From the Assessment OF Education to the Assessment FOR Education: Policy and Futures

www.tcrecord.org/Content.asp

Context: Educational reform in the United States has had a growing dependence on accountability achieved through large-scale assessment. Despite discussion and advocacy for assessment purposes that would assist learning, provide help to teachers instructional plans and execution, and give a broader perspective of the depth and breadth of learning, the general focus still remains on accountability, now elaborated with sanctions for schools and personnel.

Focus of Study: To generate scholarly discussion, options for practice, and grounded predictions about testing in the next decades, the Gordon Commission on the Future of Assessment in Education.

Participants: Convened over a two-year period and with 30 people on the steering committee, the commissioners included well-known scholars grounded in psychometrics, assessment design, technology, learning, instruction, language, subject matter, and teaching in discussion. Commissioners, additional authors, and reviewers were largely drawn from universities, private profit, and nonprofit institutions. Professor Edmund W. Gordon was the Chair of the Commission.

Design: A knowledge acquisition and synthesis study, the product design relied on papers authored by expert scholars describing their understanding of productive student testing in their own domains. The commission funded papers on a wide variety of topics. This paper focuses on two of the major topics of the reports, the emphasis on shifting assessment to help rather than simply to mark progress and how future contexts, including technological change, may impinge on testing options.

Conclusions: The paper calls for a transformation of assessment purpose and use, from annual, time-controlled accountability assessments to more continuous assessments used in the course of a learners acquisition of understanding, motivation for learning, collaboration, and deep application of knowledge in problem solving, communication, and authentic settings. Assessments should emphasize helping students of varying backgrounds and goals as well as their teachers. The role of technology as an assessment design, administration, and reporting toolset is described, in the context of changing knowledge expectations and a global competitive environment.

Education policy and practice in the United States as well as the world attends intensely to levels of student achievement as measured by tests. In the United States, the problem has a large and complex face: Our national data show essentially no improvement over a number of years, most recently reported in the 12th grade results (NAEP, 2014). Unlike countries with traditional reliance on tests and competitive performance, such as China, South Korea, and France, in the United States our distinct political divisions result in crosscutting tensions. Many would prefer minimal testing whereas others would rather use testing practices in the most successful international economies. . In the period of less than three decades, or a generation and a half, American students have been eclipsed by students on the international front from countries both larger and smaller, poorer and as diverse, with widely varying disparate educational and pedagogical practices. Despite controversy about international tests, successful performing countries share a national commitment to the learning of children and young adults, that has not been evident as a consistent thread through the last few decades, and many use both formative and summative assessments. This article will address how the measurement sciences, assessment, testing, and their useful interpretation by teachers and students could partially solve the problem.

For most of the 21st century, the nation has been generally aware of the under-productivity of schooling, bringing at best about 60% of our students to a level of expected competence as measured by current tests. In addition, dropouts have not dramatically reduced and are differential for students of low income and varied backgrounds. As the standards for intellective competence have risen and become more complex, in the newly generated Common Core State Standards (CCSS) and the new array of assessments developed to measure them, the goal of achieving universal competence of American students presents associated challenges. Our nation has responded to this general problem by demanding more of schooling, by increasing its mechanisms of accountability, and by depending more and more on the data from standardized tests of academic achievement. Explanations for poor performance have featured the increasing diverse membership of school age children, not at all a unique phenomenon to the United States.

Reflecting on the work of our colleagues on The Gordon Commission on the Future of Assessment in Education (2013), the authors of this article assert that measurement science can do better if allied with more enlightened educational policies. We can improve how we measure academic achievement, with assessments that are more inclusive and have demonstrably greater validity for the range of students, settings, and purposes to be served. The goal remains the same, underscored by emerging learning concepts, research findings, and new technologies. It is now very possible to use measurement science to analyze, document, and appraise the teaching and learning processes so their results can inform teaching and learning, with the caveat that the assessments themselves must have value. Indeed, we believe that in most instances, the measurement itself should be part of instruction. Thus, measurement science should be as much concerned with the cultivation of the process of intellective and non-cognitive abilities as it has traditionally attended to the measurement of developed ability or the inferred results from tests of outcomes. There are two major points with which we initiate this chapter. We advance the position that instead of assessment OF education as a sole goal, we should concern ourselves with assessment FOR education as our primary target. That slogan to begin to be real demands that the system weaves in measurement, tests, and assessment as regular features of its educational fabric, with changing patterns, which encompasses designs for learning, for context, for new knowledge and values, and for public policy. To do this, measurement must get fashioned anew.

We are in the middle of an important debate concerning education in the nation. There is obvious restiveness among many concerning the often-inappropriate use of standardized tests to drive high-stakes decisions concerning pedagogical policies and practices. Some people also remain confused about the new common standards and their implications for learning and assessment given recent revelations about privacy, they are wary of the uses of any test data.

Many sense that too much is happening too fast, a characteristic of much of U.S. society. We too worry that speedy action may be addressing the wrong parts of the problem without sufficient information and reflection. US efforts to employ higher educational standards have recurred, restarting most recently in 1991. Almost 25 years later, accountability-focused assessment has relatively little to show for itself. We are in a fundamental holding pattern, despite surface changes. Consider that Emily, an eight year old, is in third grade but once in her life. If she misses a chance to tackle new learning, it will only be harder for her a year or two down the road. Slowing the process down is not what we need. We need a rapid, thoughtful approach that focuses on the key issues rather than one that simply repeats the past. Instead we need assessment to progress in palpable ways to affect the real growth of students.

New assessments, if they measure new challenging goals, will appear to be difficult, in part because these concepts are just beginning to be taught. Any drop in performance must be analyzed to determine which elements are associated with assessment tasks that are measuring only students non-school background experiences and which can be profitably addressed by revamping instruction, motivation, and materials. Recent studies by Baker, Cai, Choi, and Buschang (2014) help show aspects of tasks more amenable to instruction, i.e., instructionally sensitive, and which probably need redesign.

Rather than anxiety and reflexive pushback, we could go forward now, emphasizing the transparency of assessments for students, parents, and teachers. Transparency of well-designed measures can reveal measurement of challenging goals, goals that rise to the level of those of our economic competitors and that create real feelings of efficacy for learners.

How should such progress be made? For whom are results most important? How will new assessments position us better for the future? These are the fundamental topics of the Gordon Commission deliberations.

Much of the public debate has been directed at the Common Core State Standards initiative (National Governors Association Center for Best Practices, Council of Chief State School Officers, 2010). Misinformed reactions should be separated from legitimate debates concerning (a) reform in educational standards and testing practices, (b) the evaluation of teachers, and (c) the high-quality measurement and uses of student achievement results. Although these three points are related, they also address very distinct areas. Much of what is reflected in the newly adopted standards, CCSS or State developed, should be applauded and vigorously embraced. On target are their emphases on deeper learning, that is, thinking, problem solving, teamwork, reflection, and communication, as these mental abilities are used in and outside of school (National Research Council, 2012). These learning skills will be integrated with the usual content of school learning, such as conceptual understanding of both familiar and modern subject knowledge and skills. Some content will need substantial reorganization to take into account the ever-changing understanding, as knowledge both grows and modifies. Well-clarified and organized content, linked with deeper learning skills will help learners succeed, and States and localities are developing their own approaches to teaching. As we proceed further into the 21st century, our young people's personal development, their participation in the labor force and in most forms of political and social life will depend on their command of these abilities. The values reflected in the newly developed standards deserve our enthusiastic support as do the newly released Next Generation Science Standards (National Research Council, 2013).

For the past three years, we served as chairperson and executive council member of the Gordon Commission. The Commission consisted of some of the most able scholars and thought leaders in the nation. Among the several issues debated by the Commission, the topic given special attention in the debate centered on the CCSS and the uses of assessments associated with them. Most members of the Commission welcome the direction in which the Standards guide us. Few of us are completely comfortable with the idea that assessments are principally for the purpose of accountability. Instead, as presaged above, assessments being developed to measure student achievement of the Standards must serve pedagogy well, a challenge taken on by State Consortia and individual districts and States.

Commission members discussed the possibility that modern measurement science provides us with a wide variety of alternatives to the heavy dependence on traditional standardized tests, and that the nation would be well advised to explore new assessment strategies. The research and development agenda in the measurement sciences, within a modest bit of time, is capable of providing models that show great potential for effectiveness. In this transition period we are aware that some innovation has begun.

Now, in addition to the research and development work underway in the field, and from very different perspectives, members of the Commission have raised a variety of issues that deserve serious examination as we consider the changing relationships between assessment, teaching, and learning. The determination of education and education-assessment policy for the future needs to be informed by some of these issues. As examples, the Commission put forth such ideas as the following:

The major function of assessment in education should be to inform and improve teaching and learning. If we buy into that assertion, the current strength of measurement scienceits claim of precise measurement of developed abilitymay be impeding students development of their own capacity to analyze, document, appraise, and understand the processes of learning. From this perspective, governments role should be to support understanding and improving pedagogical interventions by teachers and students instead of leveraging substantial efforts to monitor outcomes and penalize or reward those results. We must study and improve the processes of teaching and especially of learning in addition to the reporting the status of achievement.

Attributions, contexts, perspectives, and situations so greatly influence human behavior that these correlates of human performance must be fully represented integrally in educational assessment. This suggestion implies that the validity of data from any significant test may be largely dependent upon these contextual factors. Progress in psychometrics suggests greater contextualized interpretation as well as objectivity, reliability and validity.

Traditionally, we have placed an emphasis on student mastery of specific subject matter content. We suggest that the far more important effort is helping the learner develop the mental abilities and capacities that are both integrated with and the byproducts of one's having mastered such subject matter. The focus of the Standards is on such underlying mental abilitieslogical reasoning, understanding cause and effect, organizing knowledge, and problem solvingalong with mastery of key subject-matter knowledge and principles. Our colleagues advise that questions related to transfer in learning are involved here and can continue to be addressed as we refine the targets of assessment. Transfer may include subject matter, but also involves changes in the situation or context through which students demonstrate their learning. Transfer tasks are important for two reasons: first, to determine the students ability to draw from his or her patterns of learning and apply them to new settings; second, to synthesize learning of different tasks, combining elements in order to solve new, innovative, and unforeseen problems. These goals not only reduce the likelihood of cheating or mindless test practice but also challenge traditional notions concerning the primacy of subject-matter mastery as the major goal of education.

Dropped-in-from-the-sky standalone tests have not produced timely, appropriate, and adequate evidence to help us draw the inferences that bear on the instructional decisions we must make. The Commission considered the advisability of distributed differentiated systems of assessment, some that occur throughout teaching and learning during the school year. In plans for new assessments, these could be integrated with timely feedback to learners, teachers, parents, and administrators. Data from these systems of process and status-sensitive assessments could thus be used to assist teachers to adapt instruction, to designs of new pedagogical interventions, and even to make administrative decisions. In addition to emerging assessments, this approach requires renewed and broad confidence in teacher professionalism.

Some prevailing measurement models are anchored to our traditional commitment to meritocratic values rather than in pursuit of democratic opportunity and do not appear to be sufficiently interesting or inclusive to students of different backgrounds. It is true that there are questions about the compatibility of functional meritocracy in the service of the democratization in societies where opportunity is unequally distributed. This set of ideas may confuse purposes of assessments, those focused on supporting learning, i.e., criterion-referenced measures, and those reported in terms of competitive rank, i.e., norm-referenced reports. A continued faith in meritocracy complicates decisions to prioritize capacity building as the central function in assessment for education. Nonetheless, meritocracy should not be shorn from schools if it is in wide use to advance careers and further education of all. The mix must be much more subtle and depends on the centrality of teaching and learning.

In the course of our deliberations, members of the Commission realized that measurement science has not been asleep at the wheel, but has been concerned with much of the advanced thinking around such issues. In fact, there has been a strong movement to integrate the heretofore separate disciplines of measurement and student learning, so that they are focused on the same desired processes. However, a preoccupation in public policy circles with the important problem of accountability has stalled public education. This focus privileges the measurement of status and neglects the analysis, appraisal, and documentation of the processes of teaching and learning along with the contextual correlates of their effectiveness. This problem area begs for attention as we continue to struggle with the juxtaposition of such values in education as diversity, equity, and excellence in a democratic society. Why is this so? It may be simply a matter of money, as traditional tests are less expensive, and statehouses are under fiscal pressures. The reliance on old, less expensive approaches that undermine high-quality pedagogy occurs against a backdrop in which biennial political campaigns in a given states elections may cost more than the states total annual educational investment.

The above five issues raised are only some that are deemed by the Commission to require more serious attention as the field of education seeks new and more effective approaches to what the Commission calls assessment FOR education. Persons interested in the reports of the Commission, as well as the trove of rich intellectual capital generated by Commission members, may find them at http://www.gordoncommission.org. As is suggested in the Commissions Policy Statement, appropriate forward movement in the assessment of and FOR education may require considerably more conceptual and empirical inquiry before new policy decisions are made and professional instrumentation and practice regulations are implemented.

MISSED OPPORTUNITY IN EDUCATIONAL TESTING: WE MAY BE IN THE WRONG DEBATE

Some of us observed the coalescing of conservative and liberal political forces in the provisions of the No Child Left Behind (2002) education legislation, where the law emphasized greater use of standardized testing for increased accountability in education. There was rare disagreement that all students should be expected to learn; more debate that all would be tested; and some controversy the data from these tests would be disaggregated to reveal disparities in academic achievement among children of different social and economic circumstances. Many welcomed this collaboration in the interest of social justice. What we paid insufficient attention to was the possibility that some forces may have been interested in demonstrating the limitations of the public school in the interest of the privatization of the lucrative education enterprise in this country. Documenting the failures of the public school, by using "objective" standardized tests, could provide the scientific evidence needed to reduce support for public education.

If there is any validity to this speculation, it is not so strange that the tests used to measure learning and school effectiveness have not been much investigated to determine their susceptibility to instructional change. If the capability of tests to measure students capacities at low levels of success on a task depends more on out-of-school background and motivation than classroom learning, it would be logical that the standardized tests could not legitimately function as the evidence for sanctions in a program of accountability. Should tests be shown to exhibit instructional sensitivity, we would be more comfortable with the use of assessment for accountability as national education policy.

We may have been much too passive in our acceptance of data from educational testing; we may have been too trusting of statistics that attempt to prove the tests credibility. In medicine, we use test data primarily for diagnosis and to monitor progress, but with growing emphasis on accountability. In education we use test data to select, place, and hold accountable. Modern measurement science is capable of far more. We have models for diagnosis, prediction, instructional sensitivity, and fairness, and these models can be applied to tests focused on support of learning.

In addition, let us not forget that the best instruction relies on a teachers understanding of where a student is on a learning path, the teachers ability to probe for misunderstandings and to provide guidance toward next steps, a process now termed formative assessment. Although long included in the definition of good instruction, in more recent years this process has been extracted from the rhythm of learning and given a special place of its own. The emphasis on testing and assessment has given rise to the culling of formative assessment from the rest of the instructional and learning process, so that finding out how students are doing and helping them in the course of instruction have greater policy value in a testing society. Whether this is an advance in the minds of teachers is not known, but in the reports of the Commission it was clearly asserted that Assessment is best structured as a coordinated system focused on the collection of evidence . . . that can be used to inform and improve the processes and outcomes of teaching and learning (The Gordon Commission on the Future of Assessment in Education, 2013).

The Commission's focus on process as well as product is not new but renewed. In an enormously rich paper on formative assessment prepared for the Commission, Robert Calfee (Education Week, 2014) quotes Hartels (2013) contribution to the Gordon Commission as presenting a more detailed picture of the possibilities of formative assessment to foster learning:

[In my vision of the future of schooling], classroom assessment will be truly integrated with instruction, based on student pursuits that are educationally useful and intrinsically meaningful in the classroom context. Observations of the processes as well as the products of student learning activities will inform assessment inferences. Records of students actions will be captured and analyzed to support high-level inferences about their reasoning and expertise.

Calfee also quotes McManus: Formative assessment is a process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve student achievement of intended outcomes (McManus, 2008, p. 5).

He expands the definition with several helpful comments:

1.

Formative assessment is a process rather than a particular kind of assessment . . . There is no such thing as a formative test.

2.

The formative assessment process involves both teachers and students . . . both of whom must be actively involved in the process of improving learning.

3.

Learning progressions [or a plan for how learners should move from where they are to where they need to be] provide teachers with the big picture of what students need to learn, along with sufficient detail for planning instruction to meet short-term goals. [These progressions are not unlike scope and sequence guidance given to teachers in earlier times]

4.

Teachers must provide [students] the criteria by which learning will be assessed . . . using readily understood language, and realistic examples that meet and do not meet the criteria to advance significantly and to reshape previous conceptions. (McManus, 2008)

This description has proved to be a major step forward, because it captured the main themes and focused on the distinctive features of instruction, now labeled as formative assessment. Clearly, measurement science is capable of serving teaching and learning better than we have. We have permitted the nation to focus on less than the most productive purposes of assessment in education. Using punishment and reward approach in accountability is proving to be disastrous as national education policy. A casual review indicates that measurement science, using existing tools can mount an active program of Assessment FOR Education (see http://www.gordonCommission.org).

The imperative for a modified approach to assessmentone that focuses on assessment of the processes of teaching and learning, embedded in instruction and that enables cognitive developmentis demanded by the rapidly evolving world in which students are expected to function. In order to succeed in todays marketplaceand indeed, to have a fully actualized lifestudents need to develop skills far beyond subject-matter mastery. The challenge is how to stimulate those teachers who still need to embrace the goals, mission, skillsets, and knowledge for students to succeed. How are they best recruited? How are they sustained and valued? How do they get credit for more sophisticated, differentiated instructional processes? As a backdrop for its call to reimagine assessment in the service of education, the Commission first undertook to define what the goals and objectives of education in this century might be; in short, we asked, what does it mean to be an educated person in this century?

WHAT WILL IT MEAN TO BE AN EDUCATED PERSON IN THE MID-21ST CENTURY?

In speculating what it will mean to be an educated person in the middle of the 21st century, we begin by taking stock of potentially far-reaching behavioral changes resulting from new kinds of social communication. Extreme personalization and fragmentary communication would appear to be antithetical to what quality education has traditionally stood for, but the consequences of a shift toward greater person-centeredness seem increasing at this time, spurred by the lure of inferences made from Big Data, or the accumulation and interpretation of micro behaviors. Longer media forms and shorter attention spans may be different manifestations of the same trenda declining interest to do sustained, integrative thinking. Text is gradually being replaced by transmedia options. Improved media design may make it easier to recover a line of thought. Ultimately our sights are on larger educational challenges: to heighten metacognitive awareness; to help learners manage their own thinking in the midst of distractions from devices; to help students keep cognitive purposes in mind; and to evaluate their current mental states against them. These are but brief examples of the growing educational challenge to promote *sustained work with ideas* and to promote *work with other people*, that are also challenges for educational technology design.

Being educated traditionally had two aspects: academic knowledge and skills and personal qualities like character or intellect. Rapid growth of knowledge and the general uncertainty about what the future will demand has raised doubts about the perpetual value of some current knowledge and skills. The changing corpus of academic knowledge also detracts from teachers comfort levels with what they know. And they are also learners. But in every other profession, the world is changing and the practitioner is required to change almost as fast. The scope of the term educated may be narrowly limited to testable knowledge and skills or expanded to include everything that constitutes being a good citizen, but we do not believe it is wise to burden the term with every desirable human quality. Better to acknowledge that there is more to being a good person than being well educated, and being educated can take many forms. Eliminating moral perfection from the definition of educated does not mean eliminating moral reasoning but rather frees teachers to consider constructively the role it plays in cognitive processes, alongside knowledge, skills, and aptitudes.

Being knowledgeable, or having the ability to retain knowledge in the individual nervous system, has come under scrutiny. A legitimate sub question to our broader query is what will it mean to be a knowledgeable person in the mid-21st century? How much of knowledge at any one point should be uniform or overlapping? How do social and intrapersonal knowledge, balance, and focus play into our definitions? Where does the skill of finding relevant and useful knowledge reside?

During its deliberations, the Commission found ourselves pondering what C. Wright Mills called "what if" questions, summarized below.

What if dependence on memory were reduced by the use of great stores of knowledge and the capacity to find or generate needed information by association, augmented by the cybernetic capacity to access needed information from electronic storehouses? We would not reward the use of mental energy on the mastery of more and more knowledge and technique, but we would privilege the development and assessment of accumulated knowledge and electronic retrieval and association skills. In such a world the targets of assessment will quickly become the breadth and depth of one's knowledge; one's capacities for generation using relational and structural cues; and one's ability to carefully adjudicate relationships rather than the extent of one's short- or long-term memory.

What if knowledge and technique did not have to be stored in the human brain? Could brainpower be freed for imagination less constrained by preexisting knowledge? We know that prior knowledge enables, but also limits, human thought. Thinking tends to be guided by, if not restricted by, schemata that have been previously developed. The brain cannot hold in active memory all extant schemata. Humans tend to select from those that we can quickly access. The schema and associations that surface tend to enable and/or preclude access to others. Often the existence of a familiar schema prevents the creation of new schemata. The mindfulness community is trying to determine how appropriate assessment embraces both the constrained and flexible mind.

Logical reasoning is almost universally regarded as a characteristic of intellective competence. It appears to require conceptual generation and computational facility in the analysis of possible relationships between concepts. What if we achieved a division of labor in which the human mind continued to generate reconceptualization and the computer did the computational analysis? What if the mind were freed to do the work of relational generation and adjudication? Such speculation led us to think about future visions for assessment in which we could anticipate movement:

From measurement of status to appraisal and documentation of processes of teaching and learning;

From decontextualization to the recognition of phenomena in relation to specific attributions, contexts, and perspectives;

From a focus on the mechanisms by which things work to understandings of the meanings of the phenomenon in question;

From standardization of responses to systematic strategies in original probes so as to enable comparisons across persons, populations and situations;

From mastery of knowledge and technique to command of the capacity for meta-cognition;

From attention to the mastery of knowledge and skill to greater attention to the learner's command of the mental processes that Snow (ed. Cronbach, 2002) and Martinez (2000) thought are the larger products of having studied and learned any content.

Assessing development, which necessarily must be done over a time span and which typically considers global traits and dispositions, obviously calls for looking beyond testing programs as we know them today. Teachers are essential to the assessment process, not only as observers or evaluators but also as enablers of the kind of activity that is intended to be assessed. Such activities can be extended to many developmental objectives, whether formulated as skills, habits of mind, intellectual virtues, attitudes, or dispositions. In reviewing the five core competencies we have discussed, we find that there are well-recognized varieties of *knowledge creation* that fall within the demonstrable capabilities of the young; that the ability to work with *abstractions* may have considerable generality and is dependent on formal education; that there will be effective ways to teach and test widely generalizable *systems thinking abilities*; that regardless of the benefits and detractions of new media, *cognitive persistence* can be improved by helping support students reading of long and complex text and assessing their understanding; and that *collective cognitive responsibility* will need to be assessed at the group level, likely with supportive technology. We are also interested in students ability to integrate and sustain an interest, perhaps signified by attainments of qualifications or badges rather than by test scores (Baker, 2007). The kind of change that would make technology truly supportive of educating the mid-21st century person is a change that places human development goals as central and knowledge, skill, attitude, and other goals as instrumental.

School reform should not proceed on the basis of folk theories of learning, cognition, and action, largely oblivious to the past 35 years of relevant scientific and pedagogical advances. Fortunately, the Partnership for Twenty-First Century Skills has evolved toward designing intellectually enhanced approach to school subjects, a definite step toward broader human development objectives derived from a conception of an educated person in the mid-21st century. We argue that the five competencies we have highlighted are the real 21st century skills needed for a productive and satisfying life in an innovation-driven society.

What might such a society look like? 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 Behrens & DiCerbo, 2013). Correspondingly, technology will drive options in education and testing far more strongly than ever.

TECHNOLOGY AND POTENTIAL INFLUENCES ON TESTING FUTURES

Without apology, technology is now part and parcel of testing and is likely to grow both in its obvious and less salient functions. Because the growth and diversity of technology have accelerated beyond our expectations, we can only speculate on its role in education, learning, and testing in the future. Technology has presented us with options, hundreds of new apps or device a week, that captivates us, whether we knew we needed them or not. Yet, even schools, institutions historically slow to adapt to change, have been acquiring new technology. Technology has usually been adopted by an analogy to push, where something new is discovered and we find a way or many ways to use it. In contrast to technology push, schools have also been driven by requirements and slow to acknowledge opportunities that do not fit it into their predesigned structures. As a result, schools heretofore have been reluctant or only partial users of technology. But because society has been immersed in technology in many ways, from sensors in clothes and refrigerators to interactive novels, the schools now must begin to see technology as a requirement, needed to keep at least part of the attention of its student audience. So what are the big issues here and how do they connect to the major premises of the Gordon Commission discussion of standards and assessments? A major point, of course, is that the teacher and frequently the parents are no longer gatekeepers of knowledge. Students can access and explore content on the web by themselves. Furthermore, independent access to the web, without mediation by adults, creates a broader potential assessment domain. This access changes the very content of knowledge available to students. One obvious use is to teach students, in the terms above, intellective capability to learn to discern useful from irrelevant, to avoid distraction, to use multimedia to construct arguments, design innovative projects, learn about and with other people, and do so in a mindful way. But what is the implication of uniformity of knowledge to be tested, when knowledge to be learned may become increasingly diverse, or, if you like, personalized.

Un-curated access to the web leads as we have recently seen to new potential threats by technology: to individual privacy, the capacity for misdirection and untruths, the mysteries of who is at the other end of the system, and, through amazing graphic, challenges, puzzles, and stories, the lure of escape. Clearly teachers of this generation of students also must deal with the expectation of personalization, of the lack of self-control in display and drawing attention to self, exploring appropriate and inappropriate topics, and the insatiable development of digital social contact, often at the expense of real human contact. So in addition to managing cyber bullying, designing robots, and surfing the web while reading a book, teachers need to be able to help develop teams and groups and respect for all people and acquire higher level learning, motivation and ability to sustain interest in technological contexts. Will tests be able to support teachers in these tasks?

What is on the immediate horizon of assessment using technology? We can consider test or task design, administration, reporting, validation, and other topics. In the very short term, today and tomorrow schools are trying out computer-adaptive assessments. It is worth noting that computer-adapted testing will now be a regular part of many assessments of the CCSS; even though this testing innovation has been used elsewhere for about 40 years, it is novel for many schools. They may nonetheless confront technical problems.

Current technology-based assessments are likely to change. As costs for simulation design, new response modes, graphics, and innovation continue to drop, and emphasis in assessment FOR learning continues to rise, we should begin to see in assessments marked technological innovations. The computer adaptive tests, noted above may expand from difficulty level to including other options as a basis for adaptation. As technology capabilities widen other options, such as content, task framing, language demands, and visual complexity could be used. Second, at the present time, technology is used in not only the administration and scoring of selected responses, but for scoring short answers or essays administered online. To score or evaluate these essays, early forms of scoring using artificial intelligence have been brought into play; again, as technical capabilities expand and costs decrease, more advanced approaches may be able to be applied. Third, there are at present and beginning to be widely available, formative assessments integrated in video games and simulations (Baker & Delacruz 2013; Chung, 2014). For instance, Quellmalz et al. (2013) and Linn and Eylon (2011) have developed sophisticated science simulations and more are in the works. In games, game levels and in-game performance serve a formative assessment function, along with outcome and transfer measures, so long as game specification represent the educational goals intended to be learned. As a side effect from design of game formative assessments, Baker and Delacruz (2013) discovered a very comprehensive and inexpensive way of designing performance assessments, the hoped-for, but yet unrealized, linchpin for new complex standards. These approaches, useful in games and validated with student data, make it possible to assess complex learning economically and accurately. In the above case, primary age children learned to manipulate sophisticated physics variables, such as vector addition, gravity, and propulsion, in problem solving settings. Such game or simulation-based assessments, if properly designed, will overcome current limitations of establishing comparability both by design and detailed analysis of students series of responses

(Baker, Chung, Delacruz, Griffin, & Madni, 2013). What is useful about these systems is that they provide instruction (although granted with help and support for teachers) that make it possible for children to learn content and skills unlikely to be ordinarily available to them

What might we expect in five years or 10 as the routine ways in which assessments will operate? What tools on todays horizons will have impact in assessment design? How will new approaches adapt to individuals 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?

Lets posit the following precepts from the Gordon Commission findings: Learning skills will be embedded in content or multidisciplinary problems. Technology will be used to help design, administer, score, store, and report findings to appropriate users. Schools and education systems will not be the only source of assessments. Schools will use business, community, and other resources, both for learning and assessment ideas, to provide authentic tasks for students to work on, and to participate in the judgment of performances, either presented, available through technology or in some combination. Students will make things, not just give answers. And they will be working in teams, with partners in classrooms or in other States or countries.

Because of the proliferation of devices, there will be an expansion of the focus on infrastructure needed to build assessments through technology. For example, there may be a range of theoretical and practical models for assessing problem solving at various levels and for different content. It is possible that these may share common elements, including that domain knowledge is visually represented, that sequences have been developed empirically, and that there are supports available for learners experiencing difficulty. Just as the multiple choice task has become a standard template, we believe that new infrastructure is the place where research and practice may converge to determine the degree assessments can be prepared with strong validity, credibility, and at lower cost. Common infrastructure elements may be used even if surface features or topics of the assessment differ. There are several assessment design supportsboth methods and contentwhich have emerged in recent years. One such support has been the use of ontologies. At the present time, there have been approaches that use detailed representations of expectations, either translating the Common Core State Standards (National Governors Association Center for Best Practices, Council of Chief State School Officers, 2010) into network representations or making relatively simplified maps to assist teachers using a dashboard metaphor to monitor individual students progress. While recent efforts have had difficulty (Weiss, Bloom, & Brock, 2013), the level and detail of design and reporting is easily tailored for use by teachers and by students. Much of this infrastructure is drawn from computer science, specifically multilayered ontologies of content domains. These ontologies, used for assessment or learning design, can be rapidly changed as new knowledge emerges, and with a variety of assets can allow rapid and less costly design (Iseli, 2011; Wimalasuriya & Dou, 2010). Ontologies can be created from combined expert knowledge, such as represented in standards, through crowdsourcing or other online techniques or by querying individuals. In addition, new natural language processing systems permit the addition of content from extant text sources to ontologies. In the current environment, such ontologies provide a direct and strong method to use to assure the transparency and 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 are a key asset to be sampled. Ontologies have been developed for subject matter content, i.e., the CCSS, and for skill learning, e.g., problem solving. What is underway currently and should be standard by 2018 is the development of descriptive characteristics or settings and situations for assessments. Then targeted content or integrated skills can be embedded and actualized by using these sources of existing infrastructure. If the field is successful, then cost will not continue to impede assessment innovation and broad-scale use.

In order to keep the chapter from delving too far into the technical side of measurement, we have been using common language and real settings. But there are scholars and measurement practitioners who use various models to help structure their designs. That models will change is a given; how they will begin to use technology more efficiently is still open. Assessment design models vary by their level of specificity. Mislevy and his colleagues created a model that followed the general lines of artificial intelligence Artificial Intelligence 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 (Bennett, 2013; Baker et al., 2014; 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 Gagnes work in task analysis (1965), has a cognitive orientation, which is both appealing but highly idiosyncratic in practice (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 the earlier mentioned transparency for designers, users, and examinees, it will be possible to have a database of features part of the task set, much like vendors now have databases of completed items. These new items or tasks could be developed from elements residing in servers, triggered by the detailed data on student performance.

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. Using Boy and Girl Scout 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 wider currency outside of school. Currently the badge environment in the United States 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 those of the regular teacher, a way to accredit such individuals, to make complex task 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 of validity and 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 résumés 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 and public. Clearly, early computer support for badges would need to be supplanted by more sophisticated models that assisted learners in their acquisition of required skills and identified resources for assistance, feedback, certification, supported by validity evidence.

The Big Data, or analytics movement is another area of great interest. Its use on the web began largely as a descriptive venture, but evermore sophisticated analyses of massive data, called data mining are in play, in order to make predictions of consumer activity by subdividing user groups based on websites visited, time spent on site pages, 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 purchase profiles. Clearly web analytics and targeted marketing now depend on data and an expert base. Popovic (Center for Game Science, 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 a group on the web can display collective problem solving. Chung (2013) and Roberts (2014) have created with their teams a set of learning analytics, a top-down analogy for data mining. Using measurement-oriented computational models, these bottom-up and top-down approaches may be combined to suggest best predictions, reasons for misunderstandings, and likely content and skills the learner has acquired and may be able to skip.

In the future, we also expect that other human capacities, including affective states, may be routinely measured. The sensitivity of having access to such data should not be underestimated, and even 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. Pentlands *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 interactive learning, At a future point, these technologies may be useful to valid observable measures of learning.

There is also 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) and how to apply learning to 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 clearly diverging definitions of deeper learning, 21st century skills and come up with an explicit first set to be included systematically in education, used in a way to guide teaching practice. . 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, and more systematic in their searches and more evaluative in their judgments about the information they could value and include in their searches and applications.   Certainly, students will have to learn to live and interact with a wider variety of people, people from next door and from 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 policy more likely to succeed.

Starting from where we are and the benchmark of international performance, knowledge expansion, economic instability, and technology growth portends many possible assessment futures. We have excluded numerous important variables such as demography, 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 and State school authorities, parents, multinational organizations, or corporations with interests in maintaining well-prepared workforces? A combination? If the rules change, how will the transformation take place?

What is the role of self-assessment? Is all assessment directed to and by institutions?

How will local 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 testing, 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 by themselves be less relevant than the evaluation of the instructional experience, including the assessment? What criteria should be used to assure quality of the hybrid?

Will personalization trump standardized indicators of performance?

Will we lose too much art and humanity by fully automating assessment and instruction? Do the same automation processes need to be applied to all subject areas or just to 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 continuing issues such as privacy and security evolve? Will the public space of technology affect the entire process?

Will the public demand affirmative consent for any secondary use of data, as is now the case in the European Union?

Can assessment on technology create and enhance our own personal narratives?

Each of our dreams and fears are unique, and our bets on the future waver depending upon fluctuating optimism about individualism, a yearning for wise shared values, and observations of superficial divisions. What we should be looking toward is finding strategies by which the future and learning get better through technology and its worldwide distribution.

TOWARD EDUCATION FOR ASSESSMENT AND FINAL THOUGHTS

Measurement science has attracted some of our most able behavioral scientists, which have elevated the field of measurement to an important position among the social sciences. With a very sharp focus on the measurement of the status of a rather narrow range of developed academic abilities, this field has come to be associated with the assessment OF education. But the measurement sciences can do more. Building on an excellent capacity to measure indicators of ability, we call for equal attention to be given to the analysis of the learning and teaching processes by which intellective abilities are developed and to the use of measurement science in the cultivation of intellective competence. Assessment can be instructive. We see the emerging capacities of measurement science, combined with new developments in the sciences and the electronic technologies upon which education and its measurement can now ride, as enabling the more sophisticated quantitative and qualitative analyses necessary to better understanding the activities of learning and teaching persons. We also see a greater opportunity to measure learning deeply and in breadth, and to do so in a way that captivates students interest. With such understanding measurement science can be used to supplement its rich contributions to the assessment OF education, with even more powerful contributions to assessment FOR the improvement of education.

References

Baker, E. L. (2007). The end(s) of testing (2007 AERA Presidential Address). *Educational Researcher*, *36*(6), 309317)

Baker, E. L., & Delacruz, G. (2013). *Blended learning: Integrating instruction and assessment*. Paper presented at the Council of Chief State School Officers 2013 National Conference on Student Assessment, for session A Vital Goal: Blending Assessment with Instruction, National Harbor, MD, USA.

Behrens, J. T., & DiCerbo, K. E. (2013). Technological implications for assessment ecosystems: Opportunities for digital technology to advance assessment. In E. W. Gordon (Chair), *The Gordon*

Commission on the Future of Assessment in Education (Final Report). *To Assess, to Teach, to Learn: A Vision for the Future of Assessment* (Technical Report). Princeton, NJ: Educational Testing Service.

Bennett, R. E. (2013). Preparing for the future: What educational assessment must do. In E. W. Gordon (Chair), *The Gordon Commission on the Future of Assessment in Education* (Final Report). *To Assess, to Teach, to Learn: A Vision for the Future of Assessment* (Technical Report). Princeton, NJ: Educational Testing Service.

Baker, E., Cai L., Choi, K., & Buschang, R. (2014, April). *CRESST functional validity model: Deriving formative and summative information from common core assessments*. Presentation at the annual meeting of the American Educational Research Association, symposium 46.010 Innovative Validity Approaches for High-Quality Assessments: An Interaction, Philadelphia, PA, USA.

Center for Game Science. (2008). Foldit. Retrieved from http://centerforgamescience.org/portfolio/foldit/

Chung, G. K. W. K. (2013). Toward the relational management of educational measurement data. In E. W. Gordon (Chair), *The Gordon Commission on the Future of Assessment in Education* (Final Report), *To Assess, to Teach, to Learn: A Vision for the Future of Assessment* (Technical Report). Princeton, NJ: Educational Testing Service.

Chung, G. K. W. K. (2014, April). *Game development and design: Blood, sweat, and tears*. Presentation at the 2014 Center for Advanced Technology in Schools (CATS) conference, Redondo Beach, CA,

Clark, R. E. (2003). Fostering the work motivation of teams and individuals. *Performance Improvement*, *42*(3), 2129.

Cronbach, L. J. (Ed.). (2002). *Remaking the concept of aptitude: Extending the legacy of Richard E. Snow*. Mahwah, NJ:

Erlbaum.

Education Week. (2014, March 31). *Measurement science can do more for education!* Retrieved from http://blogs.edweek.org/edweek/assessing_the_assessments/2014/03/measurement_science_can_do_more_for_education.html

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

The Gordon Commission on the Future of Assessment in Education. (Edmund W. Gordon, Chairperson). (2013), *To Assess, to Teach, to Learn: A Vision for the Future of Assessment* (Technical Report, The Gordon Commission Final Report). Princeton, NJ: Author, Educational Testing Service.

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).

Linn, M. C., & Eylon, B.-S. (2011). *Science learning and instruction: Taking advantage of technology to promote knowledge integration*. New York: Routledge.

Martinez, M. E. (2000). *Education as the cultivation of intelligence*. Mahwah, NJ: Erlbaum.

McManus, S. (2008). *Attributes of effective formative assessment*. Retrieved from http://www.ncpublicschools.org/docs/accountability/educators/fastattributes04081.pdfMislevy, 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*, 335374.

National Assessment of Educational Progress (NAEP). (2014). *Are the nations 12th-graders making progress in mathematics and reading?* (NCES 2014-087). Washington, DC: U.S. Department of Education, Institute of Education Sciences. Retrieved from http://nces.ed.gov/nationsreportcard/subject/publications/main2013/pdf/2014087.pdf

National Governors Association Center for Best Practices, Council of Chief State School Officers. (2010). *Common Core State Standards*. Washington, DC: Authors. Retrieved from http://www.corestandards.org/the-standards

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.

National Research Council. (2013). *Next generation science standards: For states, by states*. Washington, DC: The National Academies Press.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 C.F.R. (2002).

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist*, *38*, 14.

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

Quellmalz, E. S., Davenport, J. L., Timms, M. J., DeBoer, G. E., Jordan, K. A., Huang, C.-W., & Buckley, B. C. (2013). Next-generation environments for assessing and promoting complex science learning. *Journal of Educational Psychology*, *105*(4), 11001114.

Roberts, J. (2014, April). *Advances in technology for learning and assessment: Collection, storage, computational analysis and researcher tools for early learning analytics at PBS Kids*. Presentation at the 2014 Center for Advanced Technology in Schools (CATS) conference, Redondo Beach, CA, USA.

Shanker, A. (1988). Reforming the reform movement. *Educational Administration Quarterly*, *24*(4), 366373.

Weiss, M. J., Bloom, H. S., & Brock, T. (2013). *A conceptual framework for studying the sources of variation in program effects* (MDRC Working Papers on Research Methodology). New York: MDRC. Retrieved from http://www.mdrc.org/sites/default/files/a-conceptual_framework_for_studying_the_sources.pdf

Wimalasuriya, D. C., & Dou, D. (2010). Ontology-based information extraction: An introduction and a survey of current approaches. *Journal of Information Science*, *36*(3), 306323.