell, G. Courtade, A. Mraz, and C. Flowers, "Literacy for Students with Significant Cognitive Disabilities: What Should We Teach and What Should We Hope to Achieve?" *Remedial and Special Education*, in press; C. Flowers, S. Wakeman, D. Browder, and M. Karvonen, *Links for Academic Learning: An Alignment Protocol for Alternate Assessments Based on Alternate Achievement Standards* (Charlotte, NC: University of North Carolina at Charlotte, 2007); see http://www.nceo.info; C. Flowers, S. Wakeman, D. Browder, and M. Karvonen, "An Alignment Protocol for Alternate Assessments Based on Alternate Achievement Standards," *Educational Measurements: Issues and Practice*, in press.

[16] Perie, *Understanding the AA-MAS*.

[17] Donahue Institute, *A Study of MCAS Achievement and Promising Practices in Urban Special Education: Report of Field Research Findings (Case Studies and Cross-Case Analysis of Promising Practices in Selected Urban Public School Districts in Massachusetts)* (Hadley, MA: University of Massachusetts, Donahue Institute, Research and Evaluation Group, October 2004), http://www.donahue.umassp.edu (accessed March 16, 2009).

[18] S. J. Thompson, C. J. Johnstone, M. L. Thurlow, and J. R. Altman, *2005 State Special Education Outcomes: Steps Forward in a Decade of Change* (Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes, 2005).

[19] M. L. Thurlow, R. F. Quenemoen, S. S. Lazarus, R. E. Moen, C. J. Johnstone, K. K. Liu, L. L. Christensen, D. A. Albus, and J. Altman, *A Principled Approach to Accountability Assessments for Students with Disabilities* (Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes, 2008).

[20] M. Thurlow, R. Quenemoen, S. Thompson, and C. Lehr, *Principles and Characteristics of Inclusive Assessment and Accountability Systems* (Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes, 2001); M. Thurlow, R. Quenemoen, J. Altman, and M. Cuthbert, *Trends in the Participation and Performance of Students with Disabilities* (Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes, 2007).

CHAPTER 9

Assessments and Technology: A Powerful Combination for Improving Teaching and Learning

Erin Martin Gohl, Daniel Gohl, and Mary Ann Wolf
State Educational Technology Directors Association

It is time to move the American educational enterprise toward equity of opportunity for every learner and a high return on investment for resources and time dedicated toward the development of the future body politic. To do so, policymakers and educators must promote accountability, efficiency, and individual student development. The combination of improved, focused, and aligned assessments with the transparency and communication strengths of technology can help meet these goals.

This chapter describes how the use of technology to assess students and to record and analyze performance can result in timely, appropriate, and individualized instruction for all students. It will highlight some of the innovative approaches in using technology to assess student progress, address current challenges in the use of technology, and provide recommendations to federal policymakers to overcome those challenges.

The Power of Assessments

Assessment and the reporting of performance are an intrinsic component of contemporary education. Assessment is the mirror through which students, teachers, educational bureaucracies, and communities evaluate educational effort and investment. Whether through a classroom quiz, a district-wide end-of-course exam, a state test mandated by the No Child Left Behind Act (NCLB), a privately funded college-readiness test, or an international comparison exam, judgments of educational success are made based on the performance of learners on an assessment instrument.

Because of this, assessments can be powerful tools by which educational stakeholders make decisions to improve upon the achievement of a particular student or group of students. District, state, or federal decisionmakers use summative assessments to measure the collective effort of schools and cohorts based on school entrance or intended graduation year. These collective measurements are used to analyze and inform resource allocations and judgments of effectiveness. By using certain assessments, some administrators are able to differentiate the approaches of different teachers in the same subject matter to determine those who are more effective or less effective. For an individual teacher, assessments are mechanisms for gauging how a learner has mastered the material or skill presented by a teacher. By seeing when performance is high or low, as compared to the way in which the material is presented, teachers are able to improve, in an iterative fashion, their own ability to instruct students. Because most classrooms in the United States contain students with a wide range of abilities, backgrounds, and learning styles, quality assessments are essential for teachers to adapt their teaching to varying needs. For a student, assessments provide an indicator of progress in the educational system, with high or low performance providing reward or concern. Changes in effort, approach, or instruction can be made after poor performance on an assessment, which will hopefully lead to more effective and productive learning. Effective instruction—and the teacher training and resources that support it—should enable a teacher to accommodate students with different learning styles, provide both enrichment and remediation, and allow for personalized instruction for each individual student rather than a single instructional mode for the whole class or group.

Technology Supports Assessment That Improves Teaching and Learning

The use of technology can improve the assessment process through the delivery, sharing, comparison, and analysis of assessment instruments and assessment effectiveness. The efficient implementation of assessment through technology, and the decisions upon which the assessment results are made, will dramatically increase the time available for direct and individualized instruction to students. Technology use can increase the efficient use of classroom time for assessment administration, reduce the human workload for the grading of assessments, improve the recording of assessment results, improve the communication of immediate and longitudinal performance for every student, aggregate and analyze performance within and across cohorts of assessment takers, and allow for practitioners and policymakers to share data to inform decisions on resource allocation. The following section describes the role technology can play related to specific assessment types.

Snapshot assessment: Technology can enhance typical snapshot assessments administered to capture student knowledge at the end of a teaching unit, semester, course, or grade. For example, through technology-enabled snapshot assessments (summative or interim), students can be asked to engage in a broader range of prompts (such as video and audio) than is possible with static print. Writing tools such as the word-counting feature and spellcheck can be incorporated into assessments so that students can focus on the quality of the submission to a free-response prompt. Students can also be expected to demonstrate their knowledge by, for example, generating databases or graphs that reflect underlying structures of information provided to them, and to display creativity with form and color in brief periods of time that would not be possible without technology. These technology-based prompts are better reflections of the tasks required by universities and employers of postsecondary graduates.

Portfolio/performance-based assessment: In addition to improving the delivery of snapshot assessments, technology greatly enhances the implementation of portfolio and performance assessments. Through the utilization of technology, the use of multiple visual formats (text and graphics) is broadened, the extension to video and audio is enabled, and

the sharing of products is normalized. In addition, students have the opportunity in technology-based portfolio and performance assessments to include responses from audiences beyond their immediate teacher and community. Websites and social networking technologies open the process of critique to the world of informal response. This allows for students to make iterative change in portfolio products, or revisions in the performance, that is informed by teacher collaboration, community response, and personal reflection. Assessments at each point of iteration can serve as formative appraisals. The use of technology allows for the archiving of student work, the documentation of student alterations to work products, and the navigation of collections of student work.

Classroom assessment: Through the use of technology, classroom teachers can conduct innovative formative assessments of all students for the purpose of improving instruction. This provides exciting new opportunities for the remediation or enrichment of each and every student, helping *all* students reach their highest potential. Given the way that technology can now alter the speed and location of assessment, many options now exist to embed "on the fly" assessment into curriculum content and lessons themselves. The days of handwritten records and paper copies of classroom assessments are quickly fading. With increased curriculum content to cover, most teachers do not have the time to use paper methods for formative assessment. Widely available technology tools provide an efficient and effective option for formative assessment. Handheld devices for reading assessment, electronic response systems, and software are all technology-based formative assessment tools that have the power to help teachers effectively individualize instruction for all students. Blogs, chats, and Wikis used in the classroom environment help teachers gain an understanding of what students know and don't know. Many districts and states are using technology-based programs and systems that provide teachers with formal and informal assessments to track student progress weekly or even daily. These types of formative assessments help keep students on track with achievement, while also providing opportunities for students to participate in engaging activities based on abilities and needs.

Adaptations: Technology can be used for the administration of adaptive, assistive, and alternative assessments that make use of an extensive range of

modifications and accommodations. Whether by restoring physical ability or translating language, for example, technology enables more students—such as those with disabilities or limited English proficiency—to participate in assessments and demonstrate their proficiency. This helps ensure that more students are held to universal educational expectations (a prerequisite for an equitable society) and that students are evaluated in an equitable manner that maintains the validity of the evaluation for inclusion in group-based comparisons.

Sharing of best practices: The use of technology for improving formative assessment is not limited to teachers and students in individual classrooms engaging in iterative improvement in isolation. Technology enables the sharing of assessment instruments, the rapid dissemination of innovation, and the tools through which to align classroom, district, state, national, and international assessment questions. By enabling assessment instruments to be viewed, and therefore shared, the challenge of developing new assessments can be addressed simultaneously with the need for transparency. Technology can create communication mechanisms, test question repositories for public access, and secure domains for test question development.

Technology in Action

Highlighted below are a few examples from states and districts using technology-based assessments to individualize instruction to improve student achievement, remediate before it's too late, track individual student growth and progress, and achieve school-improvement goals.

- **Texas TAKS**
  Texas uses a computer application as an electronic bridge between state test results—which identify each student's strengths, weaknesses, and areas of needed improvement—and the supporting instructional software. Each student's individualized learning path is created, and student assignments are based on objectives that were not mastered on the state test. Optional progress assessments may be administered during the year to allow teachers to monitor and modify student progress within the learning paths as needed. Further, teachers have an opportunity to add learning activities or create alternative learning paths based on classroom priorities.

Summative assessments are provided at the end of each year to gauge student progress and readiness for the next grade level.

- **Virginia's Algebra Readiness Initiative**
  Virginia's Algebra Readiness Initiative (ARI) assists in preparing students for success in algebra through a computer-adaptive test. School divisions are eligible for incentive payments to provide mathematics intervention services to students in grades six through nine who are, as determined by diagnostic tests, at risk of failing the Algebra I end-of-course test. The diagnostic test results allow teachers to individualize the content for intervention. A pilot study conducted during the 2005–06 school year to explore the efficacy of this approach in grade five showed that students improved by more than eighty scale score points between the Algebra Diagnostic Test given at the start of the year and the one given at the end. Teachers reported that the ARI helped determine the learning styles of their students (i.e., a preference for formula-based learning versus more hands-on math activities) and ultimately allowed for appropriate teaching modifications.

- **Indianapolis Public Schools, Indiana**
  Beginning in the 2007–08 school year, Indianapolis Public Schools adopted a software and reporting platform that fully integrates their current core curriculum with formative assessment data, helping to take data-driven decisionmaking even further by providing teachers with explicit support in using individual student data to pinpoint appropriate and effective basal lessons. By creating a strong link between the software-enabled assessment and the district's curriculum, educators were better able to craft instructional plans targeted specifically to their students' learning needs. The district has made consistent gains year after year. During the 2007–08 school year, 49 percent of the K–3 students who were identified as being at high risk for reading difficulty at the beginning of the year left the high-risk category by the end of the year, with 27 percent of those students reading at or above Benchmark level; in addition, 46 percent of students identified as being at some risk for reading difficulty were reading at Benchmark by the end of the year.

- **Arizona's Formative Assessment Item-Bank**
  Through its IDEAL (Integrated Data to Enhance Arizona's Learning) Web portal (http://www.ideal.azed.gov), Arizona provides a range of assessment tools, accessible by all teachers in the state. This includes a formative assessment item bank with over 5,500 items, and more than one hundred pre- and post-assessments, including performance objective snapshots, all aligned to the state's standards.

- **Alaska's GLE Item Sampler**
  Alaska's Formative Assessment GLE Item Sampler provides a bank of formative assessment items aligned to the Alaska Grade Level Expectations in math, reading, and writing for grades three through ten. These assessments are intended for use by all Alaska teachers to guide and adjust their instruction during the learning process and to differentiate classroom instruction so that the needs of each student are met.

- **NAEP Test: Problem-Solving in a Technology-Rich Environment (TRE)**
  Through its Technology-Based Assessment Project, the U.S. Department of Education is exploring the use of new technology in administering the National Assessment of Educational Progress (NAEP) to measure skills that cannot be easily measured by conventional paper-and-pencil means. For example, as part of the Problem-Solving in a Technology-Rich Environment (TRE) pilot, tasks to measure eighth-grade students' mastery of the kind of problem solving done with computers in educational and work environments was embedded within a physical science assessment. Students were given two extended scenarios designed to measure their ability to solve problems using technology. The assessment required students to search the Internet (using a simulated World Wide Web environment) and locate and synthesize information about scientific helium balloons. The "simulation" scenario required students to conduct experiments of increasing complexity about relationships among buoyancy, mass, and volume. These scenarios were delivered via school computers or on laptop computers

brought into the schools. The assessment produced a total score and separate scores for computer skills, scientific inquiry, scientific exploration, and scientific synthesis.

Barriers to Technology Integration and Assessment Improvement

Though there are these and other cases around the country where schools, districts, and states are using innovative technology-enabled assessments to improve teaching and learning, the full potential impact of technology to assure effective assessment in support of educational achievement has yet to be realized. In order to improve the quality of assessment and the integration of technology, there are a number of challenges that must be overcome, including the ones listed below.

Technology infrastructure: A major challenge for many school districts in using technology to administer assessments is the lack of an adequate technology infrastructure to support the broad-based implementation of such efforts. Many of these assessments require high-speed broadband as well as classroom access to equipment, which many schools lack. Thus, assessments that require broadband would be sluggish or unusable in these schools. In many districts, access to computers and other hardware to utilize technology-based assessments is also limited, with some schools' technology primarily available in labs, apart from classrooms. Until the technology is in the classroom and teachers can use it as a natural part of their teaching, the full potential of technology-enabled assessment will not be realized.

Interoperability: As assessments systems (tests, projects, etc.) are developed, it is important that information is able to be shared among practitioners, appraisers, and researchers. Currently, there is no standard format for presentation and distribution of results. Without a standard framework, the data collected from the resources cannot be appropriately used for research and evaluation.

Teacher training: In order for teachers to effectively use technology-enabled assessments, they must be properly trained to integrate these practices into their everyday classroom instruction. Many states and school districts, however, do not sufficiently fund sustainable professional development opportunities around effectively using technology to assess students. The

National Staff Development Council advocates that "at least 30 percent of the technology budgets be devoted to teacher development because technology purchases have increased dramatically in many school districts during the past decade, often with little attention given to the development of teachers' abilities to use the technology." Opportunities for teachers to learn, plan, and practice are critical to maximizing the potential of technology to improve student achievement.[1]

Lack of communication with all stakeholders: Technology is often seen as separate from mainstream curriculum. Administrators, curriculum specialists, professional development leaders, teachers, and technology support staff often work in silos. However, in order for technology-enabled assessments to be effective and fully integrated into teaching practice, stakeholders must communicate regularly so that all parties understand and commit to a comprehensive educational and professional development process. For example, information technology staff members need to work with other members of the educational team during the planning and budgeting process so that broadband and access issues are addressed.

Deficiency of pre-service programs to address technology integration: Training for the use of technology and technology-based assessments needs to begin during pre-service teacher training programs. Currently, these programs rarely employ technologies that are utilized in assessment. Colleges of education must modernize their pedagogical instruction to best prepare teachers for twenty-first-century classrooms, including technology-integrated instruction and assessment. Pre-service programs for both teachers and administrators must establish the expectation that technology is critical to improving the range of assessment, provide experiences using technology-enabled assessments, and offer training to improve the use of such assessments in the classroom and analysis and use of their results.

Inertia of vision: The procurement, implementation, and use of technology in schools has often been perceived as an optional expenditure with local funds. The federal E-Rate program, which provides need-based discounts to help U.S. schools and libraries obtain affordable Internet access and telecommunications, has allowed local and state funders to rapidly expand the infrastructure of connectivity. No subsequent funding commitments or

policy priorities have built upon the platform of connectivity constructed through E-Rate. As a result, there is a strong foundation to use technology for delivering, communicating, and analyzing assessments and results. Leadership is required to finish the fulfillment of this vision.

In addition to the challenges of integrating technology into classroom instruction and assessment, more general improvement must occur in particular areas in order to maximize the potential for assessments to improve teaching and learning. Some of them are listed below.

Teacher and administrator training: In order to be effective, teachers must be able to analyze the data produced by the assessments. Many teachers and administrators, however, have not received this training in either pre-service or in-service coursework. The skills of planning, delivering, and analyzing the results from technology-enabled assessment need to become—along with classroom management and standards-based instruction—a standard expectation for all educators.

Lack of classroom time for assessment analysis and reteaching: In the current climate, time demands caused by the breadth and depth of standards and curriculum limit the time for assessment and are so severe as to almost preclude reengagement with material after its initial presentation. Further, teachers are expected to be the primary designers, implementers, and graders of assessments. This limits time for engagement with students.

Curriculum: The current standards-based curriculum used by most districts, with daily pacing charts and dominant use of heterogeneous grouping of large student classes (not allowing for factors such as learning style, special needs, or language fluency), causes teachers to feel pressured to teach everyone in the classroom as a unit. This results in teachers not being able to address any gaps in understanding revealed by formative assessment.

State longitudinal data systems: With the federal testing requirements of the Elementary and Secondary Education Act, most states have been administering standardized tests for more than a decade. Several states, with some federal support from the Institute of Education Sciences State Longitudinal Data Systems grant program, have begun to grow these

state data systems by integrating test scores with key demographic and achievement information from students. However, even these states have lacked sufficient time, resources, support, and training to effectively utilize that data to intervene in student achievement across the state. Statewide longitudinal data systems are crucial both for accountability and for providing comparative data across district and state lines to ensure that all students are receiving relevant instruction aligned to baseline academic standards. However, state systems are not designed to drill down to the student and teacher levels for the purpose of individualizing instruction.

The Role of Policy in Overcoming These Challenges

The development of coherent policy at the local, state, and federal levels that allows for assessment development and implementation to be aligned is critical to ending the isolated efforts that characterized twentieth-century educational accountability. Policy frameworks and adopted policy need to direct professional efforts toward transparency of the assessment instruments, alignment with standards, and the manner in which achievement is reported to students, families, and schools. The gap between the level of achievement measured in both formative and summative assessments in the classroom and the collective measures of achievement reported by No Child Left Behind requires a policy-based solution. All citizens need educators and education funders to improve education. Policy designers and implementers can improve the relevancy of administered assessments, and ensure that the improvement of assessment outcomes is a standard expectation of professional educators.

Key Federal Policy Recommendations

1. **Achievement Through Technology and Innovation (ATTAIN) Act**
   Provide federal leadership to support states and districts regarding technology's role in school reform by passing the ATTAIN Act authorized at $1 billion. The ATTAIN Act would revamp and replace the current Title II-D of the Elementary and Secondary Education Act by building on its successes and focusing resources on those practices known to best leverage technology for educational improvement. The program works to

- ensure that through technology every student has access to individualized, rigorous, and relevant learning to meet the goals of NCLB and to prepare all students and America for the twenty-first century;
- increase ongoing, meaningful professional development around technology that leads to changes in teaching and curriculum and improves student academic achievement and technology literacy; and
- evaluate, build upon, and increase the use of research-based and innovative systemic school redesign that centers on the use of technology, leads to school improvement, and increases student achievement.

The ATTAIN Act would sustain this support for the federal investment in school improvement through technology and innovation.

2. **E-Rate**
   In order to strengthen the technology infrastructure of our schools, policymakers should increase funding for E-Rate, a program under the universal service fund that provides schools and libraries with discounts for telecommunications services, Internet access, internal connections, and maintenance of internal connections, based on the socioeconomic need of the school, to meet current and future high-speed broadband needs. At a minimum, federal policymakers should adjust the E-Rate pool of $2.25 billion for inflation.

3. **State longitudinal data systems**
   Federal policymakers must support the coherence of data systems among the state, district, and school levels. Policies should encourage states to align systems so they are able to drill down in the data to the student and teacher levels for the purpose of addressing teacher quality or individual instruction. Each state should redefine its role of "data compliance officer" to "data leader," and work to help stakeholders throughout the system use data to improve education at all levels.

As data leader, states should support districts in tying together their own data systems with formative assessment through the use of learning management systems, providing training to teachers on how their formative, interim, and summative assessments should be aligned with and contribute to the longitudinal data for student performance; and training educators on how to mine data for decisionmaking and changes to instruction and interventions.

States must also begin helping schools and districts address how relevant formative assessment and demographic data can "flow up" to the state to inform systemic changes in policies regarding school reform and student achievement.

Federal policy can support these activities by requiring that the federal data policy agenda—including funding—addresses these issues.

4. **Development of state assessment banks**
   The federal government should encourage states to create electronic assessment repositories that contain interim, formative, and summative assessments aligned with state standards. States should provide funding for master teachers to come together for the initial development and review of the items in the bank. This resource should be accessible to all teachers, students, administrators, and parents both within schools and remotely, and should allow for contributions from the field. Federal policy should fund a pilot program for the development of such banks and evaluate the impact of their usage on teaching and learning.

5. **Elementary and Secondary Education Act**
   In order to ensure that teachers and administers can effectively carry out assessments, districts should be encouraged to use their Title I and II funding to train teachers and administrators on using and analyzing data, administering quality and innovative assessments, and integrating technology into their classroom evaluations to support teaching and learning. For districts and schools found to be "in need of improvement," additional school improvement funding

should be used for this kind of professional development to help raise achievement for the targeted population(s).

6. **Higher Education Act**
   Funding through the Higher Education Act should go toward pre-service professional development programs that foster individualized instruction through the use of data. Federal funding should encourage the development of analytical skills around both data collection and analysis and the creation and use of effective assessments in teacher and administrator training programs.

The views expressed in this chapter are those of the authors and do not necessarily represent those of the Alliance for Excellent Education.

About the Authors

Erin Martin Gohl has worked at multiple levels of the educational system. Most recently, she was the director of education policy for Bernstein Strategy Group, a government relations organization focused on education policy issues. She has consulted with the New York City and Palm Beach County school districts; conducted school-, district-, and state-level evaluations of school choice programs in New Jersey; and assisted in the research, planning, and production of toolkits for the State Educational Technology Directors Association's National Leadership Institutes and the International Society for Technology in Education's National Education Computing Conference. Ms. Gohl has focused her academic research on issues of educational equity and access and the political dynamics involved in creating sustained education reform in urban school districts. She has also presented papers at various education-related conferences. Ms. Gohl received a bachelor's degree in American history from Princeton University and did her graduate work in politics and education at Teachers College, Columbia University.

Daniel Gohl is an accomplished educator who has experience in secondary school reform, the role of technology in education, STEM program design, and school leadership. He has served as a teacher at the secondary and college level in the fields of physics and astronomy, a high school principal, an information technology coordinator, and a director of secondary school reform. With experience in consulting with schools, districts, and state education agencies throughout the United States, he has developed expertise in collaborating with communities to address issues of achievement gaps between subpopulations, integrating academic and work skill

instruction, and ensuring meaningful learning in math, science, and technology. He has received a number of awards for teaching and school leadership, served on national committees on technology utilization and laboratory design for schools, and currently serves as president of Ontic Education LLC. He holds a master's degree in science education from the University of Texas at Austin and a bachelor's degree in physics from Vassar College.

Mary Ann Wolf is the executive director of the State Educational Technology Directors Association (SETDA). In this position, Dr. Wolf works with educational technology directors in all fifty states and the District of Columbia and works with policymakers to share models of how technology is critical to transforming our education system. SETDA contributed information and research for the development of the Achievement Through Technology and Innovation (ATTAIN) Act.

Dr. Wolf has led the Education Forum and has specifically contributed to SETDA's National Trends Reports on EETT; *Maximizing the Impact: The Pivotal Role of Technology in a 21st Century Education System*; SETDA's annual July issue with *THE Journal*; and the Class of 2020 Action Plan for Education. She taught fifth grade and worked for KPMG as a consultant for federally funded grant programs. She has a PhD in education from the University of Virginia, a master's in elementary education from the George Washington University, and a bachelor's in accounting and marketing from Georgetown University.

---

[1] D. Sparks, "Plugging Educators into Technology," *Results* (Alexandria, VA: National Staff Development Council, 1999), http://nsdc.org/library/publications/results/res2-99tech.cfm (accessed October 15, 2008).

ALLIANCE FOR
EXCELLENT EDUCATION

www.all4ed.org