The Gordon Commission
on the Future of Assessment in Education

Testing in a Global Future

Eva L. Baker
University of California, Los Angeles
Center for Research on Evaluation, Standards, and Student Testing (CRESST)

Introduction

This paper will approach the future of testing in a globalized context by addressing factors central to predicting it in the midterm (plus five years). Treated here are the current and short-term development of assessment options, the roles of international comparisons, demography, knowledge expansion, job changes, and of technological growth. Options to adapt learning to unpredictable contexts are offered in the discussion of assessing transfer and skills that underlie content learning, and social and personal development. Technology is given extensive treatment, as it will strongly continue to impact the expectations and options of learners. At the end of the paper, unanswered questions are offered to the reader to stimulate and clarify prospects, contexts and consequences of future assessments.

Some Context: Schooling and Testing in the U.S. and Beyond

Predicting the future is dangerous, but to launch such an analysis, understandings of the past and present are essential. The first part of this paper will briefly review the place of testing and assessment today, focused largely on education. The majority of children now enroll in schools dedicated education. Schools depend on formal systems of goals and standards, curricula in many subject matters, professional teachers, and buildings with desks and computers. The specific purposes of individual schools, their organization and ambience, and the groups of students served differ substantially; but they hold one thing in common. All formal schools assert authority and manage the majority of formal learning options available to students. The school has been the principal gatekeeper on content and arbiter of accomplishment. Giving and reporting external tests is one way to reinforce school authority. In addition, classroom authority historically has depended on teacher-made tests and quizzes. These were given to check up on and improve learning, and to assign grades to indicate levels of accomplishment of students.

In many countries, external sources exert far more authority than in the United States, for centralized ministries have long used testing to compare students, schools, and teachers. The United States is just now catching up. Assessments given once a year are now supplemented by external exams administered on an interim or formative basis.

Interim tests are intended to predict performance, apparently, assuming no effective instructional interventions upset predictions. Formative assessments, integrated with instruction, may be in traditional test formats or projects, papers, and performances. Their effective use requires teachers to develop high levels of content and pedagogical expertise so to understand reasons for students' errors and to determine the best strategies for giving feedback, and supporting adaptive instruction.

Describing Tests and Assessments

One way of classifying tests or assessments is to look at the purposes of assessments, the inferences, and use drawn from their results. Test findings may be informational, descriptive, relevant to research, or used to make decisions about individuals or institutions. There is a continuum of the degree of personal impact test results may have, for instance, in affecting an individual student (getting into college, or closing down his neighborhood school). Within this continuum, there is a raft of traditional test purposes involving achievement, including selection and admission to programs, certification of individuals (students with diplomas, doctors with licenses, etc.), classification for purposes of assignment to courses, jobs, or opportunities, diagnosis of weaknesses and strengths during instruction, and promotion in some schools and in the commercial sector. In addition to academic achievement goals, tests are beginning to be used to measure cognitive skills, interpersonal abilities, and metacognitive or internal personal states focused presumably on those components, which can be systematically learned. Research continues, for example, on the measurement of motivation or engagement, as explanatory variables for levels of performance. For years, tests followed an aptitude or predictive model, even when they served a nominal achievement purpose. A sea-change recommending the way tests be designed and used was articulated by Glaser (1963) after more than 20 years of experience by the training community in developing clear tests that could represent estimates of learning domains, rather than comparisons with other students on a more general construct. The implications of Glaser's work will continue to have impact in almost any vision of future testing. It addressed three key points: (1) circumscribed domains of knowledge and skills, (2) the setting of levels denoting proficiency, and (3)

the reporting of results in the light of these levels rather than referenced to the achievement of other examinees. Not all of the intent of Glaser's work has been manifest, but in the last 30 years, test results have come to be seen as the principal indicator of educational effectiveness, and a single test used for a variety of purposes, often reported in terms of frequencies of students attaining specified performance levels.

The *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999), relying on research by Messick (1989) and Cronbach (1988) argue that the validity of inferences from test results depends upon the purpose the results are intended to serve. Purpose extension or "creep" results in more complex validation procedures, for each purpose should have corresponding evidence to support it (see chapter by Ho, in this volume). In personnel evaluation involving teachers, students test performance sets up an obvious conflict of interests. If these tests are to determine teachers' job continuation, promotion, and financial incentives, teachers have both a legitimate interest in the choices of tests and a dangerously irregular line to walk to preserve the integrity of their relationships with students and their profession. Schools have, in the past attempted to manipulate performance results of students as well, most obviously by suggesting certain low-performing students stay at home. As a consequence a spate of requirements in No Child Left Behind (2002) intended to assure that the correct number and subgroups of students were tested. Accountability in the United States remains the driving force for testing, implying but overshadowing the importance of measuring learning in schools.

Accountability. The worldwide expansion of educational accountability is associated with a neo-liberal view of education. This perspective, in brief, holds that deregulation, open markets, and reduced government intervention will result in stronger educational enterprises. Consider the introduction of charter schools, voucher experiments, and *laissez-faire* approaches to teacher development and induction. Open markets nominally assume a fair distribution of consumer knowledge and wherewithal, a dubious principle in the United States (and other countries) with grossly inequitable wealth and educational attainment levels. One's vision of future economies may or may not see an equalization or a greater gap in wealth in certain economies.

The policy frame (or theory of action; Argyris & Schon, 1978) for accountability has four connected elements, extrapolated from Ralph Tyler (1949). They begin with the

idea that the tests represent larger intentions and serve the purpose of improving the quality of education and of students' life expectations, i.e., as codified in written standards. Second, attention to testing is justified by the logic of communicating desired ends (operationalized on examinations) to teachers, parents, policymakers, and the public. Third, communication of standards or goals presupposes the use of practices appropriate to improve the learning of desired ends. This assumption is based on a dependent logic that the match of instructional options to communicated goals is understood by teachers, that good ways exist to improve learning and can be readily used by teachers, that such approaches are equally accessible to teachers and students, and, finally, that both teachers and students are willing to expend effort to attain the goals by legitimate means. Incentives and sanctions are put in place to affect effort. The fourth element, improvement, is managed by the first three, and may be reflected in changes within instruction, for individual students, or for elements of the system as a whole. This last set of improvement activities rests on bureaucratic, political, financial, and equity interests, often at odds with one another.

Evidence so far is not encouraging about the success of any of the core elements of this theory of action. To take point three above as illustrative, schools have devoted long periods (a month or two) to practice test activities prior to the test administration window. Second, cheating stories are found and many more are not surfaced. The commonplace use of smartphone cameras has ramped up the students' rather than the school and teachers' use of inappropriate means to raise test scores. As technology extends its reach and flexibility, users will find ways to work around the system, and reduce the degree to which the test results have meaning.

Nonetheless, the common emphasis on assessment and accountability, in the United States and elsewhere, continues. It is driven by obvious substantial differences in the educational quality of schools, instruction, and student performance. Many reasons are given for these differences. They include those related to student background (such as language or dialect in the home), financial constraints, for families or schools, uncertain student motivation and effort, and the need to improve the quality of teaching in certain subject fields. Accountability is a relatively inexpensive (at least in financial terms) reform.

Testing Now and in the Next Three Years

To clarify the meaning of terms, "testing" usually entails formal, uniform methods of observing student performance relevant to ability or a domain. Assessments or tests (used interchangeably here) consist of samples of items or tasks intended to estimate larger learning domains with quality, efficiency, and fairness. The major content for a test may be derived from the identification of a construct, like mathematics ability or an explicit domain, such as calculus or both. The niceties of the logical relationship of the test to content or cognitive elements it intends to measure, its task or item-development processes, its reporting metrics and evidence of its validity, utility, and fairness for various student subgroups, are rarely of interest to more than professionals in the relevant fields. To the average person, a discussion of tests begins and ends with their superficial features, largely the format the student confronts. Is the test multiple-choice? Is the test open-ended? A recent United States president, riffing on Gertrude Stein (1922), said, more or less, that a reading test is a reading test is a reading test, thus discounting test purpose, utility, or evidence of quality. In the United States, many tests use multiple-choice formats, for they are optimized for mass testing and use distributional statistics to infer elements of technical quality, such as reliability. Multiple-choice tests may measure significant content and complex thought processes. They have a reputation as devices focused on retained information, facts often learned through drill and practice prior to the testing period. Test items can systematically be chosen from item pools and then arranged into different configurations of test booklets. Answer sheets can be efficiently scored from keys used by optical imaging methods.

Other formats coexist, but less frequently. Since the advent of externally administered tests in the United States, "performance-based" test formats periodically recur as the new, "next best" thing. These assessments mirror some aspects of teacher-made tests before the advent of widespread external measures. They consist of essays or extended projects connected to local teaching and learning goals (November, 2009). They have reappeared about once a decade since the 1970s. They are here, once again, with the usual claim of anticipated success, this time attributed to technology platforms designed

for essays, simulations, and games. Performance tests use student constructed responses generated from systematic or idiosyncratic directions or prompts, or questions.

Included in the genre are, for instance, student-led research projects in a range of content, multidisciplinary problem solving, finding and using multiple sources, complex decision making, or written analyses of text, film, or art products. They may also include performance, such as a speech, playing a sport, or acting in a play. Performance-based tasks require more steps by a student to achieve completion and, therefore, need longer administration time. Because of required intense or at least prolonged student engagement, they are memorable and not as easily forgotten by students as a pound of multiple-choice items. Therefore, their tasks can be easily shared among students, and new assessments need to be developed continuously. How such measures can be developed in other than expensive, handcrafted ways will be discussed in a later section of this paper. Technical, political, and values divide constituencies on whether tests should have uniform "right" answers, whether processes should be prescribed, and whether students' reasoning and other independent thinking should be evaluated as well.

Cost and quality habitually sink performance assessment use in large-scale administration. Improvements to date related to technology range from relatively trivial to ambitious (but not yet achieved). Some approaches simply train raters and distribute papers via computer for hand scoring. Renewed attention has been given recently (Chung & Baker, 2003; MacArthur Foundation, 2011) to computerized scoring of open-ended performance. Some approaches have been advocated for decades (Landauer, 1998; Page, 1968), both for essays and for analyses of graphical products (Chung, Baker, Brill, Sinha, Saadat, & Bewley, 2006; Mousavi, Kerr, & Iseli, in process). Frequently, methods depend upon syntactic and statistical properties of the essay (such as sentence variability and length), all of which can be scored relatively easily and also manipulated by wily students to achieve relatively high scores from. Often, such statistical approaches use regression analyses to map to a sample of hand-scored papers to establish "validity." However, these regression models require teachers to score significant numbers of tasks, with their scores ascribed in the "gold standard."

Task design of essays and open-ended products remains, to a large degree, idiosyncratic to teams of task authors, and stronger approaches to the automated

assessment of *meaning*—using improved natural language processes—will depend as much on the clarity and systematic design of the tasks given to students and the degree to which elements in them have been successfully taught, as on technological advances applied in scoring them. The ultimate goal is a good computational model of task design related to automated, open-ended scoring; one that relies on the meaning of the essay and its quality relationship to the domain of interest. This problem will be solved in the next three years or so. Obviously to keep costs down, teacher or human rater involvement should be minimized, although it is important during the task and scoring development phase. The recurring dependence on human scorers will also be driven by the need for constituency credibility during a transition period to automated scoring. Thus, even with new developments, growing accountability demands and shrinking educational budgets will force the reliance on multiple-choice and other selected response formats. They are applied in computer-adaptive tests (CAT) that target test performance at each student's level of achievement. CAT algorithms result in shortened testing time, but they depend on a multiple-choice base for the most part. CAT has been used for decades and with greater educational scope that may attract more appropriate design and analyses of their properties and meaning (Bjorner, Chang, Thissen, & Reeve, 2007; Gibbons, et al., 2008; Reeve, et al., 2007; Wainer, et al., 2000).

Testing What?

Starting with vague models, such as "math ability" through an *über* behavioral period of micro objectives, tests and assessments in the last twenty years have been connected to middling clear statements of educational intentions, called curriculum goals or standards. These have been written in common language and have been subject to substantial variations in interpretation. For a period, at the beginning of the 21st century, U.S. policy regressed to content goals focused on mathematics and literacy only.

Goals and standards had moved in the 1980s from the province of school districts to state control. Now the process takes a new turn and reaches its natural conclusion, moving from goals chosen by classroom teachers, then school districts, then to discrete and variable standards created by each of the 50 states. In the last three years, the majority of states (now numbering 45 and 3 territories; http://www.corestandards.org/in-

the-states) have joined together under the auspices of the Council of Chief State School Officers and the National Governors Association. This collaborative effort began with the development of standards in mathematics and literacy, the "Common Core State Standards" (Standards). State adoption, spurred by federal incentives and waivers of regulation, has resulted in a national (not federal) set of goals. State "volunteerism" veers around the conflict between federal and state roles in education. States adopting the Standards have agreed to use assessments managed by consortia (Partnership for Assessment of Readiness for College and Careers [PARCC] and Smarter Balanced Assessment Consortium [SBAC]) to measure progress in English-language Arts and mathematics. The Common Core Standards for Science are also being vetted and will result in assessments also developed by consortia. The consortia of states rely on commercial publishers to generate, according to their specifications, assessment plans, item types, scoring rubrics, administration, and analysis and reporting components. The development of the initial set of measures is funded through federal support. Much is made about how such new standards and assessments will at last improve the relatively stagnant performance levels in the United States. Let us hope so.

In the past, test content and format have strongly guided what is taught in schools, even though those tests were usually administered for accountability purposes once a year. In the new plans, there are some changes. Approaches to graphical mapping of content (to understand the deeper, nontextual relationships among standards) have been undertaken both on a research level (Iseli, 2011) and in commercial, online follow-the-dots displays. Purposes of testing have nominally been expanded to include formative assessment, or the use of assessments during the instruction to identify students' strengths and weaknesses, and to help the teacher develop optimal plans for each student. Recent investments in **s**tandards and assessments accountability will remain in the short run the major strategy for educational reform in the United States —applauded by some for its outcome focus and decried by others, noted earlier, as inadequate reform on the cheap.

On an international note, the use of standards and assessments was partially spurred by comparisons with competing nations. With this segue, let's review international comparisons and their direct and indirect effects on assessment plans, beginning the discussion of the "global" aspect of the chapter title.

International Comparisons of Educational Achievement and Processes

Although comparative education has long existed as an academic specialty, its study often depended on descriptive qualitative studies or policy analyses. In this section I will consider two approaches addressing testing, one of more than 50 years of effort and the other most active in the last decade: the work of the international community to compare its educational achievement on common measures and indicators.

International Association for the Evaluation of Educational Achievement (IEA)

The pioneering collaboration of the International Association for the Evaluation of Educational Achievement (IEA) began in the 1960s, led by a small group of international scholars who banded together to learn about each other's approaches to testing, learning, and achievement. The academic purposes of these early IEA studies were to describe differences in student attainment by country and to explain, if possible, achievement variations related to curricula, teaching practices, student background, and cultural contexts. Early on, studies were planned and fielded to enact these goals, including the First International Mathematics Study (FIMS, 1964; Bloom, 1956). A more ambitious endeavor involved comparing countries on six different subject matters. The Six Subject Survey consisted of culturally sensitive measures, relevant to the curricula of participating countries, on the topics of reading comprehension, literature, civic education, French as a foreign language, English as a Foreign Language, and science. As a time marker, data collection for the Six Subject Survey occurred in 1970–1971. The study designers confronted differences in meaning of common educational and testing terms, traditions of test taking that varied across culture, as well as some of the nuances of appropriate language translation processes. Successive IEA studies were conducted on various topics. For example, the Written Composition Study begun in the late 1970s was groundbreaking in its use of open-ended responses, culturally-sensitive prompts to obtain writing samples, and the design and use of common scoring rubrics to compare performance of cross-national performances on different writing genres. To place this activity on the timeline, data for the Written Composition Study was collected from

1984–85. A full record of IEA activities, including key scholarly documents, study designs and findings, can be found on its website (http://www.iea.nl/home.html).

From its beginning, IEA was largely a scholarly venture, with renowned academics from the United States, Sweden, Germany, and the United Kingdom at first leading the design, implementation, analysis, and reporting of studies. As noted, their overriding purpose was discovering new knowledge from these investigations. Academic foci created its own series of problems. For example, in the early days of IEA, differential funding and intellectual resources were available in participating countries, with each national steering committee responsible for securing whatever funding it could. Even in the United States, until the mid 1980s, the U.S. national steering committee was often on its own to find funding. Foundations were usually solicited and responded providing funds for the United States and, sometimes, government chipped in. The lack of centralized funding placed study decision making in the hands of the senior scholars from each country, and limited the ability of early IEA leaders to enforce expectations regarding sampling, uniform administration, data quality, and analysis. As a consequence, the numbers of participating countries and aspects of technical quality varied by study and by economic cycles as well. In the United States, in the 1980s, efforts were made to secure full U.S. funding from the government, and the Second International Mathematics Study (SIMS) found a financial home in the Center for Education Statistics. This relatively secure funding source raised standards for data quality, analysis and reporting, and studies investigating explanations of performance differences. It also opened the door a crack to government intervention. In SIMS, 26 countries participated in data collection, and its reports focused more strongly on ranking students' performance by country. The Third International Mathematics and Science Study, renamed "Trends in International Mathematics and Science Study" to preserve the TIMSS acronym, is in its fifth four-year cycle, with a next study to take place in 2015. Its reincarnations also focus on country-by-country rank orders, but it has persisted in efforts to describe curriculum and teacher preparation differences as explanations for differences (Schmidt, McKnight, Houang, Wang, Wiley, & Cogan, 2001). Subsequently, IEA has supported the Progress in International Reading Literacy Study (PIRLS, http://www.iea.nl/pirls_2001.html). IEA still regards itself as an international cooperative with, notably, its locus of primary

leadership vested in an international steering committee for each study, supported by a national research coordinator for participating countries. These individuals are employed by educational ministries, universities, or other institutions. IEA has on its docket a number of future studies.

How one entered the inner circle of IEA changed over the years. At first, IEA operated as an old boys club, and entry was controlled carefully but informally. Illustrious members of this club included well-known researchers Bloom, Noah, Postlethwaite, Keeves, and Coleman, among others. Later, as funding sources changed, access to the inner sanctum of research management was more transparently managed. The TIMSS study had a strong effect on historic U.S. notions of its academic "exceptionality." Our rank was not at the top, and in successive studies, the performance of U.S. students has continued to slide.

OECD and PISA

The continued success of IEA drew the attention of the members of the Organisation for Economic Co-operation and Development (OECD). Venturing into education because of its key role in affecting the economic status of member countries, OECD began with explorations of educational indicators, studies focused at the outset on describing the comparative statistics of member education systems (e.g., how many schools, distributions of students, and teacher preparation). Many of these findings were clearly summarized in the "Education at a Glance" series. When OECD moved toward the creation of an educational directorate responsible for a range of descriptive investigations, a stronger investment was made by member countries to achieve high quality, cross-national comparisons. As OECD is a membership group of governments, its studies displayed an early and persistent interest in the policy implications of findings, in addition to its willingness to support a large number of provocative scholarly analyses.

OECD created its Programme for International Student Assessment (PISA) in the mid-1990s, with its first achievement surveys administered in 2000. Successive studies ensued, usually focused on students at age 15 (the last common year of compulsory education). Studies with rotating topics (e.g., science, mathematics, and reading) used increasingly complex background characteristics, designs, and analytical approaches.

Studies were fielded, first on a three-year cycle, and now on a two-year interval with generally increasing numbers of participating countries, drawn from both OECD members and non-OECD countries.

The leadership of the PISA effort has been vested in an OECD senior director and team, with international consultants employed at the design, analyses, or reporting stages. In contrast to IEA steering committee, the PISA governing board determines priorities for topics of policy interest. The PISA governing board is made up of individuals appointed by the OCED member education departments and ministries. Its membership therefore shifts to reflect the policies and politics of those in power, with membership changing as a function of new governments. In recent years, for example, U.S. support of PISA altered depending upon the leadership of the National Center for Education Statistics (NCES). Current membership of the PISA Governing Board is split between governmental officials and delegated university representatives. PISA contracts out the design of surveys and the development of items through a call for proposals, and historically a relatively small number of contractors have been used.

At the outset PISA focused reports on the ranking of country performance with respect to achievement and other variables. PISA produces a prodigious number of analyses and crystal clear charts available on its website (http://www.oecd.org/pisa) where interested individuals can interact with its databases. While IEA continues an admirable analysis of cross-national curricula, PISA (with its governmental linkage) has expanded its interest beyond achievement to explanations and measures of attitude, teacher practices, among other variables. But its ties to governmental membership in OECD continue to link to policy. Chief among these is the horse race between countries, where comparative educational vitality and growth is inferred from rank order.

The link of PISA reporting and U.S. testing activities is fairly straightforward. For example, the U.S. was a strong performer in early PISA surveys, at the top for high school completion, for instance. Over the last few cycles, the United States has maintained its numbers, but has dropped to the middle of the pack or below the OECD average because of the rapid improvement of other countries (http://www.oecd.org/pisa/46643496.pdf). It is true, of course, that many OECD countries have small numbers of students (the Finnish population, for instance, easily fits

into a suburb of Los Angeles). However, it is clear that PISA has undermined reliance on the usual explanations for poor performance (or in the OECD case, flat performance). Targeting immigration levels and poverty have lost some cogency. Other English-speaking countries, albeit smaller in population, with similar high rates of immigration like Australia and Canada, do far better on equity indicators (small differences in performance by immigration, language, or economic status) and on overall performance profiles. The static nature of U.S. academic performance has been sadly reified in the National Assessment of Educational Progress (NAEP) (http://nationsreportcard.gov/ltt_2008/ltt0001.asp). Data are so remarkably stable, that small changes are greeted with overwrought excitement. Achievement differences persist among subgroups of students, for instance, on our NAEP and on our state tests (see, for example, *Los Angeles Times*, September 1, 2012 report on equity results for California), we find different attainment of Black and Hispanic students compared with White and Asian students. This disparity, perhaps chief among many achievement-related problems, has yet to be resolved by the U.S. educational system, although apparently solved to some extent by other countries.

What have been the most obvious consequences of international comparisons? First, there have been a great number of study tours to highly ranked countries in order to discover their "secrets." The focus of these visits tends to concentrate on types of measures, autonomy, teacher quality, and induction (see Darling-Hammond. 2010; Sahlberg, 2011). Second, international rankings provide only a weak rationalization for the roiling U.S. policy development, which pivots on more political grounds. PISA findings have massively motivated many OECD countries. For instance, the "PISA shock," occurred when Germany discovered in 2001 it fell below the average of student achievement in OECD countries, generating serious and systematic reforms (http://www.germantimes.com). In Japan, its recent showing ranking sixth in math and 10th in science prompted the declaration that its last 10-year experiment in "pressure-free" education was a great failure, and tougher curricula were needed.

Throughout the world, there is puzzlement about the lack of U.S. response to its place in international performance, 35th in science and 29th in mathematics (see the

fuller treatment; http://blog.commoncore.org/2010/09/08/japans-pisa-shock/). This excerpt provides a German perspective from an earlier PISA study.

In the Unites States, however, the three cycles of the PISA study and their results have gone largely unnoticed. The general public and the media did not pay attention to this study and its results — despite the fact that the United Staes performed as badly as Germany, being ranked below the average of the participating countries. In fact, the United States scored worse in the 2006 study than it had in the previous two studies. In the latest PISA study, for example, the United States ranked only 24th in mathematics among the thirty OECD member states. But no one seems to be concerned: out of eight major U.S. daily and weekly newspapers, only eight articles over the period of 2001 to 2008 actually covered PISA at all — and of these eight articles only three deal with the poor results of U.S. students. During the same time period, the German daily newspaper *Süddeutsche Zeitung*, for example, published 253 articles on this topic. Why did the PISA study not have the same effects in the United States as it did in Germany? Or to put it differently: why is there no PISA shock in the U.S.? (http://www.aicgs.org/publication/why-is-there-no-pisa-shock-in-the-u-s-a-comparison-of-german-and-american-education-policy/). (for more details on the OECD's role in education, see Kerstin Martens [2007]: "How to Become an Influential Actor – the 'Comparative Turn' in OECD Education Policy," in: Kerstin Martens, Alessandra Rusconi, & Kathrin Leuze [Eds.], *New Arenas of Education Governance – The Impact of International Organizations and Markets on Education Policy Making*, Palgrave Macmillan, 40-56.)

The United States lack of response to PISA has been given at least one benign interpretation — that the United States began its reforms earlier, with *A Nation at Risk* (National Commission on Excellence in Education, 1983) and thus knew it was behind, tracking its NAEP test more carefully than PISA. That we are behind is now an indisputable fact. U.S. credibility is undermined internationally and its educational stasis presages to many the future of shrinking U.S. vision, power, and economic innovation. If performance were important, one would expect other than half-measures or inaction to help its poorly performing school systems. The U.S. response, noted earlier, has been to propagate charter schools, undermine teacher security, reduce teaching qualifications,

while heightening accountability and, in many states, cutting education budgets and services, as well.

Yet a recent quote once again focuses U.S. bets to improve its education system largely on assessment:

> "If America is to have a public school system second to none, each state needs a first-rate assessment system to measure progress, guide instruction, and prepare students for college and careers." (http://www.ed.gov/news/speeches/beyond-bubble-tests-next-generation-assessments-secretary-arne-duncans-remarks-state-l)

Unfortunately, lessons about equity, public and political support for education have not been learned in the United States as well as it has by other countries. PISA claimed to focus much of its assessment on cross-curricular skills. In the last 20 years there has been growing awareness of the potential of learning complex skills, a topic of interest in the discussion of future efforts.

Framing the Global Context of the Future

As is evident, governments vary in the ways they read and interpret data, assert authority, use language, and implement, or not, democratic institutions. Different governments have varying rules about the roles that spiritual or religious affiliation play in education, tolerance for diversity, and expectation for high levels of student performance, all of which have serious day-to-day implications for educational practices and their assessment. Let us start with some basics. What will the student population look like in the next 10 years?

Demography

Change in demography is always a fact of life, and is reflected in cycles both within and among countries. How does demography affect assessment? For example, in Japan, Taiwan, and elsewhere in the developed and developing world, the birth rate decline is an incontrovertible fact. In Japan and Taiwan, fewer students have begun to be reflected in a concurrent surplus of places in existing universities and colleges. The assessment implications of this situation are straightforward: the percentage of students

admitted to Japanese universities without taking heretofore high status examinations has soared (Arai, 2011). In Japan, the relaxation of standards in higher education conflicts with the system's interest in raising standards of performance in pre-collegiate settings stimulated by earlier described PISA shock. Raising standards for pre-collegiate education is obviously at odds with relaxed admission standards.

On the other hand, oversupply of educated individuals in India, after years of frustration, is showing up in advancing gross national product for services as well as products. India, Brazil, and China have rapidly growing economies (as well as persisting gaps in wealth among its populations). Yet, even with policy-driven, low birth rates, China, because of its sheer numbers, produces numbers of students qualified for higher education that currently exceeds national high education capacity. Whether their birthrate problems will ever catch up to them, and alter their educational opportunities depends in part on their government's action. For example, the city of Shanghai, of more than 20 million (http://www.oecd.org/countries/hongkongchina/ 46581016.pdf, p. 4), has an aging population much like Japan and a birthrate of 1.3 (http://www.foxnews.com/world/2012/05/11/lack-babies-could-mean-extinction-japanese-people/; Li, 2010). Japan reports its birthrate 1.35. Other Asian countries cite similarly low birthrates (e.g., Hong Kong; 1.07, Singapore; .9, and South Korea; .9). Similarly low rates are found across Europe (e.g., Germany, 1.33; Spain, 1.04; Russia, 1.09). Countries interested in raising their birthrates are taking into account the recognized limits of sustainable worldwide growth. Some have looked within their borders for explanations of dropping birthrates (in those societies without government regulations or targets). For example, in Taiwan, where incentives were given for larger families, the birthrate continued to fall. Among the hypotheses are the advent of more women in the workplace, the deferral of marriage, shorter, unaided female fertility intervals, the cost of raising children, the desire for possessions (a technology-driven issue), and women's wishes to lead independent lives (https://www.cia.gov/library/publications/the-world-factbook/fields/2054.html).

These figures, linked to lower mortality rates and longer life spans for the aging, suggest that, depending upon vitality of national economies, older individuals will need to work longer in order to support their extended life spans. Evidence in the United States

suggests that older workers are already competing with young people for desirable jobs. If birth rates in some countries continue to drop, the ability to replace individuals in certain emerging work sectors will diminish as well. These population changes presage new strategies for multi-national corporations seeking workforce within and across national boundaries. The outsourcing of manufacturing to "cheap" labor is a short-term solution, although likely to be rotated through poorer Middle East and southern hemisphere nations, whose birthrates are in multiples of two, three, and four of developed nations. The tighter linkages among workforces emphasize the importance of developing and meeting international educational standards in order to sustain international economies. This is one small reason that even relatively weak articulated goals, such as measured on PISA, are likely correlated with future goals of a deliberately international nature. Undoubtedly, any country's educational goals will have a large dose of localism and tradition. But if economics is the driver, it is likely that more attention will be paid to international expectations.

A second potential issue derived from the global shift in birthrates is the likely need to resort to the testing of older workers first, to assure that specific skills and general cognitive functions remain at a sufficient level to perform tasks, and, second, to provide a nominally fair procedure to help employers choose among the very young, newly displaced, and older workers.

Increasing attention is being given by some developed nations to help improve education in developing countries. For instance, the successful economy of Korea is helping one of the most needy African nations (Burkina Faso; http://legatum.mit.edu/). One potential self-serving reason for engaging in educational improvement of poor countries is to create a cadre of well-educated workers who can fill in for countries with low birthrates. Another obvious option is for countries to open up immigration to groups with high birthrates, for example individuals from strongly Islamic and Catholic nations. Such decisions, at least in the present, seem to come with social and political consequences.

Bum Mo Chung, an illustrious Korean scholar, has written about whether education systems can transfer to different country environments, and how the interaction of levels of economic development, birthrate, authoritarianism in government, types of

schooling available, and per capita income relates to educational change (2010). His data are lodged, for the most part, in the 20th century, but imply workers can be trained by methods outside of the countries' traditional educational approaches. Clearly, no developing nation will go through the same historical sequence of development that Chung describes, particularly in a time of the rapid changes and growing technology. To date, it seems whatever their development level, emerging economies will begin their education systems and policies under an umbrella of test-based accountability (Baker, 2010). It is likely that they will adopt commercially developed measures, with an eye to PISA or similar comparisons to provide an external indicator, along with political stability, of their readiness to contribute to business growth with the investment of skilled people, that is, individuals with the proficiencies required for economic participation.

Further modifications in careers and work will be prompted by changes other than demography and incentive for development. Immigration affects jobs at all levels of activity, from the most routine to those demanding the highest level of creativity. The ratio of youth to the aged will require new configurations of work or partnerships on a regional or international level. Demography presages vast changes in high education selectivity, border crossings, costs, and quality. These factors interact with the uncertainty attached to specific knowledge growth and to job evolution.

Changes in Knowledge and Career Options

A second factor with global reach considers the explosion of knowledge and its implications for predictable job markets. Many modern careers create and depend upon new knowledge. Now cliché, it is true that knowledge is expanding at an unprecedented rate. In school, or at least in the university, we are well aware that most content is undergoing revision and expansion, especially in the sciences and engineering fields. New strategies and applications of mathematics are present. There are few areas that are particularly stable, as literacy, for instance, now includes verbal, graphical, and interactive components, attributes which depend on increasing skills in searching, classifying, authenticating, and applying knowledge.

As the academic and applied knowledge changes they affect the predictability of jobs and career options. In many areas, there are new jobs that no one could have

imagined even five years ago. We are told that in five years, half of the future top ten jobs for adults have not yet been invented. Job change is caused by knowledge expansion, by the globalization of businesses, technological advances, and the interdependencies of technical and service sectors. While there will always be a combination of global and local components in jobs and careers, as in education, common expectations are on the horizon and will affect school or informal learning requirements.

Furthermore, as one progresses through a life of work, we observe that individuals now have many successive jobs or careers, rather than just one. (My son has had 12 different employers [he works in Internet games and systems] since he finished the university. I have had three jobs, one lasting many decades.) Some change is due to economic instability (for instance in the dot com world), which is sure to reach all domains, affecting recently even the tranquility of university life. Yet, the change is more pervasive. Even those who continue working in one business or for one employer are expected to adapt to new task expectations. Twenty years ago, assistants typed my work from handwritten yellow pads and receptionists answered the phone before transferring it to a suite of offices. Now most people draft their own texts, answer their own mail and phones, and manage their own schedules. I have had to learn much about computers, graphics, and software that earlier would have been the sole province of experts. People will be forced to learn throughout their careers. Lifelong learning will not be an old saw intended to spur the reading of Oprah's latest book recommendation. It is the reality now.

If job and economic growth will demand cross-national skills, who might be in charge? Who will manage assessments that must cross national boundaries and determine fitness for work? In Europe, should the Eurozone survive, there may be a further consolidation of common goals in response to more open higher-education options. But another option is the preemption of authority of multinational corporations on assessment. By this I do not mean commercial test publishers with international reach, although they may be contracted to do some of the less proprietary work. Now we see foundations pushing assessments with international reach. Could multinational employers substitute their assessment authority for schools? Will students have choice and be able to opt among providers and certifiers? Will there be multiple pathways toward careers through a different form and locus of assessment? If multinational corporations remain or

grow in significance, won't their demands supersede what national educational systems expect and measure? Will we have incentives, through assessment for more "intelligent" dropouts on the order of prodigy Steve Jobs, Bill Gates, prodigy app makers and innovators? If national entities wish to preserve their traditional educational functions (other than childcare), how will they cope with the start-ups and titans of the global economy, surging knowledge, and serious lifelong learning? They need to find some agile way to allow their assessments to keep up. The relatively simple 21st century analyses of skills offered by some as a substitute for content knowledge is not enough.

Preparing the Content for a Future of Unpredictable Expectations

There are at least three rational approaches to dealing with the unpredictability of job requirements, global contexts, and emerging relevant learning required for future success. The first is for educational systems to become more agile both operationally and politically. I have no useful comment on this alternative. The third is to focus on outcomes relevant to existing standards and knowledge, but always, always, include in tests or assessments tasks that call for transfer, or the application of learning to new, unexpected tasks. The second is to focus on a set of more pervasive skills that could be embedded in unforeseen different contexts, and changing subject matter, directed toward new applications.

Transfer and Generalization. The second option transfer and generalization is based on the premise that if people are taught one thing, they may very well be able to generalize their learning to new applications. Early versions of transfer were relatively simplistic, learn "x" and measure "y" initially investigated in verbal learning (Gibson, 1940). Much research has investigated systematic application of the concepts of generalization and differentiation to verbal learning areas (Bereiter, 1995; Dancereau, 1995; Pressley, 1995).

Refined analyses of transfer have been developed addressing the extent to which transfer is affected by specific or general approaches to learning, whether it depends on the induction of schema or patterns that can be applied to unencountered problems (Gick, & Holyoak, 1987), whether it is best conceived as others explore content or structure methods by which it might be induced, e.g., analogical reasoning (Holyoak & Koh,

1987). Sweller, van Merrienboer, and Paas (1998) also have expanded earlier investigations of the role of memory, especially working memory, to facilitate the application of learning to frames of unencountered problems. In short, the research bases in transfer are wide and deep, and alternative methods for dealing with degrees of similarity from initial learning content and structure have been explored (Mayer, 2008).

If transfer (and the generalization it produces) is learnable, that is, there are methods to deal with unencountered problems, that is, the future. The first simple policy step is that transfer must be regularly included as part of tests or assessments used to measure learning. Studies of development and validity of transfer measures are ongoing, in particular in game research (Baker, Chung, & Delacruz, 2012a) and may have implications for the degree to which we are able to move from present practiced assessments to more powerful, generalized learning outcomes.

"21st Century" Skills or Learning the Core of the Core Standards

The second method of dealing with unpredictable changes in content and context is to investigate the use of cognitive, interpersonal, and intrapersonal skills (National Research Council, 2012). Heretofore this general domain has been variously named skills for cognitive readiness (Baker, in press; Baker, O'Neil, & Linn, 1993; Fletcher, 2004; Fletcher, 2007; Linn, Baker, & Dunbar, 1991; O'Neil, Perez, & Baker, in press), and most recently in the report of the NRC, deeper learning (2012). While groups such as the Partnership for 21st Century Skills (http://www.p21.org/) as well as others each have their favorite list of skills. O'Neil and Lang (in press) categorized them in educational functions, such as those easily taught vs. those difficult to teach. In other words he differentiated "skills" that were stable with trait-like predispositions such as creativity (hard-to-teach) and contrasted them with cognitive or social components that adapted to and embedded in various content, such as complex, multi-stepped problem solving or teamwork and could be specifically learned.

As with transfer and generalization, there is a remarkable history of research on skill learning, and far less on best ways to measure them as they can be embedded in new content.

One way to divide the skill pie is by the type of interaction expected to be demonstrated (improved task performance, better interactions with people, or routinized self-management of learning and behavior). Another cut that is useful for conceiving of measures reverts to the logic of more general constructs that have components which contribute and interact with one another. For instance, what if we were to posit two major constructs to encompass skill learning, one cognition and the other motivation? Under the cognitive construct, we might have related supporting skills such as situation awareness, problem solving, and decision making. Under motivation as an overarching construct, one could place, learnable and measurable subskills such as engagement, effort, self-efficacy. Both constructs require an executive set of skills or construct tied to a more general construct, self-regulation, metacognitive, or intra-personal learning (Rueda, 2011). These executive functions are internal to the individual. Teachable components of this executive construct include planning, attention, feedback use, control of fear, emotion, or anxiety, as examples. There is no doubt, however, that they will interact with temperament and preference as well as experience. Taken together the three constructs (cognition, motivation, and self regulation) bound a great deal of what can be validly measured for an individual, and applied to the collaborative, interpersonal situations.

At this point, it may not matter which particular set or organization of skills is embraced; emerging requirements will make priorities clearer. But it does matter that we understand what we mean when we intend to foster their skill learning and to assess them. For instance, in the area of problem solving, there are different positions about the situated nature of learning and whether such skills, when conceived as serious intellective processes, can transfer to other settings. Let's explore that example. Start with elements of problem solving. All must be embedded in a particular initialized state, a set of content and contexts, and invoke relevant prior knowledge, including patterns or schema. Under these conditions, some general structures of problem solving may be formed as a set of production rules, or if-then propositions. Is the student given the problem, or does she have to find it? How many distractors are in the setting? Is the problem occluded, conceptually or visually? Does the student have to represent the problem or is representation provided? Are alternative representations acceptable? More questions

pertain to the choice of solution strategies, their evaluation during their implementation, and whether the student needs to reconsider decisions, either in a backward chaining manner or is required to begin anew. All serious problem solving needs to be taught within contexts and in content requiring various levels of prior knowledge. In this argument, we would want elements of the problem solving process and structure to be embedded in different situations and different content, again looking toward transfer and generalization. If we take seriously the unpredictability of the type and speed of change in the world, we very much need to focus on learnable skills. These skills may generalize, in part, or be combined with other skills, or undergird new ways of conceiving or attacking a particular problem in particular contexts; in other words, they need to give students a leg up, the wherewithal to be successful in unfamiliar tasks. But we must not invoke the learning of such skills at the expense of high-quality expectations of how they will be integrated with challenging subject matter. It is an error to think these skills only apply to settings where there is no right answer. Answers may be uniform, with varying pathways to them, they may be variable or expressed in different ways, or they may be judged to satisfy requirements and solutions will take many different forms.

Reprise

To this point, we have considered as influences on the future of testing, (1) the state of testing now and in the short term, (2) international comparisons and what they portend for American students and our place in the academic and competitive world, (3) the role of demography, changing knowledge, and career loci expectations, and (4) a focus on new outcomes, i.e., transfer and generalization and the acquisition of learned skills as mitigations of uncertain knowledge and careers. Consider these interacting elements as preface to the factor that will make all the difference in the future of testing: technology.

Technology and Potential Influences on Testing Futures

In the unpredictable future we know one sure fact: technology, whether silicon, biologically based, or found in the ether will continue to expand and to surprise us (see Beherns, et al., this volume). This section will consider technology, as it now affects

learners and education, its likely influence on tests in the very short run, and more speculatively, how it may play out in the future. Technology will drive options in education and testing far more strongly than ever. As Tom Glennan (1967) presciently put it, "technology-push" will determine much of the nature of educational delivery and assessment systems, rather than our requirements-driven traditions. The technology-push has started and come from unexpected places, faster and at more oblique angles than anyone could anticipate. Furthermore, independent access to the web, without adult or school authority mediation, either to browse, use, or to upload personal material, has changed the content of knowledge, normally the partial province of educational systems. Unsupervised personal access to knowledge portends a massive and continuing change that will debilitate many efforts to maintain control and authority over learning. Let's look at a sector involving people and technology with implications for education. First, consider an analogy of technology development at the present moment: games. Second, consider how technology already has changed expectations of everyone, particularly students and how will new expectations affect assessment. Third, let's look at how technology will be relevant to the design part of assessment.

Games as a Harbinger of Testing Future: An Analogy

Begin with games, an activity seeming far afield from the seriousness and frequent formality of testing and assessment in our current accountable world. Only five years ago, the modal game was an application played on a console or PC that was typically purchased as a shrink wrapped disk, or obtained through subscription and downloaded. Players were usually individuals in the 18–30 age range, although younger kids were responding to handheld platforms and single purpose games, e.g., shoot something, find something, steer something. These games evolved to include many simultaneous players, more complex narratives, increasingly realistic graphics, and special effects to create a persistent user group that used email, chat, and voice interactions with other participants to forward their play. Interactive tasks required collective strategic planning to make progress or to impede the progress of competitors in attaining the objectives of the game. A great fictional example of such a game can be found in *Reamde* by Neal Stephenson (2011) that illustrates both positive and negative

roles for individuals and groups or teams of players. How do such games, of the quest and role playing type, teach strategic thinking, interpersonal skills, and require problem solving and situation awareness (think 21st century skills)? A lecture, by John Seeley Brown (2012), contains a plausible answer. Brown argues he would rather have a successful player of *World of Warcraft (WoW)* than a Harvard MBA to run his business. He points out that WoW and games like it use after action reviews (AARs) as a form of collective formative assessment. Here, the review process is to replay events and evaluate performance. This is a routine practice in training using complex aircraft simulations, military exercises, or, for those who can remember, fighter pilot exercises in the *Top Gun* film. Leaders of the exercise are contractors who have built the simulations. Through captured action, that is replayed, they can point out performance strengths, weaknesses, and turning points. In military field exercises, the AARs are conducted by ranking personnel. Brown argues that the public self-reviews of performance in multiplayer games is essential to serious learning. More importantly, because they are collaboratively managed by players, rather than status authority figures, all participants have an opportunity to identify and suggest reasons for errors, unexpected outcomes, or surprising successes. One idea to be pursued in a subsequent section is the use of collective wisdom in the conduct of formative and other assessments.

But let's return to the game development narrative: in a very short time, console games were enhanced by sensors that monitored and represented the physical movements of the player, in dance, in sports, and other physical endeavors — first with handheld or other devices and then using camera technology to "read" action and record action, and then to play it back so the player could see his performance much like AARs above.

Such subscription multiplayer games remain popular. The numbers of devotees are astonishing. WoW numbers dipped from a high over 11 million monthly subscribers to 10.3 million earlier this year, and the next 12 role-playing games are estimated to account for less than three quarters of the annual total of WoW. Zynga® claims 180 million players a month on Facebook®, and a recent release of *Call of Duty* is reported to support 40 million worldwide players a month; and a popular game, *Halo*, had a recent release that approached the annual subscriptions of WoW.

Yet, unexpected change came to the game scene with the advent of the iPhone® and similar smartphone technology. The development of iTunes® as a distributor for game applications, a source for application (app) developers and the emergence of the iPad® and similar devices added greater portability and direct haptic control of game play with touch screens. Game applications could be downloaded, from the iTunes and comparable sites, with free trials (if one could tolerate advertisements), and low purchase prices (far less than an order of magnitude cheaper than the cost of console games). New apps were inexpensively built with the development kit provided by Apple®, allowing products to be easily tried and discarded. Another order of magnitude shift occurred in number of players related to any one game. For example the app series, *Angry Birds*, reached more than one billion downloads internationally (http://www.slashgear.com). Developed by Finns, wouldn't you know, it includes an accurate physics model, impressive graphics, and a game mechanic that rewards success. It is also smart, in that it rewards patience by including delays instead of developing even more frenetic rates of play. Apple reports in 2011 that 425,000 applications have been developed for iTunes, that they had had 15 billion downloads, and counted an additional 200 million international users in 90 countries. Furthermore, they reported that 100,000 apps had been developed for the iPad alone (Apple computer, press release, May 2011). In the first part of 2012, the number of apps on iTunes increased by over 100,000. Android™ smartphones and pads have their own system of downloads, and report over 400,000 apps available for free or low-cost downloads.

Let us put this rapid change into an educational research and development perspective. When the team at the Center for Research on Evaluation, Standards, and Student Testing (CRESST) began its learning game research for the Office of Naval Research (ONR) and the Institute of Education Sciences (IES) about eight years ago, games for children were intended to be played in schools on desktop or laptop computers, using newly available systems to program interactive game play. Many game "mechanics" did not work across platforms. Numerous technology games now use pads, Smartphones, social networks, voice, real time observation of scientists, geographical regions, and the overwhelming set of options available on video through YouTube®. Students are expected to collect information in place-based environments and bring it into

their game settings. They are also encouraged to work together, either through chat, social networks, or particular functions built into the game so they can do a better job of learning. Numerous games are now attempting systematically, rather than incidentally, to affect learning. The plan is to blend entertainment production values with high-quality instruction, including videos of lectures or key demonstrations. In the game area we are only now seeing many beginning to yield interesting results. But imagine if the growth curve for downloads with games applied as well to achievement measures (for parental review) or, even more importantly, sets of expectations developed by industry (to get jobs), or the higher education (to attain admission).

Personalized and Public Space

The game example is also meant to convey how much choice an individual now has in this domain and to provide an explicit contrast with formal pre-collegiate education. With the advent of the iPod, and the music and video form of iTunes, individuals had, legally, and on a mass scale, the option to buy only tracks of albums rather than the whole compilation and to create special, personalized versions of music consisting of their favorite songs, artists, and renditions. Did you like Michael Jackson singing "Never Can Say Goodbye" when he was a child or an adult? And people embraced the personalization of music. They created multiple "playlists" for travel days, sad days, music for kids, sing-alongs, dance music, and these playlists could be dropped, modified, or enlarged. Unless they are shared, either privately or over the web, and many were, the set of songs is personal, unique, and unlike any other in the world.

Personalization has become a watchword of web interaction, whether it involved changing one's Facebook photo with the seasons, or personal mood, choosing which photos to share on anyone of a number of sites, showing others preferences in food, shopping, décor, movies, and friends, and viewing videos on YouTube, from fascinating TED talks to inane videos of kittens, apparently the number one topic found in a Google® image search (Novig, 2012). As important, everything was changeable and under the control, read authority, of the user. And the users were urged to express their opinions.

Not only was personalization a priority and the way of the web, so was public display. Although one could choose in many cases with whom to share preferences, art

products, personal videos, many people, and many students, were very comfortable with seeking the most public of displays. Some of these involved uploading videos of their singing, comments, and photos, especially of cats. Other less benign use saw hostile or embarrassing comments about others, attacking individuals, bullying them subtly or aggressively and resulting in serious difficulties for the target persons. Some see the web as providing models for antisocial behaviors. But the linked message is that a new expectation for personalized rather than uniform experiences has developed; and perhaps, for many the web has provided a way for them to be separately identified.

Personalization is not a new concept to educational technology. In games, intelligent tutoring systems, simulations, and other technologies, options to personalize interactions have been plentiful. How relevant they have all been to learning is not so clear. For instance, is it of instructional value for students to design avatars? Motivational value? It reminds me of an earlier technology epoch when I was captivated by changing font styles and sizes. In addition to changing looks, sounds, and voices (an innovation well-known on one greeting card site), learners can select pathways through experience and acquire needed resources by search, provision, or winning, if demonstrations of proficiency are included. Projects such as Mobilize, a current research project of the National Science Foundation (http://www.mobilizingcs.org/about) is one of many studies focusing on place-based learning, using phones, sensors, and other techniques to monitor behaviors selected for the most part by the participants.

Reviewing games development history and status has illustrated the first two lessons for assessments of the future (Scacchi, 2012): (1) Assessments will need to change rapidly, to take advantage of the technology, and to meet learners' expectations; (2) they will need to be personalized and fully adapt to the interests, formats, and expectations of individual learners. This stricture goes well beyond the use of algorithms in ITS or CAT to adapt content appropriate to learning levels. Personalization and point-to-point communication, unmediated by authority, is one major feature of widespread technology.

In describing technology and its potential educational impact, one must, simultaneously take into account how these new options will be configured or combined for learning. In the short term, it is safe to predict that learning through technology will

be based on a continued set of connected elements: (1) longer tasks involving both independent and collaborative learning; (2) mobile or device-free connections to technology through camera and sensors; (3) use of virtual tools, including sources, analyzers of text, problems, and (4) automatic ways of modifying difficulty. In addition a major blurring should occur among classroom and other informal sources of learning and assessment. The direction of this shift may favor informal learning or at least widen the types of unsupervised activities that allow users to explore, interact, connect, and manage their own learning resources. This change will place increasing responsibility on students to be responsible for their own learning, rather than conform to directives from adults. However, they will need to learn how to be successful with different requirements. Technology may also assist and record student performances. For instance, it could read journals and blogs for meaning and progress, estimate engagement and motivation, and provide cues to maintain focus (derived from noninvasive biometric instrumentation). There are current projects that can read gesture and engagement through cameras and EEG recordings involved in games. There are systems that can process real-time text and voice, over phone and social networks, to estimate level of proficiency and emotional state. Recent reports on current technology (Edudemic.com) describe a "10-fold increase" in the use of smartphones for studying between 2011 and 2012. The article gives examples of social networks for students who connect online in study sessions, planning, or project activities. These sites have varying degrees of management by teachers and have become a recognized part of higher education. There may be some assessment monitoring on these sites, but it is fairly routine. Standard technologies such as Skype® for visual contact, Dropbox® for collaboration, are presently available and their use bridges educational and day-to-day use. Another existing technology 3-D printing has great promise. More than ten years ago, at Raytheon, I saw a demonstration of three dimensional printing, where designed objects were created by a printer using a resin to fabricate products that could not be developed in any other way. Its early use was for missile parts that required connecting metal parts far more precisely than could be accomplished with welding technology. Such technology is now more routinely used in commercial products. The inevitable advent of desktop technology using multiple materials for fabrication is now enabling students to make products and to exercise

creativity that has a conceptual and palpable value. (See http://www.fool.com/fool/free-report/18/sa-3dprintingaudio-ext-184238.aspx.) Three-dimensional printers allow students to design and create usable objects, or objects that could not be made in another way. Learning to use cad-cam and newer design software will result in tangible hold-in-your-hand creations, another release from the two dimensional space of a display.

How will current developments affect learning futures? Students may be able to explore pathways that focus on proximal goal states as well as longer-term objectives, and develop increasing expertise in choosing among instructional options useful for life-long learning. Even though we are at the start of a long-awaited emphasis on high-quality standards to guide the schools, the standards may be employed as much to guide the students' paths as teachers' instruction. The result is that proximal learning goals and processes will be more personalized than standardized, although standards will provide the meta-menu for options needed to succeed in school.

Four assessment thoughts should intrude at this point: First, personalization is the opposite of formal, standardized and uniform, the current descriptors of much of our testing approach. If tests are to mirror instruction, some big changes are needed. Second, embedded, automated testing and scoring will save time and likely increase the accuracy, speed of feedback and accumulation of validity evidence for inferences. Its benefit of reducing the ceremonial side of testing has a downside. Running tests below the conscious surface as part of seemingly everyday activities raises questions of the meaning of test performance. For example, how will performance preceded by prompted (or scaffolded) experience be treated? Are we looking for average performance over many "natural" trials, or situations where the learners are motivated to do their best? For instance, when students believe that they are responding in an everyday environment, will they be judged on their average level of engagement and performance across multiple trials and technologies? Will they be offered the option to select their own best performance? Fully embedded assessments seem to be based upon a different performance model than one we believe is in place today and implies a modification in the ways in which validity data may be collected.

Third, when testing becomes totally hosted on the web, the security of individual performances is at some risk, as it pertains to their personal level of learning and

performance. In today's schools, students check at least some of their privacy at the door. Their test scores are routinely shared with other teachers and relevant instructional help. The web permits the accumulation of potentially permanent individual-performance information. How long will these records last, how will they be protected, and who will have access to them? Moreover, there is a strong chance that student hacking of computer-based testing systems will occur. I have great confidence that students who seem to be able to get into banking systems will have little trouble with our State examination technology.

It should be an option for students and parents for every technological program to determine whether they want to protect today's performance from others who might make inferences from it later in their lives. Ironically, the learners themselves may forget that comments made on social networks can be accessed by millions of unknown others. Data from these sources could follow students forever, and limit their ability to reinvent themselves as they mature. Consider the phenomena that universities and employers read applicants' social network history to learn about them. Yet, in some European countries, there is a mechanism to allow students to erase formally earlier performance, so when they look for a job, they can put their best self forward.

Fourth, limitation of technology-based assessment, and probably only short-term concern is test security. In this case, one is not concerned with the privacy of the individual's response, but the opportunity for the test questions or answers to be shared with others. The technical term for this is cheating, and the availability of tests on computer networks heightens its probability. Test security assumes that access to the test is fair for all students, but we have seen in the United States., some cases of inappropriate practice of test content. More recently, smartphone pictures of test items have been posted on the web. In a situation where many may share the same or comparable test content, but administer them to students during a relatively broad interval, this sort of cheating is likely to increase. I think an excellent solution (Baker, *Teachers College Record*, in press) is to make test items and tasks available to all *before the test is given.* To make this work, the testing tasks should be very hard, differ in format, consist of thousands of particular items, and published with no answers given.

Assessment Development

Technology will help design, score, and monitor different student activities through the analysis of web experience (data mining) and determining with various learning analytics the types of progress individuals are making along differing goals and content trajectories.

Using technology to make assessment is not an original idea (see for instance, Millman & Outlaw, 1978), nor to date, has it been executed in a fully satisfactory manner. Model-based design as it has been described (Baker, 1997) was based on this idea, but only partially implemented. The advent of new technology now has allowed "model" to be used in two ways: the design model and the computational model needed to create any particular task or task set, whether used in instruction or in assessment. Structured model-based assessment design depends upon the availability of bounded sets of assets, that when sampled can be combined to construct an assessment with its own scoring model. The resulting task should systematically relate to a standard or instructional goal or serve as an integral part of a predesigned or learner-constructed instructional sequence. The bounded domains form the structure of the meta-domain of assessment. At this point, in trials at CRESST, they include content and cognitive ontologies, sets of narratives, texts, scenarios, motivating events, problem formulations, impediments, paths to solutions, and relevant criterion elements, posed as questions. Not all elements will be used for any single assessment, and they will vary by purpose, age, topic, expectation, and standard. But the approach may allow us to solve some current problems. For instance, cost. Simple combinatorial models can transform elements, even with constraints on numbers of options that can be used on the task from each category, into very large domains, i.e., 6 with 16 zeros (Baker, Chung, & Delacruz, 2012b). Following Pearl (2012), we call the resulting combination of assets a signature. Of key importance is how these elements, which exist at different levels of specificity, roll up into coherent tasks. Part of the answer is a set of computational models (Cai, in press) which will permit — if successful, the generation of sample tasks. It is our goal to create such tasks on-the-fly, in real time, by combining assets for particular purposes. Segments of these componential, combinatorial, computational development process are undergoing a series of trials now. Within a learning system on-the-fly creation of the

personalized scenarios, game levels, or assessment tasks (that is, what the student sees next) would be inferred from the student's performance.

There are empirical questions to be answered about whether the degree of refinement generated by such a personalized approach is worth the effort when compared to pre-stored instructional or assessment options. Obviously, the same system of assets, objects, and computational models can be used to create the original set of tasks to be stored. However, it is clear that the concept should be evaluated and assessed. In our trials at CRESST (based on a position paper by Cai, Chung, Delacruz, Baker, & Iseli, in process) tasks have been created far faster and for far less cost. Even without the technology administration and or scoring, we may be on the verge, as are others, in creating more flexible and efficient assessments, for example, to be applied in a performance assessment setting. The system extends our earlier research on reusable templates, employing cognitive demands, content ontologies, and formalized tasks conditions (Baker, Freeman, & Clayton, 1991; Vendlinski, Baker, & Niemi, 2008). The new effort is a serious improvement because we are adding situational variables with known values, and a system of verification using layered computational models. This approach seems to build nicely on Evidence Centered Design (Almond, 2010 p. 97; Gee, 2010; Hickey, Honeyford, Clinton, & McWilliams, 2010; Mislevy, Steinberg, & Almond, 2003), using technology differently. It also comports with Kay (2012) and Pearl (2012) who have advanced thinking about higher order generation of usable and generalizable solutions. With high-quality automated scoring addressing meaning of written responses, that is, propositional understanding, is essential to the guts of the system, can be developed in at least a semi automated way, with people acting as editors, interpreters of data, and determining the extent to which validity and comparability claims can be met.

Whether or not the system should be composed in real-time, the data summaries flowing from individuals use over time and groups of students allow the analysis of the impact of numerous variables derived from a combination of student data, using probabilistic methods. Settings could be disparate (for instance, classrooms and mobile phones) and be examined separately or weighted in concert, depending upon the standard/skill under study. The key function of computational models is to move from lowest level of data to indicators intended to represent particular purposes, such as to give

students feedback, to provide teachers with guidance or actual lessons, and to give parents individual and group results during and following key instructional units (Shute, 2008). The data also are examined from a theory (or learning and instruction) perspective and the interaction between the bottom-up analyses derived from student databases and the top-down expectations, will suggest experimental or on-the-fly modifications that can help resolving any disparities in view points. Just as teachers become attuned to the individual behavior of different students, in a technology framework, systems will become more sensitive to students by using aforementioned sensors, gestures, and audial information in order to adapt systems to increase engagement, attention, and persistence, if needed to support learning.

Assessment in Summary

Technology will revolutionize assessment, from design of assessment to scoring, and monitoring different student activities through the analysis of web (instructional) experience, and determining through a kind of learning analytics the types of progress students make. These analyses can take into account differing goals, skills, and content trajectories. To survive the continuing knowledge explosion there will be greater emphasis on cognitive or thinking skills, as well as self-management skills. Various statistical summaries allow the combination of student data in different settings, to provide indicators for different purposes. Whether these indicators will be dramatically suboptimal for particular purposes will be the topic of future students.

Future Visions of Assessment Technology

What will we expect in five years or ten as the routine ways in which assessments will operate? What tools on today's horizons will have impact in assessment design. How will new approaches adapt to individual's desires for personalization, growing transparency and openness of expectations, the need for comparisons, and the speed needed for assessment development to respond to changes in knowledge, jobs, and technology?

Let's posit the following: learning skills will be embedded in content or multidisciplinary problems. Technology will be used to design, administer, score, store, and report findings to entitled users. Schools and education systems will not be the only source of assessments. Students will make things, not just give answers. And they will be working in a more globalized environment.

Assessment Design Supports: Methods and Content

Method: Ontologies. At the present time, there have been approaches that use detailed representations of expectations, either translating the Common Core State Standards (the Standards) into network representations, or making relatively simplified maps to assist teachers on a dashboard (see the Literacy Design Collaborative, http://www.literacydesigncollaborative.org/about/). The use of multi-layered ontologies of content domains, e.g., calculus, derives from the design of artificial intelligence systems. Ontology content was typically elicited from experts using clear protocols that could be instantiated in software rules. Ontologies can be created from combined expert knowledge, such as represented in the Standards, or by querying individuals. In addition, new natural language processing systems permit the addition of content from extant text to ontologies (Iseli, 2011; Wimalasuriya & Dou, 2010). In the current environment, such ontologies provide a direct method to use to assure the representativeness of assessment tasks, to recognize relationships and sequence options, to operationalize "big ideas" by looking at nodes with most connections, and to determine fundamental requirements by mapping subordinate layers. In the discussion of real-time game level or assessment task development, ontologies were a key asset to be sampled.

The ontology methodology has been applied to problem solving, situation awareness, and communication metacognitive components, such as planning, in game development. While such ontologies in no way fully represent a content system, they provide an easy mechanism to modify content as new knowledge develops. Small features or whole sections of the ontology might change as a result of new discoveries. These carefully developed ontologies to be used by assessment development systems should not be confused with maps intended to guide teachers, students, or parents through desired learning sequences. They are obviously connected, but most maps are generated

at a palatable level of specificity. Ontologies may be deeper layered and propagate specificity. Next on our list is the improvement of the ontology capturing the requirements of transfer situations, an idea generated by Lauren Resnick many years ago.

Assessment design models vary by their level of specificity. Mislevy and his colleagues created a model that followed the general lines of AI systems (a student model, a content model) but elaborated it with the need for evidence to support the validity and utility of resulting assessments (Mislevy, Steinberg, Breyer, Almond, & Johnson, 1999). Other models have focused on the relationship of cognitive processes to content in specific task types (Baker, 1997; Paas, Renkl, & Sweller, 2003). The main focus of such models is to develop approaches that affect learning through assessment, either by the induction of schema, or through specific features explicated in tasks. The latter approach, a spinoff of Gagne's work in task analysis (1965) has a cognitive orientation, which is both appealing but highly idiosyncratic (Clark, 2003). It is reasonable to expect a new wave of assessment design and improvement, focusing on combinations of sophisticated computational modeling. They may make the design as well as the administration of assessments a dynamic process.

To support transparency for users, examinees will know what features are part of the task set, but each item could be developed from elements residing in servers. Preliminary analyses imply that early performance, say on assessments embedded in a game will predict performance in other games related to states (Popovic, 2012, DARPA telecom August 20, 2012), common content, cognitive demands or all three. So models embedded with dynamic quality metrics, will be one way to keep up with changes in technology options.

Method: Badges. Another method of obtaining credit for performance other than by taking tests, or even completing validated instructional sequences has been a topic of recent interest: badges signifying specific achievements. Badges are awarded upon the completion and verification of a predefined set of integrated tasks. The idea of earning a qualification derives from external certification, such as a network administrator by technology companies, by career requirements involved in the German educational system, or extrapolations from scouting badges (http://dmlcompetition.net/media/4/BadgesforLifelongLearningAnnouncement.pdf).

Using Scouting Badges as an analogy Albert Shanker more than three decades ago (1988) argued that conceptually linked or practically driven tasks would have greater meaning for students and currency outside of school. An approach to badges implemented in a school system can be observed in the New Zealand effort to substitute some required testing for particular applied learning (NZ: http://www.minedu.govt.nz; http://www.ial.ac.nz). Currently the badge environment in the U.S. is rather complicated and uneven. It has not yet obtained the legitimacy needed by the educational system. For full implementation, it will depend upon opening schools up to evaluators with specialized skills beyond the regular teacher, a way to accredit such individuals, to make difficulty comparable, and to figure out the substitution patterns needed to exchange uniform tests for systems that count badges attained in category systems related to adopted standards. The complexity of this transformation, however, is likely of value in the quest to personalize education, for students as they grow, will have greater and greater choice (and motivation) to accomplish real tasks. If such accomplishments become part of student resumes and sets of employer requirements, the link between learning of some topics and job needs will be tighter. The badges also support the idea of transparency because requirements and criteria are explicit. Clearly, early computer support (such as Mozilla's backpack site for badge collection) would need to be supplanted by more sophisticated models that assisted learners in their acquisition of required skills and identified resources for assistance, feedback, and certification.

Method: Crowdsourcing. Crowdsourcing is a current term to describe a process where users of the web generate information, suggestions, judgments, and other inputs to particular tasks and their views are (often) made publically available. Probably the easiest way to think about crowdsourcing is to reflect on Amazon® ratings of products purchased by buyers. Here the views of users is compiled and fed back to potential buyers, and the processes replicates itself on most shopping websites. It can be contrasted with ratings given to descriptive aspects of products by Consumer Union in *Consumer Reports*. While both approaches want to affect buyers' choices, and they both use simplified rating metrics, i.e., stars vs. red to black ratings, they vary in important ways. One is based on expertly obtained evidence, the other by collective wisdom of the unknown user. One is

explicitly criterion-referenced, although reporting product data comparatively while the other is based on normative judgments by unknown individuals.

The "analytics" movement around web use began largely as a descriptive venture, but evermore sophisticated analyses of massive data, called "data mining" are in play, subdividing user groups based on websites visited, time spent, money spent, and inferred demographic information. Much of current data mining uses inductive logic, to detect clusters of meaningful performance. Not unlike exploratory factor analyses, inferences drawn from large data sets are used to customize (personalize) email and offers to individuals with various patterns of web journeys and purchases profiles. Clearly web analytics and targeted marketing now depend upon popularist rather than expert base. Popovic (2008) found in his *Foldit* game, involving complex protein folding options that many of the "crowd" developed solutions that had escaped experts, validating at least partially the idea that there is a collective intelligence.

Recently, crowdsourcing and web analytics have been used in designing assessments (the use of icons for children who cannot read) and in evaluating games and other educational implementations (Chung, this volume; Roberts, Chung, & Baker, 2012). This methodology can be used in connection with careful design to evaluate the target implementation, with one reservation: it is not clear who is actually providing data.

Today, researchers and developers are refining approaches using natural language processing (NLP), speech recognition, and gesture analysis to make inferences about cognitive and emotional states of individuals. Pentland's *Honest Signals* (2008) is an excellent introduction to this field. More invasive, at this point, are imaging studies, conducted with a functional magnetic resonance imaging (fMRI) apparatus to monitor attention, focus, and other cognitive states while engaged in activities, including test performance. Similarly, researchers engaged in imaging using fMRI equipment are moving beyond the study of individuals and into the world of learning in interactive settings, but these technologies will improve and perhaps provide more direct measures of learning, when combined with behavioral or product creation.

Content: Cognitive and Affective Skills

Although treated elsewhere in the volume, there is considerable interest in dealing with the use of cognitive skills as a way to anticipate the changes in content or knowledge that is the inevitable future of this world. Teaching students how to acquire and store new knowledge (in schema), how to apply it in new or unforeseen situations (transfer and generalization) are obviously within our reach. What we have yet to do is to take the ever increasing and ever diverging definitions of deeper learning, 21st century skills, and come up with an explicit first set to be included systematically in education. We can tell by the overlap that teamwork, communication, and problem solving are on most lists. The lists diverge when they begin to include personal traits rather than learned skills. But of course that may be an old way of seeing things. We may very well be able to teach students to be more creative, less impulsive, more systematic in their searches, more evaluative in their judgments about what information they should value and include in their searches and applications. They certainly will have to learn to live and interact with a wider variety of people, people from next door and the next continent. Having a shared set of such skills, locally instantiated, but having some general properties may make collaboration and joint problem solving in environmental studies, history, and politics more likely to succeed.

Summary of Future Options in Assessment

Starting from where we are and the benchmark of international performance, unceasing changes in demography, knowledge expansion, job instability, and technology growth portends many possible assessment futures. We have excluded numerous important variables such as politics, changing preferences, individual differences, and any significant focus on changes in values, character, or emotional and spiritual, outcomes. Nonetheless, this less than perfect prediction attempts to unite emerging developments with ways assessment might evolve. The wisest path is to formulate predictions as questions. Think of these as thought experiments. Consider the alternatives and play out the consequences as you see them.

- Who will be in charge of assessment in the future? Current national school authorities, multinational organizations, or corporations with interests in

maintaining well prepared workforces? A combination? How will the transformation take place?

- What is the role of self-assessment?
- How will demography affect assessment? Will it change who is assessed outside of school? What will it do to higher education within countries? What will happen if appetites for international experiences continue to grow? What are the implications for instructional strategies, cultural diversity, and interpersonal skill development?
- Will investing in transfer and generalization of learned outcomes as well as skill development applied to new situations be a useful strategy to address the change and unpredictability of content? Is it sufficient? How will the changes in content (data) be connected to the changes in application and general understanding of non-specialists?
- Will assessments be integrated fully into instructional systems? Will validity of assessments be less relevant than the evaluation of the instructional system itself? What criteria should be used?
- Will personalization trump standardized indicators of performance?
- Will we lose too much art by automating assessment and instruction? Do the same processes need to be applied for all subject areas or just some? For particular application domains? Differentially depending upon age and experiences of students?
- Can choices be offered to students about how and when to develop their competencies? How early? How much?
- Will performance and extended tasks supplant the persisting preference for small, discrete test items? Can badges or systems of qualifications be developed and gain credibility along with documented validity? Can they be linked to surface systems, such as standards, and deeper representations like ontologies? Can their quality, difficulty, and ultimately bands of comparability be managed successfully?
- Will neuroscience affect the way in which accomplishments are measured or validated? How will that work in a less invasive manner? How soon?

- What will be the relationship of expertise and populist versions of knowledge?
- How will old issues such as privacy and security evolve? Will the public space of technology affect the entire process?

Final Note

Papers should answer more questions than they raise. Or should they? My answers are not yours; my dreams and fears are unique, and my bets on the future waver depending upon my optimism about individualism, a yearning for wise shared values, and my observations of superficial division. What we should be looking toward is finding strategies by which the future gets better through technology and its distribution and effects worldwide.

References

Almond, R. G. (2010). Using evidence centered design to think about assessments. In V. J. Shute, B. J. Becker (Eds.), *Innovative assessment for the 21st century* (pp. 97). New York: Springer.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Arai, K. (2011). Discussant presentation at the session "K-12 and University Education – New Approaches to University Entrance Examinations," International Symposium of Organization for the Study of College Admissions (OSCA), National Center for University Entrance Examinations (NCUEE), Tokyo.

Argyris, C., & Schon, D. A. (1978). *Organizational learning: A theory of action perspective*. Reading, MA: Addison-Wesley.

Australia Department of Education: Australian Government: Department of Education, Employment and Workplace Relations: http://www.deewr.gov.au/Pages/default.aspx

The Government of Western Australia: The Department of Education Western Australia: http://www.det.wa.edu.au/

NSW Government: Education & Communities: http://www.dec.nsw.gov.au/home/ Queensland Government: DETE Education: Department of Education, Training and Employment: http://education.qld.gov.au/

South Australia Government: South Australia Department for Education and Child Development: http://www.decd.sa.gov.au/

State Government Victoria: Department of Education and Early Childhood Development: http://www.education.vic.gov.au/

Baker, E. L. (in press). The chimera of validity. *Teachers College Record*.

Baker, E. L. (in press). Learning and assessment: 21st century skills and cognitive readiness. In H. F. O'Neil, R. S. Perez, & E. L. Baker (Eds.), *Teaching and measuring cognitive readiness*. New York: Springer.

Baker, E. L. (1997, Autumn). Model-based performance assessment. *Theory Into Practice, 36*(4), 247–254.

Baker, E. L. (2010). *Summative and formative evaluation in educational accountability* (working paper prepared for the work of the Advisory Council on the Evaluation and Incentive Policies, OECD). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Baker, E. L., Chung, G. K. W. K., & Delacruz, G. C. (2012b). *DARPA ENGAGE program review, CRESST – TA2*. Presentation at the ENGAGE PI Meeting (Phase I Review): Defense Advanced Research Projects Agency, Arlington, VA.

Baker, E. L., Chung, G. K. W. K., & Delacruz, G. C. (2012a). The best and future uses of assessment in games. In M. C Mayrath, J. Clarke-Midura, D. H. Robinson, & G. Schraw (Eds.), *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research* (pp. 227–246). Charlotte, NC: Information Age Publishing.

Baker, E. L., Freeman, M., & Clayton, S. (1991). Cognitive assessment of history for large-scale testing. In M. C. Wittrock & E. L. Baker (Eds.), *Testing and cognition* (pp. 131–153). Englewood Cliffs, NJ: Prentice-Hall.

Baker, E. L., O'Neil, H. F., Jr., & Linn, R. L. (1993). Policy and validity prospects for performance-based assessment. *American Psychologist*, *48*(12), 1210–1218.

Bereiter, C. (1995). A dispositional view of transfer. In A. McKeough, J. Lupart, & A. Marini (Eds.), *Teaching for transfer: Fostering generalization in learning* (pp. 21–33). Mahwah, NJ: Erlbaum.

Bjorner, J. B., Chang, C.-H., Thissen, D., & Reeve, B. B. (2007). Developing tailored instruments: Item banking and computerized adaptive assessment. *Quality of Life Research, 16*, 95–108.

Bloom, B. S. (Ed.). (with Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R.). (1956). *Taxonomy of education objectives: The classification of education goals. Handbook I: Cognitive domain*. New York: David McKay.

Brown, J. S. (2012). *How World of Warcraft could save your business and the economy*. YouTube lecture: http://www.youtube.com/watch?v=BhuOzBS_O-M.

Cai, L. (in press). Potential applications of latent variable modeling for the psychometrics of medical simulation. *Military Medicine*.

Cai, L., Chung, G. K. W. K., Delacruz, G. C., Baker, E. L., & Iseli, M. R. (in process). *Computational model for integrating learning, instruction, and assessment* (Position Paper). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Chung, B. M. (2010). *Development and education: A critical appraisal of the Korean case*. Seoul: Seoul National University Press.

Chung, G. K. W. K. (2012, in preparation). *Toward the relational management of educational measurement data*. Invited chapter for ETS Gordon Commission on the Future of Testing.

Chung, G. K. W. K., & Baker, E. L. (2003). Issues in the reliability and validity of automated scoring of constructed responses. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 23–40). Mahwah, NJ: Erlbaum.

Chung, G. K. W. K., Baker, E. L., Brill, D. G., Sinha, R., Saadat, F., & Bewley, W. L. (2006). *Automated assessment of domain knowledge with online knowledge mapping* (CSE Tech. Rep. No. 692). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Clark, R. E. (2003). Fostering the work motivation of teams and individuals. *Performance Improvement, 42*(3), 21–29.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Erlbaum.

Dansereau, D. F. (1995). Derived structural schemas and the transfer of knowledge. In A. McKeough, J. Lupart, & A. Marini (Eds.), *Teaching for transfer: Fostering generalization in learning* (pp. 93–121). Mahwah, NJ: Erlbaum.

Darling-Hammond, L. (2010). *The flat world and education: How America's commitment to equity will determine our future*. New York: Teachers College Press.
FIMS. (1964). http://www.iea.nl/fims.html

Fletcher, J. D. (2004). *Cognitive readiness: Preparing for the unexpected*. In J. Toiskallio (Ed.), *Identity, ethics, and soldiership* (pp. 131-142). Helsinki: Finnish National Defence College.

Fletcher, G. (2007, October 19). *Assessing learning from a holistic approach: Creating a balanced system of learning assessment*. Paper presented the Congreso Internacional Evaluacion Factor de Calidad Educativa, Queretaro, Mexico.

Gagne, R. M. (1965). *The conditions of learning* (2nd ed.). New York: Holt, Rinehart, & Winston.

Gee, J. P. (2010). Human action and social groups as the natural home of assessment: Thoughts on 21st century learning and assessment. In V. J. Shute, B.J. Becker (Eds.), *Innovative assessment for the 21st century* (pp. 13–40). New York: Springer.

Gibbons, R. D., Weiss, D. J., Kupfer, D. J., Frank, E., Fagiolini, A., Grochocinski, V. J., Bhaumik, D. K., Stover, A., Bock, R. D., & Immekus, J. C. (2008). Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatric Services, 59*, 361–368.

Gibson, E. J. (1940). A systematic application of the concepts of generalization and differentiation to verbal learning. *Psychological Review, 47*(3), 196–229.

Gick, M. L., & Holyoak, K. J. (1987). The cognitive basis of knowledge transfer. In S. M. Cormier & F. D. Hagman (Eds.), *Transfer of learning: Contemporary research and applications* (pp. 9–46). Orlando, FL: Academic Press.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes. *American Psychologist, 18*(8), 519–522.

Glennan, T. K. (1967). Issues in the choice of development policies. In T. Marschak, T. K. Glennan & R. Summers (Eds.), *Strategies for research and development* (pp. 13–48). New York: Springer-Verlag.

Hickey, D. T., Honeyford, M. A., Clinton, K. A., & McWilliams, J. (2010). Participatory assessment of 21st century proficiencies. In V. J. Shute, B.J. Becker (Eds.), *Innovative assessment for the 21st century* (pp. 107–138). New York: Springer.

Holyoak, K. J., & Koh, K. (1987). Surface and structural similarity in analogical transfer. *Memory & Cognition, 15*, 332–340.

Iseli, M. (2011). *Ontology development: Overview and example* (Draft CRESST Whitepaper). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Kay. (2012). Presentation at the "Celebration in Honor of Professor Judea Pearl, Winner of the 2011 Turing Award," UCLA Computer Science Department, Los Angeles.

Landauer, T. K. (1998). Learning and representing verbal meaning: The latent semantic analysis theory. *Current Directions in Psychological Science, 7*(5), 161–164.

Li, Z. (2010, April). *Shanghai coming to grip with its aging population problems* (EAI Background Brief No. 517). Available at http://www.eai.nus.edu.sg/BB517.pdf

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, *20*(8), 15-21. (ERIC Document Reproduction Service No. EJ 436 999)

MacArthur Foundation. (2011). http://foundationcenter.org/pnd/rfp/rfp_item.jhtml?id=354300034

Martens, K. (2007). How to become an influential actor – the 'comparative turn' in OECD education policy. In K. Martens, A. Rusconi, & K. Leuze (Eds.), *New arenas of education governance – The impact of international organizations and markets on education policy making* (pp. 40-56). Houndmills, Basingstoke, Hampshire, UK: Palgrave Macmillan.

Mayer, R. E. (2008). Applying the science of learning: Evidence-based principles for the design of multimedia instruction. *American Psychologist, 63*(8), 760–769.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education/Macmillan.

Millman, J., & Outlaw, W. S. (1978). Testing by computer. *AEDS Journal, 11*(3), 57–72.
Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessment. *Measurement: Interdisciplinary Research and Perspective*, *1*(1), 3–62.

Mislevy, R. J., Steinberg, L.S., Breyer, F.J., Almond, R.G., & Johnson, L. (1999). A cognitive task analysis, with implications for designing a simulation-based assessment system. *Computers and Human Behavior, 15*, 335–374.

Mousavi, H., Kerr, D., & Iseli, M. R. (in process). *Unsupervised ontology generation from unstructured text* (CRESST Rep.). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

National Commission on Excellence in Education. (1983). A nation at risk: The imperative for educational reform. *The Elementary School Journal, 84*(2), 112–130.

National Research Council. (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. Committee on Defining Deeper Learning and 21st Century Skills, J. W. Pellegrino and M. I. Hilton (Eds.), Board on Testing and Assessment and Board on Science Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

New Zealand Ministry of Education: http://www.minedu.govt.nz/
No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 C.F.R. (2002).

November, A. (2009). *Creating a new culture of teaching and learning*. Available at http://novemberlearning.com/wp-content/uploads/2009/02/creating-a-newculture-of-teaching-and-learning.pdf

Novig. (2012). Presentation at the "Celebration in Honor of Professor Judea Pearl, Winner of the 2011 Turing Award," UCLA Computer Science Department, Los Angeles.

O'Neil, H. F., & Lang, J. (in press). What is cognitive readiness? In H. F. O'Neil, R. S. Perez, & E. L. Baker (Eds.), *Teaching and measuring cognitive readiness*. New York: Springer.

O'Neil, H. F., Perez, R. S., & Baker, E. L. (Eds.). (in press). *Teaching and measuring cognitive readiness*. New York: Springer.

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist, 38*, 1–4.

Page, E. B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan, 47*, 238–243.
Pearl, J. (2012). *In quest for the grammars of science*. Presentation at the "Celebration in Honor of Professor Judea Pearl, Winner of the 2011 Turing Award," UCLA Computer Science Department, Los Angeles.

Pentland, A., with Heibeck, T. (2008). *Honest signals: How they shape our world*. Cambridge, MA: The MIT Press.

Popovic, Z. (2008, December 17). *CASP8 results*. *Foldit blog*. http://fold.it/portal/node/729520.

Popovic, Z. (2012). DARPA Telecom August 20.

Pressley, M. (1995). A transactional strategies instruction Christmas carol. In A. McKeough, J. Lupart, & A. Marini (Eds.), *Teaching for transfer: Fostering generalization in learning* (pp. 177–213). Mahwah, NJ: Erlbaum.

Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., Thissen, D., Revicki, D. A., Weiss, D. J., Hambleton, R. K., Liu, H., Gershon, R., Reise, S. P., & Cella, D. (2007). Psychometric evaluation and calibration of health-related quality of life items banks: Plans for the patient-reported outcome measurement information system (PROMIS). *Medical Care, 45*, S22–31.

Roberts, J., Chung, G. K. W. K., & Baker, E. L. (2012, January). *Using game-based data for performance monitoring*. Joint PBS KIDS / UCLA/CRESST white paper prepared for the Office of Science and Technology Policy (White House).

Rueda, R. (2011). *The 3 dimensions of improving student performance: Matching the right solutions to the right problems*. NY: Teachers College Press.

Scacchi, W. (Ed.). (2012, July). *The future of research in computer games and virtual worlds: Workshop report* (Tech. Rep. UCI-ISR-12-8). Irvine, CA: University of California, Irvine, Institute for Software Research, University of California, Irvine, Irvine, CA. July 2012. http://www.isr.uci.edu/tech_reports/UCI-ISR-12-8.pdf

Schmidt, W. H., McKnight, C. C., Houang, R. T., Wang, H. C., Wiley, D. E., Cogan, L. S., & Wolfe, R. G. (2001). *Why schools matter: A cross-national comparison of curriculum and learning*. San Francisco: Jossey-Bass.

Shanker, A. (1988, November). Reforming the reform movement. *Educational Administration Quarterly, 24*(4), 366–373.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153–189.

SIMS. http://www.iea.nl/sims.html

Six Subject Survey. http://www.iea.nl/six_subject_civic_education.html; http://www.iea.nl/six_subject_french.html; http://www.iea.nl/six_subject_english.html; http://www.iea.nl/six_subject_literature.html; http://www.iea.nl/six_subject_reading.html; http://www.iea.nl/fiss.html

Sahlberg, P. (2011). *Finnish lessons: What can the world learn from educational change in Finland?* New York: Teachers College Press.

Stein, G. (1922). Sacred Emily. In G. Stein (author), *Geography and plays* (p. 187). Boston: The Four Seas Company.

Stephenson, N. (2011). *Reamde: A novel*. New York: HarperCollins.

Sweller, J., van Merrienboer, J., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*, 251–296.

Tyler, R. W. (1949). *Basic principles of curriculum and instruction*. Chicago, IL: University of Chicago Press.

Vendlinski, T. P., Baker, E. L., & Niemi, D. (2008). Templates and objects in authoring problem-solving assessments. In E. L. Baker, J. Dickieson, W. Wulfeck, & H. F. O'Neil (Eds.), *Assessment of problem solving using simulations* (pp. 309–333). New York: Erlbaum.

Wainer, H., Dorans, N., Eignor, D., Flaugher, R., Green, B., Mislevy, R., Steinberg, L., & Thissen, D. (Eds.). (2000). *Computerized adaptive testing: A primer* (2nd Ed.). Hillsdale, NJ: Erlbaum.

Wimalasuriya, D. C., & Dou, D. (2010). Ontology-based information extraction: An introduction and a survey of current approaches. *Journal of Information Science, 36*(3), 306–323.
Written Composition Study. http://www.iea.nl/written_composition_study.html
United Kingdom Department for Education: http://www.education.gov.uk/
United States of America: U.S. Department of Education: http://www.ed.gov/