The Gordon Commission **Final Report**

To Assess, To Teach, To Learn: A Vision for the Future of Assessment

Technical Report

The Gordon Commission
on the Future of Assessment in Education

TO ASSESS, TO TEACH, TO LEARN: A VISION FOR THE FUTURE OF ASSESSMENT

Technical Report

Table of Contents

The Gordon Commission
on the Future of Assessment in Education

INTRODUCTION

The experience of thinking aloud in the presence of highly competent colleagues who have been deliberately selected to be one's advisers and critics is at one and the same time a privilege, a challenge and an inspiration. To have that opportunity at the beginning of one's tenth decade of living, when one has a richly etched tablet on to which to inscribe the meta-productions of that experience is as unique an experience as is likely to befall most of us who try to make a career of scholarship and service. This report is an account of my having had such an experience and my take on what might well be taken from an inquiry into thought and speculation concerning what is likely to be happening in the education enterprise as we move through the 21st-century and, especially, what demands on the field of assessment in education might well be expected.

Educational Testing Service has very graciously enabled me to convene scholars and thought leaders of my choosing to advise and challenge me as I have conducted this inquiry. I was honored when friends at ETS suggested that the initiative be called the Gordon Commission under my Chairpersonship. It is with deep appreciation that I acknowledge a personal debt to Kurt Landgraf, President and CEO at ETS, for his decision to support the Commission and me as we inquire into possibilities for the future of assessment in education in the 21st-century. There is no way in our reporting to adequately reflect the work involved in conceptualizing, planning, orchestrating and implementing the contributions and efforts of the estimated 100 persons who have done the work. It is obvious that it is more than the work of a ninety year old Chairperson. The organizing conceptual and managerial force behind this work is the Gordon Commission's Executive Officer, Paola Heincke. The Gordon Commission carries my name. The Technical Report is my statement endorsed by my Co-chairperson, Professor Jim Pellegrino, but the work of the Gordon Commission has been orchestrated by my Associate, Paola Heincke. The Commission members and I are indebted to her.

The Commission was organized early in 2011, as a virtual study group of thirty members and an additional fifty consultants. These excellent human beings have been involved in various ways and degrees of intensity. I have been promiscuous in my search and tapping of the thinking and scholarship of these people. The identifiable contributions of several scholars are included in this Technical Report and are represented in the four-volume collection of the papers that were written as part of the work of the Gordon Commission Knowledge Synthesis (KSP) Project. This original written work is the substance of my report along with my own commentary on and interpretation of what I have heard and understood as I have tried to use my well aged mind and about seventy years of informed experience in related activity to make sense of and gain some perspective on what is happening in education, what I sense will happen in the field, and to suggest implications for the future of assessment in education. I have concluded that building upon a long and extraordinary history of achievement in the assessment ***OF*** education, the future of assessment in education will likely be found in the emerging interest in and capacity for assessment to serve, inform and improve teaching and learning processes and outcomes. Shall we call that assessment ***FOR*** education in addition to the assessment ***OF*** education?

The Technical Report begins with a reprise of the substantive work of the Commission, reflected in what we have called the Knowledge Synthesis Project. Digests of the several papers and findings drawn from these papers are reported. This work is followed by a summation of Professor Carl Kaestle's essay concerned with the history of assessment in education. Since this history reflects the emphasis in educational measurement that has been placed on the assessment of education, Kaestle's history is complemented by a commentary from the Chairperson on a possible future history, of assessment that is in the service of education. The history and future history of assessment in education introduces two futurist essays, one having to do with shifting epistemologies and changing paradigms, and the second concerns what it will mean to be an educated person in mid-21st-century. Three additional essays are included. They address issues related to shifting epistemologies, ”Postmodern Test Theory;” “Technological Implications for Assessment Ecosystems;” and a vision of “What Educational Assessment Must Do.” These essays are followed by a vision for the future by the Chairman of the Commission, “To Assess, to Teach and to Learn: A Vision for the Future of Assessment,” in which I seek to capture ideas and perspectives to which I was exposed in my work with the Gordon Commission. This vision of the future introduces the Recommendations from the Commission, followed by technical information concerning the Gordon Commission and its operations.

I must acknowledge that the work of the Gordon Commission is incomplete. We have initiated discussion and inquiry into possible futures for education and its assessment as we move through the 21st-century. We have identified several of the issues that we feel must be addressed as we proceed along this course. We have commissioned several scholars to develop knowledge and thought synthesis papers concerning these issues. We have not had an opportunity to debate and deliberate concerning the findings that are grounded in this work. It will be left to some subsequent forum to seek consensus concerning the meaning of this information for recommended action in the 21st-century. This new century will be a period for which we cannot make precise predictions, but we can make the prediction that things will continue to change. The changes this time will be as consequential for human societies as were the introductions of the printing press, mass communication and industrialization. As best as we can tell the changes will involve:

- The nature of knowledge and human access to it;
- The quantity and quality of scientific information, its digitization and its electronic exchange;
- The nature and control of the political economies of the nations of the world;
- The nature of human social intercourse and the distribution of world populations;

All of these changes will be occurring concurrently. To this dialectical predicament we bring a philosophy of science that rests on the assumptions of the availability of universal principles, consistency and fixity, orderly relations between phenomena, reliability, validity, and veridicality. Some of us members of the Commission, as do a growing number of scholars, believe that some of these values may be challenged by or may require some accommodation in the light of changing ways of thinking about the realities of the future we are beginning to envision.

Already we see signs of conflict and contradiction between many traditional notions and respected practices in assessment, teaching, and learning, on one hand, and on the other, knowledge and thought that are emerging from new developments in science, technology and scientific imagination. The Gordon Commission found itself struggling with a set of paradoxes similar to those faced by Columbus and Magellan, i.e., navigating a world that was known to be flat at a time when the evidence was beginning to indicate that the world is round. How do we operate in a system that we have come to know from a positivist science, but are beginning to understand will require a contextualist and relativist science? From this growing sense of chaos, the members of the Gordon Commission have been trying to make sensible judgments and speculations concerning the future of assessment in education.

Edmund W. Gordon

Edmund W. Gordon
Chairperson,
Gordon Commission on the Future of Assessment in Education

TO ASSESS, TO TEACH, TO LEARN: A VISION FOR THE FUTURE OF ASSESSMENT

EXECUTIVE SUMMARY[1]

1. Critical Issues for the Future of Assessment in Education

In the initial meeting of the Gordon Commission, attention turned to questions having to do with why we assess, what we assess and how we assess in education now and in the future. The members of the Commission quickly agreed that the answers to these questions should form the context for our inquiry into the future of assessment. One of the initial activities of the Gordon Commission involved the identification of what commissioners agreed were the most critical issues facing the field. It was thought that the encirclement of extant knowledge and thought concerning these issues should inform the work of the Gordon Commission as it inquired into the current state of assessment in education, the best of extant theory and practice, and our understanding of the changing nature of education and its assessment in the present and anticipated future.

This decision led to the conduct of the central activity of the Gordon Commission that has been referred to as the Knowledge Synthesis Project. This initiative consisted of the commissioning of 25 reviews of extant knowledge and thought concerning the issues that were identified as most important. The papers that resulted from this work are listed in this report. These papers will be published in a four volumes series, *Perspectives on the Future of Assessment in Education*. Under the guidance of our two senior research associates, Rochelle Michelle, PhD, and Ernest Morell, PhD, these several papers written specially for the Gordon Commission were subjected to analysis and digest by six emerging scholars who served as pre and post-doctoral Commission Fellows.

[1] This Executive Summary was prepared by Paola Heincke, Executive Officer of the Gordon Commission on the Future of Assessment in Education, based on the content of the Technical Report of the Commission "To Assess, to Teach, to Learn: A Vision for The Future of Assessment" http://www.gordoncommission.org/publications_reports.html

Developing Perspectives on Assessment

The papers contained within this section (Kaestle, 2012; Meroe, 2012; Varenne, 2012; Mendoza-Denton, 2012; Dixon-Román & Gergen, 2012; and Gergen & Dixon-Román, 2012; Torre & Sampson, 2012; Bennett, 2012) all provide varying views on the historical context for assessment, ranging from testing policies to measurement models used in testing.

Accountability and Validity Frameworks

The papers within this section (Linn, 2012; Mislevy, 2012; Gorin, 2012; and Ho, 2012) discuss the evolving uses of tests and the need to consider assessment frameworks that take into consideration the current and potential uses of test in the context of the teaching, learning, and assessment process. In addition, these papers challenge the testing industry to develop assessment systems that can capture evidence of student learning at multiple time points, from different sources (i.e., inside and outside of school settings), different types (i.e., quantitative and qualitative), and that allow for the demonstration of student learning in different ways.

Beyond the Basics

While current large scale, standardized tests focus on the basic skills of reading, writing, and mathematics, and to a lesser degree science and history, the next set of papers (Bereiter & Scardamalia, 2012; Cauce & Gordon, 2012; Armour-Thomas & Gordon, 2012; and Baker, 2012) call for a movement to go beyond these basics and consider a wider range of competencies. In addition, these papers support a more integrated approach for instruction, curriculum, and assessment that support student learning and allow students to move beyond the basics that are learned and transfer that knowledge to other contexts beyond the one in which the original knowledge was learned. These papers also highlight the importance of collaboration and acknowledge the varying social contexts in which students learn.

Lessons Learned from Testing Special Populations

While the papers within this section, (Hakuta, 2012; Thurlow, 2012; and Boykin, 2012) address specific populations of students (i.e., English-language learners and students with disabilities), their view of assessment questions the current way in which groups are identified to receive alternate assessments or receive accommodations in testing. The papers consider how some of the accommodations may be helpful to learners beyond those that have been identified as having a disability (e.g., universal design) or those who may be English language learners (e.g., bilingual class for English-language learners and native speakers of English).

Technology as a Tool to Advance Assessment

The papers within this section (Hill, 2012; Chung, 2012; and Behrens & DiCerbo, 2012) highlight how developments in technology allow for the development of more advanced, more comprehensive assessment systems that can provide varying levels of data to inform the teaching, learning, and assessment process. Specifically, technology will allow for the collection and management of fine-grained data throughout the teaching, learning, and assessment process that can be used to monitor and inform student learning.

2. A History of the Assessment of Education and a Future History of Assessment for Education[2]

The purpose of Kaestle's (2012) essay is to reflect on the development of modern testing practices in a historical context. This can spur ideas on how to shape assessments to fit our 21st-century values. We have a long and distinguished experience with the use of assessment, measurement and testing in the history of education. That history is marked by a heavy emphasis on the assessment of education through testing and the measurement of the status of one's developed ability and achievement. Rich bodies of theory and practice have been established and are currently used in the service of accountability, selection and certification. It was also noted that there are some equity and accountability goals that have been well-served by being able to pinpoint how well individual students or groups of students are doing. Kaestle also acknowledges the power of standardized, multiple-choice tests due to their cost effectiveness and efficiency compared to the more complex, more subjective and higher-level assessments. These positive qualities of standardized, multiple-choices tests stand in the way of the call for authentic and performance-based assessments that challenge existing frameworks.

[2] Abstracted from Kaestle, C., 2012, *Testing Policy in the United States: A Historical Perspective* and amended by Edmund Gordon http://www.gordoncommission.org/rsc/pdfs/kaestle_testing_policy_united_states.pdf

The claim is advanced that some of the embodied perspectives may be outdated and dysfunctional to the needs of education in the 21st-century. The Gordon Commission has embraced a parallel concern as we move through the 21st-century, in which it promotes as a primary emphasis on assessment for education through the collection and interpretation of a variety of forms of evidence in the service of the disconfirmation of inferences drawn to explain, inform and improve teaching and learning processes and outcomes. The future history of assessment in education is projected to be a history in which the best features of assessment of education will be conjoined with emerging features of assessment for education to inform and improve teaching and learning.

3. The Changing Context for Education and its Assessment–Edmund W. Gordon

Increasingly the goals of education reflect the growing concern with encouraging and enabling students to learn how to learn and to learn to continue learning; to become enquiring persons who not only use knowledge but persons who produce and interpret knowledge. The pedagogical challenge will be less concerned with imparting factual knowledge and more concerned with turning learners on to learning and the use of their mental abilities to solve ordinary and novel problems. Reading, wRiting and aRithmetic will continue to be essential skills, but thought leaders in education (Sir Kenneth Robinson is among them) increasingly point to varying combinations of Cs as essential processes in education: Creativity and innovation; Conceptualization and problem solving: Communication and collaboration; and Computer literacy. The Cs are replacing the Rs as the modern ends toward which education is directed. Learning how to think critically and creatively, reason logically, interpret relationally, and to access and create knowledge will be more and more privileged in the 21st-century.

Education and its assessment will have to become capable of capturing aspects of context, perspective and the attributions which come to be assigned to these conditional phenomena. The exactness and precision which have been gained by de-contextualization in the past will be challenged by the situative and existential sensitivities required when contextualism and perspectivism are required for understanding as well as knowing.

Yet, modern social and psychological sciences are pressing us to examine or assess human performance with greater respect for the influence of affective, emotional, situative and social processes. Evidence mounts in support of the fact that these processes influence the character and the quality of human performance, yet they are these instances of objectively documented human performance that are the source of the data of traditional assessments in education. However, assessment in education in the future will have to be more sensitive to subjective phenomena, i.e., to affect, attribution, existential state, emotion, identity, situation, etc., as will also the teaching and learning transactions in which learners are engaged.

Pressure mounts from the profession and the practicalities of educational praxis for better information to inform intervention prior to the search for better information by which to determine how well we are doing. We have known for more than a century that what we do in education is imprecise; that one model does not fit all; and that much of our intervention is under-analyzed trial and error. We believe that assessment in education can and should inform and improve teaching and learning processes and outcomes, without ignoring the importance of accountability. Whether the two purposes can be served concurrently and by the same assessment instruments and systems is one of the questions to be answered.

Humans will very likely continue to create technologies that make their work easier and that amplify and expand human abilities. Some of these, as with artificial intelligence inventiveness, could change the importance of some of the competencies for which we currently educate or, more likely, will exacerbate the need for other functions that we currently know less about enabling, i.e., agency, disposition, relational adjudication. The human ability-amplifying technologies may make some of our educational tasks easier, but they may also create monumental challenges and opportunities for the people who are responsible for assessing, teaching and learning in some well-orchestrated manner.

4. To Be an Educated Person in the 21st-Century –Carl Bereiter and Marlene Scardamalia

Bereiter and Scardamalia consider the ways in which the intellective demands on educated persons will change in this century. Attention is called to the increasing limitations of knowledge mastery in the absence of knowledgeability in a knowledge-based society. Emphasis is given, however, to the importance of knowledge repertoire and its role as a basis for relating new chunks of knowledge. They emphasize the growing demand for the capacity for adaptability and disposition to exercise agency. Their emphasis on aspects of character seems to have increasing currency. All of these concerns are addressed in the context of tremendous technological advances that will continue to affect the field of education.

Bereiter and Scardamalia (2012) identified five competencies: a) Knowledge creating where students are able to build, amend and create knowledge; b) Working with abstractions where students should be able to work with abstraction and convert them to real-world applications, going from the theoretical to the practical; c) Systems thinking where students should be able to recognize and understand the complexity of the world and consider how to take advantage of the complexity whenever possible; d) Cognitive persistence where students should be able to sustain focus and study in the face of increasing obstacles and distractions; and e) Collective cognitive responsibility where students should be able to engage in collective work that is collaborative.

The authors recognize that as theories of collaborative learning develop, learners should be given instructional space to collaborate, and assessment should adapt so that individual and collaborative contributions to solving problems may be measured and evaluated. They recommend preparing learners to engage in lifelong learning, enabling learners to gain new competencies, while adapting to the accelerating pace of change. Part of this will require education to foster breadth, depth and the ability to navigate diverse ideas, people, and culture. To this end, assessments should be developed that foster creativity. Bereiter and Scardamalia (2012) also call for systems thinking where students are able to both discern usefulness of knowledge and place knowledge within the appropriate context. The authors also recommend developing methods for assessing knowledge creation, work with abstractions, systems thinking, cognitive persistence, and collaborative responsibility.

5. Postmodern Test Theory[3]–Robert Mislevy

Mislevy addresses concerns that are prevalent throughout the work of the Commission relative to the influence of changes in contemporary conceptions of the nature of knowledge and the role of knowledge and knowing in intellective functions. The growing concern for context, perspective and situated meaning that is associated with postmodern talk constitutes a possible challenge to education and to its assessment. This paper and at least two others capture the Commission's concern with the tensions between the positivist traditions that have shaped measurement and the postpositivist, and "neopragmatic postmodernist test theory" that seem to be more appropriate to 21st-century conceptions of assessment in education. The stark contrast between formal and informal assessment arises because to understand students' learning and further guide it, teachers need information intimately connected with what their students are working on,

[3] Postmodern Test Theory (Robert J. Mislevy) is Reprinted with permission from *Transitions in Work and Learning: Implications for Assessment, 1997*, by the National Academy of Sciences, Courtesy of the National Academies Press, Washington, D.C.

and they interpret this evidence in light of everything else they know about their students and their instruction. The power of informal assessment resides in these connections. Good teachers implicitly exploit the principles of cognitive psychology, broadening the universe of discourse to encompass local information and address the local problem at hand. Yet precisely because informal assessments are thus individuated, neither their rationale nor their results are easily communicated beyond the classroom. Standardized tests do communicate efficiently across time and place—but by so constraining the universe of discourse that the messages often have little direct utility in the classroom. The challenge now facing neopragmatic postmodern test theory is to devise assessments that, in various ways, incorporate and balance the strengths of formal and informal assessments by capitalizing on an array of methodological, technological and conceptual developments.

6. Technological Implications for Assessment Ecosystems–John Behrens and Kristen E. DiCerbo

The Behrens and DiCerbo (2012) paper, *Leverage Points for "Natural" Digital Activities in the Assessment of Human Attributes*, describes three core aspects of technological developments that can be used for educational assessment: a) computers can be used to enhance human capabilities given computers' ability to store, process and mine large amounts of fine-grain data from multiple sources; b) the increased use of digital technologies makes it possible to gather new forms of data based on human interaction in digital environments; and c) digital technologies can be used to better visualize the fine-grain data so that observations, patterns and inferences can be made based on the data. These new technologies should allow new insights into student learning using computational methods of storing, analyzing and modeling student data. Behrens and DiCerbo (2012) recommend a reframing of assessment practices from identifying correctness of test questions to capturing a constellation of learning transactions using digital technologies to make inferences about student cognition and learning.

7. Preparing for the Future: What Educational Assessment Must Do –Randy E. Bennett

This essay explores the forms that summative and formative assessments will take and the competencies that they will measure in the future. Education, and the world for which it is preparing students, is changing quickly. Educational assessment will need to keep pace if it is to remain relevant. This paper offered a set of claims for how educational assessment might achieve that critical goal. Many of these claims are ones to which assessment programs have long aspired. However, meeting these claims in the face of an education system that will be digitized, personalized, and possibly gamified will require significantly adapting, and potentially reinventing, educational assessment. Our challenge as a field will be to retain and extend foundational principles, applying them in creative ways to meet the information and decision-making requirements of a dynamic world and the changing education systems that must prepare individuals to thrive in that world.

The author proposes a set of 13 claims about what educational assessment must do if it is to remain relevant and if assessment is to actively and effectively contribute to individual and institutional achievement. The author notes that in order for assessment systems to remain relevant, future educational assessment systems will need to provide trustworthy and actionable summative information for policymakers as well as formative information for teachers and students. He has identified the need for assessments that serve multiple purposes. However, a single test may not be able to meet the needs beyond which the assessment was originally developed. It may be the case that an assessment developed for multiple purposes may not work for any of the identified purposes. According to Bennett, assessment for education must:

- Provide meaningful information
- Satisfy multiple purposes
- Use modern conceptions of competency as a design basis
- Align test and task designs, scoring and interpretation with those modern conceptions
- Adopt modern methods for designing and interpreting complex assessments;
- Account for context

- Design for fairness and accessibility
- Design for positive impact
- Design for engagement
- Incorporate information from multiple sources
- Respect privacy
- Gather and share validity evidence
- Use technology to achieve substantive goals

8. To Assess, To Teach, To Learn: A Vision for the Future of Assessment in Education–Edmund W. Gordon[4]

This section of the Technical Report is bifocal. It provides the insight of Edmund W. Gordon, Chairperson of the Gordon Commission, into the substantive work of the Commission as reflected in 25 essays that were written for the synthesis of knowledge and thought that informed its work. The essays range from several that are concerned with various perspectives on assessment in education and their meanings; problems associated with accountability, reliability and validity as frameworks for assessment; and the notion of assessment as evidential reasoning. In other essays, attention is directed at changing and persistent targets of assessment having to do with just what it is that we assess; lessons learned from assessment in the education of diverse cultural groups and special populations; and the implications of emerging developments in science, technology and scientific imagination for education and its assessment. The assessment enterprise in education will become an educative service concerned with informing and improving teaching and learning, and modeling the adaptive, intellective and learning behaviors that exemplify the intended outcomes of education.

[4] The Author acknowledges with deep appreciation the editorial and research assistance of Emily B. Campbell, Paola C. Heincke and Paola A. Valencia in the preparation of this commentary.

Why do we assess? We assess in order to better understand the people we teach, the processes by which we teach them, the situations in which they learn or fail to do so, and to enhance their intellective character and competence. What then might well be the characteristics of systems of assessment in education that embrace assessment, teaching and learning as privileged processes? Gordon's preferred candidates for assessment capacity and practice by mid-21st-century are:

- A system of inquiring assessment probes embedded in teaching and learning transactions. There are at least three ideas included in this proposal: a) gradual replacement of standalone tests with systems of assessment (multiple and varied assessment opportunities), which are distributed over time and throughout the teaching and learning transaction; b) the integration of assessment probes as instruments of inquiry, instruction and mediation; c) separate responsibility for the use of data drawn from rich descriptions of these transactions for administrative and for student development purposes. Teachers would be enabled to interpret these data diagnostically and prescriptively. Psychometricians would be responsible for distilling from these in vivo learning and teaching transactions data needed for accountability.
- The integration of assessment with teaching and learning will demand a view of assessment as diagnostic inquiry, exploratory mediation, and intensive accountable exchange ("accountable talk" to use Resnick's term). There is a rich history of the use of questioning as a part of instruction. Good teachers know the art of posing questions that stimulate thought (Socratic dialogue) as well as probing for evidence of status or process. Most good teachers do not depend solely on standardized tests to know where their students are and what they need. Whimby (1980) makes extensive use of exploratory mediation through which teacher and student inquiry are used in the search for explication of meaning and processes utilized. In the integration of assessment with teaching and learning, the unique character of each of these processes may be lost, as each serve functions that can be interchanged with the other.
- The unbundling and explication of the cognitive demands of knowledge and technique mastery. What is the cargo of transfer learning? Gordon gives extensive discussion to his concern for the complementarities between the worlds of knowledge and technique on one hand, and developing mental capacities on the other. He also discussed the possibilities for distilling from the items of standardized test clearer indices to the cognitive demands of test items. In this

approach, he recognizes the importance of knowledge content in teaching and learning, but argues that the mastery of such content may be less important than is the achievement of intentional command of the mental abilities that (1), have been developed in the course of the study of this content and (2), are essential to the processing of information represented in the knowledge and technique.

- Modern information technologies afford students access to almost limitless quantities and varieties of information resources. Competence in accessing and utilizing available resources could replace the more traditional privileging of memory store. Assessment and education by mid-21st-century will be capable of documenting and determining the status of one's competence in determining resource need, accessing needed resources, help seeking, and the utilization of these resources.
- Distance learning and the use of epistemic games have already reached epidemic levels among age groups of learners under thirty. Current predictions suggest continued growth in the use of these educative and recreational media. The almost colloquial anticipation is that this genre of electronic digital information exchange carries with it a trove of information that can be used for educational purposes. In the near future such information will be distilled from the records of these transactions, even as the genre gains in sophistication relative to its capacity to generate useful information. The assessment challenge will be the systematization of relevant indicators as well as the data distillation techniques utilized.
- The author describes the digital and electronic technologies as amplifiers of human abilities, and recognizes that these technologies do not simply enhance the existing human abilities; they appear to have the potential for creating new human capacities. Future assessments in education will need to be capable of documenting human abilities in their amplified state as well as these newly emerging human capabilities. Even at this time we can anticipate increasing demands for abilities that relate to adaptation to randomization: pattern recognition and generation of patterns; rationalization of contradictions; the adjudication of relational paradoxes; and the capacity for virtual problem solving.

- In the 20th century, testing and measurement of developed abilities dominated assessment. In the 21st-century, assessment for the development of human capacities will be the demand. Assessments in that new age will need to be diagnostic, prescriptive, instructive and capable of documenting what exists, capturing the processes by which abilities are developing, and modeling the achievements that are the ends of assessment, teaching and learning. Assessments will continue to be conducted and interpreted by the professionals others, but assessment will also be ubiquitously conducted by oneself and layperson others, in what Torre & Sampson (2012) describe as cultures of assessment, where evidentiary reasoning will become a colloquial basis for action, based on data that are ubiquitously generated in commerce, in life, in play, in study and in work.
- In most of the work of the Gordon Commission, there is elaborated an essentially epistemological rationale for new directions in our approach to assessment, but there is also a deontic rationale, which may be even more powerful than the epistemological. If the intent in assessment in education is to inform and improve teaching and learning, the moral obligation is to generate, interpret and make available the relevant evidence that is necessary for intervention as action on this enabled understanding.

9. The Findings and Recommendations of the Gordon Commission

The members of the Gordon Commission have not met formally to deliberate concerning findings and recommendations that can be drawn from the work of the Commission. The co-chair persons of the Commission, however, have agreed on the following conclusions on findings and recommendations that are grounded in the consultations, deliberations, and commissioned papers conducted by the Gordon Commission. Edmund W. Gordon and James W. Pellegrino have concluded that the findings and recommendations of the Commission can be summarized as follows:

FINDINGS

Nature of Assessment

1. Assessment is a process of knowledge production directed at the generation of inferences concerning developed competencies, the processes by which such competencies are developed, and the potential for their development.

2. Assessment is best structured as a coordinated system focused on the collection of relevant evidence that can be used to support various inferences about human competencies. Based on human judgment and interpretation, the evidence and inferences can be used to inform and improve the processes and outcomes of teaching and learning.

Assessment Purposes and Uses

3. The Gordon Commission recognizes a difference between a) assessment OF educational outcomes, as is reflected in the use of assessment for accountability and evaluation, and b) assessment FOR teaching and learning, as is reflected in its use for diagnosis and intervention. In both manifestations, the evidence obtained should be valid and fair for those assessed and the results should contribute to the betterment of educational systems and practices.

4. Assessment can serve multiple purposes for education. Some purposes require precise measurement of the status of specific characteristics while other purposes require the analysis and documentation of teaching, learning and developmental processes. In all cases, assessment instruments and procedures should not be used for purposes other than those for which they have been designed and for which appropriate validation evidence has been obtained.

5. Assessment in education will of necessity be used to serve multiple purposes. In these several usages, we are challenged to achieve and maintain balance such that a single purpose, such as accountability, does not so dominate practice as to preclude the development and use of assessments for other purposes and/or distort the pursuit of the legitimate goals of education.

Assessment Constructs

6. The targets of assessment in education are shifting from the privileging of indicators of a respondent's mastery of declarative and procedural knowledge, toward the inclusion of indicators of respondent's command of access to and use of his/her mental capacities in the processing of knowledge to interpret information and use it to approach solutions to ordinary and novel problems.

7. The privileged focus on the measurement of the status of specific characteristics and performance capacities, increasingly, must be shared with the documentation of the processes by which performance is engaged, the quality with which it is achieved and

the conditional correlates associated with the production of the performance.

8. Assessment theory, instrumentation and practice will be required to give parallel attention to the traditional notion concerning intellect as a property of the individual and intellect as a function of social interactions - individual and distributive conceptions of knowledge - personal and collegial proprietary knowledge.

9. The field of assessment in education will need to develop theories and models of interactions between contexts and/or situations and human performance to complement extant theories and models of isolated and static psychological constructs, even as the field develops more advanced theories of dialectically interacting and dynamic biosocial behavioral constructs.

10. Emerging developments in the sciences and technologies have the capacity to amplify human abilities such that education for and assessment of capacities like recall, selective comparison, relational identification, computation, etc., will become superfluous, freeing up intellectual energy for the development and refinement of other human capacities, some of which may be at present beyond human recognition.

Assessment Practices

11. The causes and manifestations of intellectual behavior are pluralistic, requiring that the assessment of intellectual behavior also be pluralistic, i.e., conducted from multiple perspectives, by multiple means, at distributed times and focused on several different indicators of the characteristics of the subject(s) of the assessment.

12. Traditional values associated with educational measurement, such as, reliability, validity, and fairness, may require reconceptualization to accommodate changing conditions, conceptions, epistemologies, demands and purposes.

13. Rapidly emerging capacities in digital information technologies will make possible several expanded opportunities of interest to education and its assessment. Among these are:

    a. Individual and mass personalization of assessment and learning experiences;

    b. Customization to the requirements of challenged, culturally and linguistically different and otherwise diverse populations; and

    c. The relational analysis and management of educational and personal data to inform and improve teaching and learning.

RECOMMENDATIONS DRAWN FROM THE WORK OF THE GORDON COMMISSION

The members of the Commission recognize that the future of assessment will be influenced by what the R&D and the assessment production communities generate as instruments and procedures for the assessment in education enterprise. However, we are very much aware that equally determinative of the future will be the judgments and preferences of the policymakers who decide what will be required and what practitioners and the public will expect. In recognition of the crucial role played by policymakers, the Executive Council of the Gordon Commission has given special attention to the development of a policy statement that concludes with three recommendations directed at those who make policy concerning education and its assessment. The statement has been prepared by James Pellegrino, co-chair of the Commission, and Lauren Resnick, member of the Executive Council, with input from Sharon Lynn Kagan, consultant to the Chair, and other members of the Executive Council — Randy Bennett, Eva Baker, Bob Mislevy, Lorrie Shepard, Louis Gomez and Edmund W. Gordon — and the assistance of Richard Colvin as writing consultant.

This Public Policy statement represents the authors' sense of recommendations that are implicit in the work of the Commission. However, it has not been vetted by the members of the Gordon Commission, and thus it should not be concluded that any given member of the Commission endorses the specifics included herein.

A Statement on Public Policy Concerning the Future of Assessment in Education

The Gordon Commission on the Future of Assessment in Education was created to consider the nature and content of American education during the 21st-century and how assessment can be used most effectively to advance that vision by serving the educational and informational needs of students, teachers and society. The Commission's goal in issuing this brief public policy statement is to stimulate a productive national conversation about assessment and its relationship to teaching and learning at a time when developments in assessment and education in the US present a remarkable opportunity to reconceptualize the purposes of educational assessments.

The statement advances arguments for:

1. *Transforming Assessment to Support Teaching, Learning and Human Development*

2. *Reconsidering Assessment: Why, What, and How We Assess*
3. *Moving Forward: The Opportunity*

Recommendations Concerning Public Policy

In the Realm of State Collaboration and Policy

It is recommended that states create a permanent Council on Educational Assessments modeled on the Education Commission of the States with functions such as:

- Evaluate the strengths and weaknesses of the Smarter Balanced and PARCC assessment systems and their effect on teaching and learning.
- Conduct research on how assessments are changing, help inform states so that they make good purchasing decisions, and surface issues as they arise. The Council also would oversee the process of setting cross-state performance level targets.
- Mount a public education campaign targeting parents, educators, school board members, and the media explaining the importance of good assessment to quality education.
- Create a Study Group on the Challenges of Equitable Assessment to explore issues related to diversity, equity, and excellence.
- Commission research on policies designed to secure the privacy of assessment data while also creating protocols for making large quantities of such data available to qualified researchers.

In the Realm of Federal Policy

It is recommended that the president and Congress build on various models to encourage experimentation with different approaches to assessment and accountability.

In the Realm of National Research and Development

It is recommended that the U.S. Department of Education, the Department of Defense, the National Science Foundation, and the National Institute of Child Health and Human Development, in collaboration with the philanthropic community, not-for-profit, for-profit sector, professional teacher organizations, and universities commit to a 10-year research and development effort to strengthen the capacity of the U.S. assessment.

General Recommendations Concerning the Future of Assessment in Education

1. As is traditional in the Medical profession and is rapidly embraced as a guide for all professional activity, the recommendation is made that in assessment policy, practice and use of assessment data, this field should “First Do No Harm.” Responsibility for honoring this value falls at multiple levels – policymakers, administrators, staff and perhaps most heavily on the manufacturers of assessment devices and those of us who sell them. (See Ho’s paper on purpose drift.)

2. We could declare as consensus among the members of the Commission that assessment can serve multiple purposes. There is less agreement concerning the possibility that a single test should be so used; however, the consensus holds concerning the need for balance in the attention given to the use of assessment for different purposes. It is recommended that with the possible exception of “informing and improving teaching and learning,” no single purpose should be permitted to distort the valued goals of education. Similarly, it is recommended that fidelity to the purpose for which the instrument or procedure is designed be honored. This recommendation references, among other concerns, the difference between our traditional concern with assessment of education and the Commission’s emphasis on assessment for education.

3. Assessment in education is essentially grounded in inferential reasoning. It is a process by which evidence collected for the purpose of the disconfirmation of inferences one seeks to make concerning the phenomena being assessed. It is therefore recommended that assessment processes be held to standards similar to those long honored in the tradition of the empirical sciences. However, given the Commission’s concern for changing paradigms and shifting epistemologies, it is further recommended that the universal utility of positivist scientific methodologies as a standard for evolving assessment practices be subjected to continuing inquiry.

4. We believe that most members of the Commission embrace concern for differential validities, i.e. the idea that validity may be a relative construct, and that it’s relativity must be taken into account in policy-making and practice with respect to assessment in education. It is therefore recommended that the field embrace the notion of differential validities and the imperative that tests of validity be appropriate to the populations and situations in which the construct is being utilized.

5. It is recommended that research and development efforts be intensified around questions related to the implications for assessment in education that flow from questions related to the cargo of learning transfer. Special attention may need to be given to the complementarities between mastery of declarative and procedural knowledge and the intentional command of instrumental mental processes.
6. It is recommended that the targets of assessment in education be broadened to include a wider range of human abilities, ways of adaptation, amplified abilities and human capacities, including those that are the products of exposure to digital electronic technologies.
7. Given the considerable evidence in support of agency, disposition, cultural identities, and existential states as influences on the nature and quality of human performance, it Is recommended that research and development concerning the relationships between human performance and these variables be given considerably greater priority in inquiries concerning assessment in education.
8. Debate continues concerning the idea that intelligence is a characteristic of individuals; intelligence is a collectively produced construct best associated with social groups; and the idea that intelligence originates and is expressed in both contexts. The increased practice of collaboration in the production of knowledge and its application suggests the importance of our recommendation that research and development effort be directed at differentiating assessments to capture intellective competence as a property of individuals and as a function of collaboration between persons.
9. Considerable concern has been expressed in the Commission about the artificiality of “Stand alone” or “Drop in from the Sky” tests. Perhaps more problematic than the isolated character of these examinations is concern with the tendency to treat the data from these tests as independent and sole sources of information concerning the performance and status of students. Some commissioners argued for the greater use of systems of examinations distributed over time embedded in the ongoing teaching and learning of experiences. It is recommended that assessment in education move progressively toward the development and use of diversified assessment systems for the generation and collection of educational assessment data.
10. It is then the final recommendation, implicit in the work of the Gordon Commission, that the academic and philanthropic sectors of the society – cooperatively supported by tax levy funds, consider the creation of a Virtual

Institute on the Future of Assessment in Education (VIFAE) to continue the inquiry initiated by the Gordon Commission; to encourage broad and cross disciplinary collaboration in this work; and to support the attraction to and development of young and new scholars to conceptual, research and development explorations of the relationships between assessment, teaching and learning.

10. About The Gordon Commission on the Future of Assessment in Education

Commission Background

Conceptions of what it means to educate and to be an educated person are changing. Notions of and demands on practice in the teaching and learning enterprise are broadening and expanding. And the concern with accountability forces this dynamic and eclectic enterprise to constrict and, in the worst of instances, to compromise in the interest of meeting certain accountability criteria. These realities, coupled with changes in epistemology, cognitive and learning sciences, as well as in the pedagogical technologies that inform teaching and learning, are narrowing — possibly even stifling — creativity and flexibility in teaching and learning transactions. These are among the perceived problems that led to the creation of the Gordon Commission on the Future of Assessment in Education by Educational Testing Service in January 2011.

Although these immediate issues were foundational in the establishment of the Gordon Commission, a second more compelling contextual problem helps to drive its mission. Changing conceptions of and practices in educational assessment are making many of the capabilities of traditional conceptions and practices in educational assessment obsolete. The work of the Commission rests on the assumption that assessment in education can inform and improve teaching and learning processes and outcomes.

Mission of the Commission

The Gordon Commission was created with the mission to study the best of educational assessment policy, practice and technology; consider the best estimates of what education will become and what will be needed from educational measurement during the 21st-century; and to generate recommendations on educational assessment design and application that meet and/or exceed the demands and needs of education — present and predicted.

Given the mission of the Gordon Commission, a number of goals were outlined that focused the work of the Commission. The goals of the Gordon Commission are to:

- Inform the field and the public about the need and possibilities for change in education, as well as change in the functions, practices and roles of assessment in education;
- Increase public awareness and knowledge about assessment as an integral component of education and the possibilities for change in assessment practice;
- Encourage the field of educational assessment to strengthen its capacity to factor into measurement practice attention to the influence of human attributes, social contexts and personal identities on human performance;
- Balance emphasis on prediction, selection and accountability with equal concern for informing and improving teaching and learning processes and outcomes; and
- Inform long-term planning and product development in the field of psychometrics.

Commission Members

The Gordon Commission consists of 30 members. The scholars, policymakers and practitioners who comprise the Commission have identified critical issues concerning educational assessment, investigated those issues, and developed position and review papers that informed the Commission's recommendations for policy and practice in educational assessment.

Chairperson

Edmund W. Gordon
John M. Musser Professor of Psychology, Emeritus
Yale University
Richard March Hoe Professor of Education and Psychology, Emeritus
Teachers College, Columbia University

Co-Chair

Jim Pellegrino
Liberal Arts & Sciences Distinguished Professor Distinguished Professor of Education Co-Director, Learning Sciences Research Institute University of Illinois at Chicago

Executive Council

Eva Baker
Distinguished Professor, Graduate School of Education and Information Studies, and Director, National Center for Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles

Randy E. Bennett
Norman O. Frederiksen Chair in Assessment Innovation, Educational Testing Service (ETS)

Louis M. Gomez
MacArthur Foundation Chair, Digital Media and Learning, Graduate School of Education & Information Studies, University of California, Los Angeles

Robert J. Mislevy
Frederic M. Lord Chair in Measurement and Statistics, ETS

Lauren Resnick
Senior Scientist, and Project Director, Learning Research and Development Center, and Distinguished University Professor of Psychology and Cognitive Science, University of Pittsburgh

Lorrie A. Shepard
Dean, School of Education, and Professor of Education, University of Colorado at Boulder

Commissioners

J. Lawrence Aber
University Professor and Albert and Blanche Willner Family Professor of Psychology and Public Policy, Department of Applied, Psychology, Steinhardt School of Education, New York University (NYU)

Bruce M. Alberts
Professor, Department of Biochemistry and Biophysics, University of California, San Francisco, and Chief Editor, Science Magazine

John Bailey
Director, Dutko Worldwide

John T. Behrens
Vice President, Pearson Center for Digital Transformation

Ana Mari Cauce
Provost and Earl R. Carlson Professor of Psychology, University of Washington

Linda Darling-Hammond
Charles Ducommun Professor of Education, and Co-Director, School Redesign Network (SRN), School of Education, Stanford University

Ezekiel Dixon-Roman
Assistant Professor, School of Social Work and Social Policy, University of Pennsylvania

James Paul Gee
Mary Lou Fulton Presidential Professor of Literacy Studies, Arizona State University

Kenji Hakuta,
Lee L. Jacks Professor of Education, School of Education, Stanford University

Frederick M. Hess
Resident Scholar and Director of Education Policy Studies, American Enterprise Institute for Public Policy Research

Andrew Ho
Assistant Professor of Education, Graduate School of Education, Harvard University

Freeman A. Hrabowski III
President, University of Maryland, Baltimore County

Michael E. Martinez (1956–2012)
Professor, Department of Education, University of California, Irvine

Rodolfo Mendoza-Denton
Associate Professor, Psychology Department, University of California, Berkeley

Shael Polakow-Suransky
Chief Academic Officer and Senior Deputy Chancellor, New York City Department of Education

Diane Ravitch
Research Faculty, Steinhardt School of Culture, Education, and Human Development, NYU

Charlene Rivera
Research Professor, and Executive Director, Center for Equity and Excellence in Education, George Washington University

Lee Shulman
President Emeritus, The Carnegie Foundation for the Advancement of Teaching, and Charles E. Ducommun Professor of Education–Emeritus, School of Education, Stanford University

Elena Silva
Senior Associate, Public Policy Engagement,
Carnegie Foundation for the Advancement of Teaching

Claude Steele
Dean, Graduate School of Education, Stanford University

Ross Wiener
Executive Director, Education and Society Program, The Aspen Institute

Robert Wise
Former U.S. Governor, West Virginia, and President, Alliance for Excellent Education

Constance M. Yowell
Director of Education, The John D. and Catherine T. MacArthur Foundation

Staff

Executive Officer:	**Paola Heincke**
Embedded Journalist:	**David Wall Rice** Associate Professor of Psychology Morehouse College
Multimedia Advisor:	**Mikki Harris** Multimedia Consultant and Professor of Journalism University of Mississippi
Senior Research Scientist:	**Ernest Morrel** Professor of Education and Director, Institute for Urban and Minority Education (IUME) Teachers College, Columbia University
	Rochelle Michel Senior Product Management – Lead Educational Testing Service
Research Assistants:	**Emily Campbell**
	E. Wyatt Gordon
	Emile Session
	Paola Andrea Valencia-Cadena
Editorial Assistant:	**Maralin Roffino** Assistant to the Director of Communications SUNY Rockland Community College

WORK OF THE COMMISSION

Meetings of the Commission

There were two face-to-face meetings of the Gordon Commission. The initial meeting was held May 24-25, 2011 at the Chauncey Conference Center in Princeton, NJ and the second meeting was held February 12-13, 2012 at the Caribe Hilton in San Juan, Puerto Rico.

Consultative Conversations

The Gordon Commission spent much of its first year gathering and synthesizing information and perspectives concerning the state of the art and sciences of educational measurement and assessment. The chairman and members of the Commission have held individual consultations with experts around the country who provide input into the work and the direction in which the Commission is going. The Commission hosted more than a dozen consultative conversations with groups that advised the Commission on the identification of issues that need to be addressed and the substance of the issues to be considered.

The Gordon Commission Fellows

The Gordon Commission Fellows is a dynamic group of six emerging pre- and post-doctoral scholars in the fields of the learning sciences, anthropology, psychometrics, the sociology of education, and education technology. These Fellows were assembled to analyze and identify emergent themes, critical innovations, similarities and distinctions, and ultimately synthesize the knowledge produced across the body of the commissioned papers in brief papers of their own. The idea behind the creation of this group was that the work of the commission's experienced scholars and policymakers should be complemented by a younger generation who, in their ongoing dialogue and in their syntheses of the more than two dozen papers, would add new life and new ideas to the project. During their work together over the spring and summer, each Fellow selected overlapping cross-sections of the papers to critically analyze and present for a series of Fellows-led group discussions, all under the tutelage of Commission chairman Dr. Edmund W. Gordon and Dr. Ernest Morrell, the current director of the Institute of Urban Minority Education (IUME) at Teachers College, Columbia University.

The Gordon Commission Fellows are: Keena Arbuthnot, Ph.D. in educational psychology from the University of Illinois at Urbana-Champaign; Sherice N. Clarke, Doctoral student in education at the University of Edinburgh; Juliette Lyons-Thomas, Doctoral student in the Measurement, Evaluation, and Research Methodology (MERM) program at the University of British Columbia; Jordan Morris, Doctoral student in the Social Welfare program at the University of California, Los Angeles; Catherine Voulgarides, Doctoral student in the Sociology of Education program at New York University; and Amanda Walker Johnson, Ph.D. and M.A. in anthropology (sociocultural) from the University of Texas at Austin's African Diaspora Program. For bios and more information please go to http://www.gordoncommission.org/fellows.html.

Science, Technology and Scientific Imagination

Under the auspices of the Gordon Commission on the Future of Assessment in Education, the Arizona State University (ASU) Center for Games and Impact, the ASU Center for Science and the Imagination, and the Carnegie Mellon Project on Working Examples (funded by the MacArthur Foundation and the Gates Foundation), sponsored two concurrent symposia on October 25-27, 2012 at ASU: 1) The Perils and Possibilities of Emerging Technologies for Learning and Assessment, and 2) Science and Imagination – The Future for the Teaching, Learning and Assessment We Want and How to Get There. These symposia are based on longer-term projects related to these areas.

Excellence, Diversity and Equity

In the agreement by which the Gordon Commission was funded, the Commission was asked to give special attention to the problems posed for assessment by the concern for the concurrent privileging of the pursuit of excellence and equity in academic opportunity and achievement. Through the Excellence and Equity Project, the Commission has honored that agreement. This concern is addressed in a group of the Gordon Commission papers directed at the synthesis of knowledge and thought concerning disabling and handicapping conditions, cultural variation, differences in first language and class/ethnic diversity. In addition, a small study group has been designed to give extended discussion to this set of problems.

Communication and Social Marketing

A bifocal program of communication was developed for the Gordon Commission. As part of the internal communication plan, the Commission created a blog that was used for the Commission members. The external communication plan included: a) the creation of a website; b) the development of a bimonthly bulletin, *Assessment, Teaching, and Learning;* c) hosting of public hearings and forums; and d) the use of regular and social media for the dissemination of strategic messages to target audiences.

Bibliographic Resources

From the beginning of the work of the Gordon Commission, staff members and Fellows have worked to compile a comprehensive collection and directory of the bibliographic resources used in the course of this work. Our Resources File is not a definitive collection; however, it does represent what we think of as the most important literature that has relevance for the work of the Gordon Commission. The collected works are organized under the working categories used by staff and can be searched using common search terms and the special search terms indicated in the File. It can be found under “Resources” at www.gordoncommission.org.

Knowledge Synthesis Project

The central activity of the Gordon Commission has been referred to as “the Knowledge Synthesis Project”. This initiative consisted of the commissioning 25 reviews of extant knowledge and thought papers concerning the issues that were identified as most important. The papers that resulted from this work will be published in the series *Perspectives on the Future of Assessment in Education.* http://www.gordoncommission.org/publications_reports.html.

Assessment in Education: Changing Paradigms and Shifting Epistemologies

1. Epistemology in Measurement: Paradigms and Practices – Part I. A Critical Perspective on the Sciences of Measurement (Ezekiel J. Dixon-Román and Kenneth J. Gergen)
2. Epistemology in Measurement: Paradigms and Practices – Part II. Social Epistemology and the Pragmatics of Assessment (Kenneth J. Gergen and Ezekiel J. Dixon-Román):
3. Postmodern Test Theory (Robert J. Mislevy)[5]
4. What Will It Mean to Be an Educated Person in Mid-21st-Century? (Carl Bereiter and Marlene Scardamalia)
5. Toward an Understanding of Assessment as a Dynamic Component of Pedagogy (Eleanor Armour-Thomas and Edmund W. Gordon)
6. Preparing for the Future: What Educational Assessment Must Do (Randy Elliot Bennett)
7. Changing Paradigms for Education: From Filling Buckets to Lighting Fires to Cultivation of Intellective Competence (E. Wyatt Gordon, Edmund W. Gordon, John Lawrence Aber, and David Berliner)

Changing Targets of Assessment in Education

8. The Possible Relationships Between Human Behavior, Human Performance, and Their Contexts (Edmund W. Gordon and Emily B. Campbell)
9. Education: Constraints and Possibilities in Imagining New Ways to Assess Rights, Duties and Privileges (Hervé Varenne)
10. Toward a Culture of Educational Assessment in Daily Life (Carlos A. Torre and Michael R. Sampson)
11. Toward the Measurement of Human Agency and the Disposition to Express It (Ana Mari Cauce and Edmund W. Gordon)
12. Test-Based Accountability (Robert L. Linn)
13. Variety and Drift in the Functions and Purposes of Assessment in K-12 Education (Andrew Ho)
14. Testing Policy in the United States: A Historical Perspective (Carl Kaestle)

[5] Postmodern Test Theory (Robert J. Mislevy) is Reprinted with permission from *Transitions in Work and Learning: Implications for Assessment, 1997*, by the National Academy of Sciences, Courtesy of the National Academies Press, Washington, D.C.

Psychometric Change in Assessment Practice

15. Four Metaphors We Need to Understand Assessment (Robert Mislevy)
16. Assessment as Evidential Reasoning (Joanna S. Gorin)
17. Assessment in the Service of Teaching and Learning (Clifford Hill)
18. Testing in a Global Future (Eva Baker)
19. Technological Implications for Assessment Ecosystems: Opportunities for Digital Technology to Advance Assessment (John T. Behrens and Kristen E. DiCerbo)
20. Toward the Relational Management of Educational Measurement Data (Greg K. W. K. Chung)

Assessment in Education and the Challenges of Diversity, Equity and Excellence

21. Human Diversity, Assessment in Education and the Achievement of Excellence and Equity (A. Wade Boykin)
22. Assessment of Content and Language in Light of the New Standards: Challenges and Opportunities for English Language Learners (Kenji Hakuta)
23. Democracy, Meritocracy and the Uses of Education (Aundra Saa Meroe and Edmund W. Gordon)
24. Accommodation for Challenge, Diversity and Variance in Human Characteristics (Martha L. Thurlow)
25. A Social Psychological Perspective on the Achievement Gap in Standardized Test Performance Between White and Minority Students: Implications for Assessment (Rodolfo Mendoza-Denton)

1. CRITICAL ISSUES FOR THE FUTURE OF ASSESSMENT IN EDUCATION

In the initial meeting of the Gordon Commission, attention turned to questions having to do with why we assess, what we assess and how we assess in education now and in the future. The members of the Commission quickly agreed that the answers to these questions should form the context for our inquiry into the future of assessment. However, before we could seriously engage issues related to the future of assessment in education, a substantial number of the members of the Commission insisted that there was something wrong with the investment of considerable energy in the long-term future – mid- and late-21st-century – of assessment in education, when some members of the Commission consider that education and its assessment is currently in crisis. Substantial concern was expressed about the fact that national education policy is driven by a concern with the use of assessment in education primarily for purposes of accountability.

Many commissioners were not as troubled by the accountability focus as they were concerned with the traditional use of standardized tests in accountability and especially for high stakes decision-making. Still others were concerned that the testing and measurement enterprise may be grounded in systems of thought that are contradicted by emerging epistemologies and changing paradigms for education and its assessment. To accommodate this wide range of concerns, one of the initial activities of the Gordon Commission involved the identification of what commissioners agreed were the most critical issues facing the field. It was thought that the encirclement of extant knowledge and thought concerning these issues should inform the work of the Gordon Commission as it inquired into the current state of assessment in education, the best of extant theory and practice, and our understanding of the changing nature of education and its assessment in the present and anticipated future.

This decision led to the conduct of the central activity of the Gordon Commission that has been referred to as the Knowledge Synthesis Project. This initiative consisted of the commissioning of 25 reviews of extant knowledge and thought concerning the issues that were identified as most important. The papers that resulted from this work are listed in this report. These papers will be published in a four volumes series, *Perspectives on the Future of Assessment in Education*. Under the guidance of our two senior research associates, Rochelle Michelle, PhD, and Ernest Morell, PhD, these several papers written specially for the Gordon Commission were subjected to analysis and digest by six emerging scholars who served as pre- and post-doctoral Commission Fellows. What follows is the product of their effort.

Developing Perspectives on Assessment

The papers contained within this section (Kaestle, 2012; Meroe, 2012; Varenne, 2012; Mendoza-Denton, 2012; Dixon-Román & Gergen, 2012; Gergen & Dixon-Román, 2012; Torre & Sampson, 2012; Bennett, 2012) all provide varying views on the historical context for assessment, ranging from testing policies to measurement models used in testing.

Kaestle's (2012) paper, *Testing Policy in the United States: A Historical Perspective*, uses the history of testing to discuss the competing forces within educational testing. That is, the widespread use of standardized, multiple-choice tests to measure basic skills in various contexts such as placement, accountability, and program evaluation, which was viewed as a limiting factor. It was also noted that there are some equity and accountability goals that have been well-served by being able to pinpoint how well individual students or groups of students are doing. Kaestle also acknowledges the power of standardized, multiple choice tests due to their cost effectiveness and efficiency compared to the more complex, more subjective and higher-level assessments. These positive qualities of standardized, multiple-choice tests stand in the way of the call for authentic and performance based assessments that challenge existing frameworks.

This paper notes the costs involved with having an educational system that has such a central focus on standardized multiple-choice tests of basic skills. Although Dixon-Román and Gergen (2012) warn against blindly trusting the quality of test instruments, Kaestle (2012) recommends that new systems of assessment should clearly define why the proposed system would be better than the current system. In addition, Kaestle emphasizes the importance of being able to articulate how the new system can meet accountability goals. Kaestle also recommends that the new system of assessment should be accessible and understood by lay people and educational practitioners. Kaestle advises proponents of authentic and performance assessment to craft a narrative about the need and significance of these assessments and their value beyond traditional tests that include multiple choice questions.

Meroe's (2012) paper, *Democracy, Meritocracy and the Uses of Education*, identifies the tensions that existed when both democracy and meritocracy were developed. Meroe identified three tensions: a) full participation and notions of a deserving elite; b) majority rule on the one hand, and individual and minority rights (and protections) on the other hand; and c) expansions of freedoms in the United States. Meroe encourages the education community to look to other countries. For example, Finland's success has been attributed to high standards of excellence in both curriculum and teaching, universal social benefits and targeted supports for underperforming and comparatively underprivileged students, for examples of how to balance assessment, governmental support, and a collective ethos for educational system.

Varenne's (2012) paper, *Education: Constraints and Possibilities in Imagining: New Ways to Assess Rights, Duties and Privileges*, finds that schools have not taken into account "interaction theory" and have not recognized human agency (Cauce & Gordon, 2012) and the educative quality of everyday life moments. Varenne also points out that the life of schooling should be understood as an ongoing, interactive and changing process, which implies that assessment must also be conceptualized as ongoing, and take into consideration that various methods of self-education exist, and that there are different systems of education, assessments, curriculum, and pedagogy that occur outside of the typical school setting. Varenne (2012) recommends that we challenge the dichotomization of the "educated" and "uneducated" that ultimately imposes a deficit-model upon those with lower levels of formal education.

Varenne also makes some recommendations for states' involvement in the education process, specifically that states should protect the right to a free and public education for all; states should not use school-based assessment and certification (or degrees) as a means to grant career privileges. Although Varenne found the use of assessments for selection and granting privilege to be reasonable, Varenne noted that schools do not have to be the site for those types of assessments. He notes that some find more value in developing their own assessment systems (e.g., some employers rather develop their own in-house assessments than delegating assessment to an external body). Varenne also provided specific recommendations regarding assessments and the assessment process. Varenne recommends that assessment must not be time bound, static, bound in language of success or failure, or exclusionary, and assessment must move from being something that is rewarded in the market to something that betters the human experience. Varenne calls that, "it may be time to figure out how individuals educate themselves." Meroe's identification of social mobility through educational attainment is aligned with Varenne's view of the best-case scenario, where rights and privileges are granted by the state through the school and are based on an assessment of the merits of the person receiving the degree. Both Meroe (2012) and Varenne (2012), have noted that this is not the case and that there are other factors that play into one's social mobility, and that schooling has not equalized chances or opportunities.

The Mendoza-Denton (2012) paper, *A Social Psychological Perspective on the Achievement Gap in Standardized Test Performance between White and Minority Students: Implications for Assessment*, acknowledges the research that has examined test bias in assessment, including factors such as stereotype threat. Mendoza-Denton's review of the research suggests that academic achievement gaps, rather than reflecting test bias, may be more accurately described as reflecting societal bias. As identified in Cauce and Gordon (2012), an individual and their environment (or social context) are intricately intertwined and certain environments may facilitate productive or destructive reactions to failure. Mendoza-Denton also argued that ambiguous environments may contribute to the achievement gap more than test bias, because ambiguous environments can promote uncertainty about one's perceptions and ability and lead to disengagement because the feedback is not trusted. Mendoza-Denton provides a number of specific recommendations that test developers can follow in an effort to close the achievement gap.

Mendoza-Denton recommends that the question of biased or unbiased testing be reframed to consider threatening or nonthreatening environments. Mendoza-Denton states that the educational environment is psychologically not equivalent for minority and majority students. In addition, Mendoza-Denton identified the strong need to create environments in which people can trust the fairness of the feedback they receive, as well as their own belongingness within these environments. Mendoza-Denton recommended encouraging attitudes that see intelligence as malleable and incremental as opposed to fixed. Mendoza-Denton recommends that testing organizations increase their diversity within their organizations, as well as collaborate with a variety of stakeholders to ensure that their developed assessments are accessible, and capture differences in competence and qualifications for all future test takers. Mendoza-Denton also advocates for creating tests that assess a wider variety of skills than currently being focused on and that do not show evidence of group differences.

Dixon-Román and Gergen (2012), *Epistemology in Measurement—Part I. A Critical Perspective on the Sciences of Measurement,* find that the aims of measurement are at odds with social relational processes of education. Mislevy, Moss and Gee (as cited in Dixon-Román & Gergen, 2012) discuss a perspective where statisticians and psychometricians would treat probability models of measurement as one factor in understanding what a student knows, within a particular context. This argument is supported by Linn (2012), in his argument around the value of combining both quantitative and qualitative information to better understand students' knowledge, skills, and abilities. The contextual and social considerations of teaching, learning, and assessment have also been discussed in other commissioned papers (e.g., Bereiter & Scardamalia, 2012; Cauce & Gordon, 2012; Gorin, 2012; Hakuta, 2012; Mendoza-Denton, 2012; Meroe, 2012; and Varenne, 2012). In addition, the authors have found that policymakers have assumed that the instruments of measurement are sound, robust, and valid, but it appears that this is a blind assumption and advise against blindly trusting the quality of the instruments. Dixon-Román and Gergen (2012) recommend that the origins of practices that are still used today be explained to policymakers and other stakeholders that make use of test results. In addition, the authors recommend pointing out that other epistemologies (e.g., assessment systems that use both quantitative and qualitative information) may be beneficial for framing assessment in the 21st-century.

Gergen and Dixon-Román (2012) found that current testing practices have negatively impacted the teaching and learning process, resulting in a narrowing of curriculum and pedagogical methods. Linn's *Testing Accountability* (2012) finds this also, and discusses the underlying assumptions of the current accountability model, where sanctions and rewards are meant to be seen as motivating factors that will help to change teacher practices and eventually lead to more motivated students and greater student achievement. However, Gergen and Dixon-Román believe that this has led to reduced levels of motivation and engagement in both teachers and students, and has led to negatively shaping parents' views of their children. In addition, the authors find that education and measurement divide and stratify society and creates arbitrary hierarchies that have serious social implications. As other authors have noted (Bereiter & Scardamalia, 2012; Cauce & Gordon, 2012), collaboration is not emphasized within the current educational systems, and Gergen and Dixon-Román have found these to be counter-democratic practices that suppress pluralism and particularity. The authors also note that the current measurement system does not allow for local differences and human characteristics to be considered. Cauce and Gordon (2012) have also noted the importance of promoting human agency and incorporating it into the teaching, learning and assessment process. Gergen and Dixon-Román recommend moving toward an assessment system that uses multiple criteria, contains a formative component, and includes professional development for teachers and administrators. A number of other commissioned papers found value and also recommended these approaches for future assessments. Gergen and Dixon-Román also recommended keeping standardized testing but also expanding the availability of different kinds of testing and training education communities in participatory evaluation as a form of accountability. As mentioned in the companion Dixon-Román and Gergen (2012) paper, there was a recommendation to move toward considering measurement from a social constructivist perspective rather than the commonly used positivist perspective.

The Torre and Sampson (2012) paper, *Toward a Culture of Educational Assessment in Daily Life*, makes the case for a culture in which educators, as well as laypersons, have an understanding of educational assessment data and processes. Torre suggests that students know more about themselves, and are better at self-assessment than external assessments at evaluating their strengths and limitations. Educating students and others about how to assess their own learning, growth, and the process through which they've accomplished their goals is an effective pathway toward the development of critical thinking. Torre believes that providing learners with opportunities for self-assessment can enhance outcomes, so that the learning becomes more personal and they take an active role in the teaching, learning, and assessment process. The self-assessment process allow students to reflect on what they have learned, and allows teachers to focus on how to best teach individual students. However, Torres and Sampson do not view the self-evaluation process as one to undertake in isolation. The authors also highlight the interactive form in which self-assessment should take place. It is through these interactions that students can further develop enthusiasm, motivation and inspiration for learning the subject matter, while receiving critical feedback that can facilitate discussions on beliefs, thoughts, and objectives.

The Authors provide a view of this type of learning environment where monitoring of learning is viewed as an interaction in which the growth of the students is being evaluated while the learning environment is being monitored and revised to meet the ever-changing needs of the student. Examples are provided of what researchers, as well as higher education institutions have said about what it means to be an educated person. The key aspect was related to being able to go beyond the skills that are learned and apply it to other contexts. The authors highlighted this as a reason for favoring more broadly defined educational objectives such as critical thinking, communication, and creativity, which the authors assume will always be useful, regardless the context. However, they acknowledge that this is different than what most teachers know how to effectively teach and in some cases some may question whether or not these are teachable skills.

Torre and Sampson acknowledge that in order to change the current culture of classrooms from one of passively depending on an authority figure to tell students how they have performed, to one of mutual collaboration through which teacher and student consider the merits of what was intended compared to what was achieved. There will be a need for the development of criteria by which learners can evaluate their own work. In addition, standards, even examples of what is required, must be clear, meaningful, and relevant. Teachers can help student define the criteria for various levels of performance (i.e., excellent, mediocre, poor). The authors note that an adequate amount of time must be provided for students to develop and consider their own critiques as well as the feedback from their teachers.

Students need to find out what they have accomplished and what uncertainties remain. The result of this self-evaluation also provides feedback for teachers to improve their teaching. Torre and Sampson also provide information about possible tools to assist in the self-evaluation process, such as the K-W-L strategy (Ogle, 1989)—which allows student to go from what they know to what they think they know, use of student log books to document their learning and monitor their progress and growth. The added element of including both student and teachers comments in the log book fosters additional interaction between the students and teachers.

Bennett's (2012) paper, *Preparing for the Future: What Educational Assessment Must Do*, explores the forms that summative and formative assessments will take and the competencies that they will measure in the future. Education, and the world for which it is preparing students, is changing quickly. Educational assessment will need to keep pace if it is to remain relevant. This paper offered a set of claims for how educational assessment might achieve that critical goal. Many of these claims are ones to which assessment programs have long aspired. However, meeting these claims in the face of an education system that will be digitized, personalized, and possibly gamified will require significantly adapting, and potentially reinventing, educational assessment. Our challenge as a field will be to retain and extend foundational principles, applying them in creative ways to meet the information and decision-making requirements of a dynamic world and the changing education systems that must prepare individuals to thrive in that world.

The author proposes a set of 13 claims about what educational assessment must do if it is to remain relevant and if assessment is to actively and effectively contribute to individual and institutional achievement. The author notes that in order for assessment systems to remain relevant, future educational assessment systems will need to provide trustworthy and actionable summative information for policymakers as well as formative information for teachers and students. He has identified the need for assessments that serve multiple purposes. However, a single test may not be able to meet the needs beyond which the assessment was originally developed. It may be the case that an assessment developed for multiple purposes may not work for any of the identified purposes.

According to Bennett, assessment for education must:

- Provide meaningful information
- Satisfy multiple purposes
- Use modern conceptions of competency as a design basis
- Align test and task designs, scoring and interpretation with those modern conceptions
- Adopt modern methods for designing and interpreting complex assessments;
- Account for context
- Design for fairness and accessibility
- Design for positive impact
- Design for engagement
- Incorporate information from multiple sources
- Respect privacy
- Gather and share validity evidence
- Use technology to achieve substantive goals

Accountability and Validity Frameworks

The papers within this section (Linn, 2012; Mislevy, 2012; Gorin, 2012; and Ho, 2012) discuss the evolving uses of tests and the need to consider assessment frameworks that take into consideration the current and potential uses of test in the context of the teaching, learning, and assessment process. In addition, these papers challenge the testing industry to develop assessment systems that can capture evidence of student learning at multiple time points, from different sources (i.e., inside and outside of school settings), different types (i.e., quantitative and qualitative), and that allow for the demonstration of student learning in different ways.

Linn's (2012) paper, *Test Accountability*, provides commentary on the lessons learned from past experiences with test-based accountability systems and acknowledges that testing and accountability have grown to become influential in strategies for education reform. Linn finds that accountability systems need to include mechanisms to evaluate score inflation and guard against it. He also find that tests that include self-monitoring systems should be an effective approach in the future, and low-stakes tests should be used to monitor progress on high-stakes tests and may be used as one of the possible mechanisms to avoid test score inflation. Linn also argues that test-based accountability rests on the following assumptions: a) teachers know how to improve student achievement; b) the sanctions and rewards associated with achievement can be adequately linked to teacher performance; and c) achievement tests correctly measure student learning and the tests cannot be manipulated in practice. In light of these findings, Linn emphasized the importance and usefulness of collecting and using both quantitative and qualitative approaches to accountability. For example, test scores could be used as a trigger to identify schools where on-site visits could be conducted to collect qualitative information that might explain observed results. This would allow for an opportunity to suggest possibilities for improvement and to provide contextual information to educators regarding the basis for quantitative findings. Using the complementary quantitative and qualitative approaches will facilitate accountability systems with providing instructionally useful data to teachers and educational practitioners.

Mislevy's (2012) paper, *Four Metaphors You Need to Understand Assessment*, identifies the need for a more systematic framework for understanding, organizing, and distinguishing concepts underlying the purposes, designs, and uses of assessment to facilitate conversations between experts and the public. The National Council on Measurement in Education (NCME) issued a special edited issue (Allalouf & Sireci, 2012) related to this same, very critical area, *Dissemination of Measurement Concepts and Knowledge to the Public*. In addition, Sireci and Forte (2012) contributed by adding in some actionable steps that could be taken to make this a reality. Mislevy's framework adds to these existing sources and further highlights the need for conversations between testing experts and the public to be considered as we move into framing the future of assessment for the 21st-century.

Mislevy uses a set of metaphors to help conceptualize this new framework. He begins with four overarching metaphors that provide a framework for better understanding assessment. These metaphors are:

- Assessment as practice, where assessment is matched with real-world situations so that the inferences that are made from assessment results can be strengthened;

- Assessment as a feedback loop which emphasizes the multiple uses of assessment results and a consideration of the length of time between when the assessment is administered and when the results are actually shared with interested stakeholders;
- Assessment as evidentiary argument where the assessment results are used to provide evidence about what a student knows based on what he or she says, does or makes within a limited set of situations or contexts; and
- Assessment as measurement where a framework is provided to reason about patterns of information in context.

In addition to the aforementioned, four overarching metaphors, Mislevy (2012) identified four more precise metaphors that provide a sharper focus for understanding assessment: tests as contests, assessment design as engineering, examination as the exercise of power, and assessment as inquiry. These additional metaphors seek to address the missteps within the current educational assessment climate, and push for creating assessments that promote ways of using knowledge, techniques, and values of an individual to solve an array of problems. These new assessments would allow one to evaluate what Gordon and Bridglall (2007) have called, an individual's *intellective competence*. Intellective competence is an individual's "ability and disposition to use knowledge, technique, and values through mental process to engage and solve both common and novel problems" (Gordon, 2007). Mislevy's framework can be used to set the stage for discussions on how the various stakeholders in the assessment process can think about assessment issues. The framework also provides a logical way to organize and conceptualize those issues.

Gorin's (2012) paper, *Assessment as Evidential Reasoning*, draws on the 1999 Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1999) to support the types of information that are needed from an educational assessment to support making an evidentiary argument. The evidentiary argument described in this paper complements Mislevy's (2012) metaphor of assessment as an evidentiary argument about students' learning and abilities, given their behavior in particular circumstances. Gorin finds that there are four things that are needed from educational assessment to support an evidentiary argument: a) clear definitions of all possible states of knowledge; b) lists of behavioral evidence that would illustrate a particular state of knowledge; c) a set of data collection procedures that produces the relevant behavioral evidence; and d) scoring rules for using behavioral evidence to determine individuals' state of knowledge from among the set of possible states. In addition, Gorin questions whether the evidence that is currently being used to make inferences about student ability and subsequently, claims about knowledge and ability is the most valid and reliable evidence. Gorin calls for a more comprehensive view of educational assessment, where multiple evidential sources are provided (e.g., tests and inventories, behavioral observations, interviews). Gorin also raises two questions to which assessment designers should respond: a) what constitutes the best evidence with which the most persuasive arguments and inferences about student knowledge and ability can be made?; b) under what conditions can we produce robust evidence of student knowledge and ability in order to make these claims?

In consideration of any assessment, the validity of the scores should be at the forefront of the discussion. In the context of assessments, the validity of a test is the extent to which it measures what it claims to measure. However, it is important to note that validity is an evolving concept (AERA, APA, & NCME, 1999; Anastasi & Urbina, 1997). Ho's paper, *Variety and Drift in the Functions and Purposes of Assessment in Education*, continues the evolution by acknowledging that maintaining a test developer's purpose is difficult after the test enters the public domain where the test is used and the test's meaning can be modified. In addition, Ho notes that purpose drift presents validity issues in reference to the interpretation of test scores. In Ho's validity framework, he delineates four *measuring* purposes and three *influencing* purposes. The four measuring purposes were identified as (a) instructional guidance, (b) student placement and selection, (c) informing comparisons among educational approaches, and (d) educational management. The three influencing purposes were identified as (a) directing student effort, (b) focusing the system, and (c) shaping public perceptions.

Given the acknowledgement that test results may eventually be used for purposes other than for which they were intended, Ho makes some recommendations to address the validity concerns that will arise from this repurposing. Ho also challenges the testing community to increase their foresight into the potential routes for purpose drift, as well as the consequences. Ho advises that these potential routes of purpose drift should be thought of and explicitly taken into account when the tests are developed. Ho believes that standards should be raised for validity and validation but also recognizes that there should be a verification of test users' understanding of the current standards for validity and validation. Although there currently exist a number of validity frameworks, Ho recommends that a new validity framework needs to be developed that pays more attention to differentiating the assessment of learners and learning, as well as differentiating between the assessment of teachers and teaching.

Beyond the Basics

While current large-scale, standardized tests focus on the basic skills of reading, writing, and mathematics, and to a lesser degree science and history, the next set of papers (Bereiter & Scardamalia, 2012; Cauce & Gordon, 2012; and Armour-Thomas & Gordon, 2012; Baker, 2012) call for a movement to go beyond these basics and consider a wider range of competencies. In addition, these papers support a more integrated approach for instruction, curriculum, and assessment that support student learning and allow students to move beyond the basics that are learned and transfer that knowledge to other contexts beyond the one in which the original knowledge was learned. These papers also highlight the importance of collaboration and acknowledging the varying social contexts in which students learn.

The Bereiter and Scardamalia (2012) paper, *What Will it Mean to Be an Educated Person in the Mid-21st-century?*, finds that a mid-21st-century person should possess a range of competencies that include a wide range of knowledge and an understanding of problems of various levels of depth and complexity. Bereiter and Scardamalia identified five competencies: a) knowledge creating where students are able to build, amend and create knowledge; b) working with abstractions where students should be able to work with abstraction and convert them to real-world applications, going from the theoretical to the practical; c) systems thinking where students should be able to recognize and understand the complexity of the world and consider how to take advantage of the complexity whenever possible; d) cognitive persistence where students should be able to sustain focus and study in the face of increasing obstacles and distractions and e) collective cognitive responsibility where students should be able to engage in collective work that is collaborative. The authors also acknowledge that technology plays a role in the assessment of the aforementioned competencies. The authors also identified two areas of deficiency in that students are currently not being asked to sufficiently synthesize their knowledge and that students are not sufficiently applying critical thinking skills to place readily available information within the larger political, historical, economic or social contexts.

Bereiter and Scardamalia (2012) recognize that as theories of collaborative learning develop, learners should be given instructional space to collaborate, and assessment should adapt so that individual and collaborative contributions to solving problems may be measured and evaluated. The authors recommend preparing learners to engage in lifelong learning, enabling learners to gain new competencies, while adapting to the accelerating pace of change. Part of this will require education to foster breadth, depth and the ability to navigate diverse ideas, peoples, and cultures. To this end, assessments should be developed that foster creativity. Bereiter and Scardamalia also call for systems-level thinking where students are able to both discern usefulness of knowledge and place knowledge within the appropriate context. The authors also recommend developing methods for assessing knowledge creation, work with abstractions, systems thinking, cognitive persistence, and collaborative responsibility.

The Cauce and Gordon (2012) paper, *Toward the Measurement of Human Agency and the Disposition to Express It*, identifies human agency and its expression as a central mechanism that helps to drive educational outcomes. The authors define human agency as the capacity and disposition to recognize and act in one's own interest and that of chosen others. The authors also make the distinction that knowledge, skills and social context are pre-conditions for agency, where social context can enable or preclude the expression of human agency. Cauce and Gordon (2012) made a number of recommendations related to the promotion of human agency and how it can be incorporated into the teaching, learning and assessment process. They recommend that researchers and educators should develop methods for promoting students' agency by emphasizing to youth the links between effort, goals, and fulfillment. However, the authors also recognized the need for research that explores how educators and administrators support or sometimes preclude the development of student agency. They also thought about whether the future of assessment should consider agency as a product of education and not just a process. In terms of the assessment of human agency (not just for individuals, but also the assessment of collective agency), the authors recommended the development of measures of agency, specifically – the development of holistic measures of agency beyond survey methods, to provide robust data for understanding learners' agency, its emergence and its expression. The assessment of collaborative agency is supported by the work by Bereiter and Scardamalia (2012) related to collaborative learning and the development of assessments to measure learners' contributions to the collaborative problem solving.

The Armour-Thomas and Gordon (2012) paper, *Toward an Understanding of Assessment as a Dynamic Process of Pedagogy*, identifies dynamic pedagogy as a form of pedagogy that integrates assessment, curriculum, and instruction and finds this interaction to be instrumental for student learning. The assessment strand promotes real-time and metacognitive probes in order to provide iterative dynamic feedback to promote higher-order thinking and to inform adaptive instruction. The curriculum strand uses multiple resources to engage students to think about concepts and procedures in multiple ways, promoting the likelihood that student will learn more deeply about the content of a discipline. The instructional strand promotes scaffolding techniques and critical thinking skills.

The assessment strand of dynamic pedagogy has two components: a probing component and a metacognitive component. The probing component includes seven aspects that are able to probe into different aspects of student learning: prior knowledge, skills and readiness for new learning emerging understanding of new concepts as well as misunderstandings; acquisition of new knowledge and skills; ability to demonstrate their knowledge and skills with automaticity; how well they have consolidated their new knowledge; how well they are able to transfer to other contexts; check the mental processes encouraged during learning; and check disposition and motivational level while language in learning tasks. With the idea being – that all of this would be formative and used to inform and adapt instruction. The metacognitive component evaluates the extent to which students use and demonstrate their awareness of effective learning strategies. In addition, these students demonstrate that they know how to use them and recognize them when they are applied. Armour-Thomas and Gordon (2012) recommend a reconceptualization of pedagogy as a dynamic process in which assessment, curriculum and instruction work together to support student learning. In this process, there would be a model of formative assessment as part of the dynamic pedagogy, given that assessment should provide feedback that can be used to adapt curriculum and instruction to optimize learning. Then, the data from the formative assessments could be used to make appropriate selection of texts, tasks, and processes that promote learning of subject matter, allowing for the adaptation of the curriculum to students' needs.

Armour-Thomas and Gordon (2012) provide a set of recommendations related to learner-centered assessment. Specifically, it was recommended that they count in the evaluation of teaching and learning, that computer technologies should be used in their development, and ensure the validity and fairness of these assessments. In support of dynamic pedagogy, the authors recommend changing the way in which teachers are prepared for their service, so that assessment is viewed as a practice that is embedded in daily practice.

Baker's (2012) paper, *Testing in a Global Future*, examines the assessment options in the context of international comparisons, demographic changes, knowledge expansion, job changes, and technological growth. Given the unsurpassed personal access to knowledge presents a challenge for those who try to maintain control and authority over learning. Baker foresees that there will be a major blurring between classroom and other informal learning and assessment. This change will place increasing responsibility on students to be responsible for their own learning, rather than conform to directives from adults. This observation is consistent with the recommendations made by Torre and Sampson (2012) regarding the importance of collaborative self-evaluation. Students and teachers will have to work together to make sure that students have the tools that they need to make learning a more personal and active process, as students become more responsible for evaluating and keeping track of their own learning progressions.

Baker recommends the development of tests and assessments that require/measure the transfer or the application of learning to new, unexpected tasks. This should be done instead of focusing on outcomes relevant to existing standards and knowledge. Policies should be developed that require assessments to serve this purpose. Assessments should focus on a set of more pervasive skills that could be embedded in unforeseen different contexts and changing subject matter, directed toward new applications. Baker noted the evolving nature of games, which are now trying to systematically affect learning. Assessments will need to change rapidly to take advantage of the technology, and to meet learners' expectations. Assessments will need to be personalized and fully adapt to the interests, formats, and expectations of individual learners.

Lessons Learned from Testing Special Populations

While the papers within this section, Hakuta, 2012; Thurlow, 2012; and Boykin, 2012; address specific populations of students (i.e., English language learners and students with disabilities), their view of assessment questions the current way in which groups are identified to receive alternate assessments or receive accommodations in testing. The papers consider how some of the accommodations may be helpful to learners beyond those that have been identified as having a disability (e.g., Universal design) or those who may be English language learners (e.g., bilingual class for English language learners and native speakers of English).

In Hakuta's (2012) paper, *Assessment of Content and Language on the Heels of the New Standards: Challenges and Opportunities for English Language Learners* finds that current standards encourage interdisciplinary construction of arguments without providing linguistic scaffolding for English Language Learners (ELLs). Hakuta describes three important issues concerning English language learners: moving students from being

English proficient and fluent to having academic literacy; the role that students' native language has in English instruction, and the controversy around bilingual education; as well as considering the advantages of bilingualism for both the English-language learner and the native English speakers participating in the bilingual programs. Hakuta recommends that assessment should be formative, continuous, and interactive. Hakuta advises creating an interdisciplinary and applied field that addresses the complex needs of English-Language Learners. Specifically, Hakuta advocates the use of bilingual assessments that measure language skills through authentic translation tasks. Hakuta would like to see assessments for ELLs that not only capture basic language acquisition but also academic language knowledge acquisition.

Thurlow's (2012) paper, *Accommodation for Challenge, Diversity and Variance in Human Characteristics*, discusses recent developments in accommodations research, particularly universal design and that universal design has made accommodations a part of the process of assessment rather than separate from it. In addition, in the context of universal design, this makes the accommodations relevant for all types of learners. Thurlow poses the question of whether or not group characteristics should define accommodations rather than individual student needs. Thurlow (2012) recommends conducting research to examine the effects that accommodations have on students with and without a disability label. Thurlow also calls for a better understanding and accounting for factors that may affect student performance beyond a disability, race, language or gender. That being said, Thurlow does acknowledge the value of past findings related to factors that may contribute to observed group differences and believes that they should not be discounted. In addition, Thurlow calls for a move away from theories of differential boost and thinking of students as defined not by their disability, but by their learning needs. In addition, as a means for increasing access to college, Thurlow recommends creating more congruency between K-12 settings and assessment and college entrance exams and settings.

Boykin (2012) opens his paper, *Human Diversity, Assessment in Education and the Achievement of Excellence and Equity,* with the point that while Binet, one of the originators of ability testing was charged with developing a procedure by which uneducable students could be easily identified, assessment in the U.S. has historically been conceived as a process to provide greater opportunities for persons from diverse backgrounds. In terms of present day understanding of diversity in ethnic and social backgrounds, assessment practices can easily be seen as serving exclusionary purposes for individuals whose experiences are construed as outside the mainstream of our society. Binet's uneducatables tend to be over-represented by persons from the lower-status divisions of human societies. Boykin makes the case for not just assessment OF learning, but assessment FOR learning. He points out that greater attention should be given to classroom-based assessments directed at informing teaching and learning processes in the service the development of capacities in diverse populations of learners (assessment for learning).

According to Boykin, expanding the reach of assessments to include not just assessments of students, but also assessments of educational processes and their contexts in the service of informing teaching and learning, is an imperative. To do so will require expanding attention to issues of validity to include matters of consequences for the diverse populations served by education and its assessment. Boykin argues that educational assessments and formal education should emphasize human capacity building rather than sorting and selecting, unless the sorting and selecting is directed at the meeting the special needs of students so identified. The author asserts a special role for assessment having to do with providing access to specialized opportunities for individuals from diverse backgrounds and of varied constellations of abilities and challenges. The special challenge to assessment that better serves diagnostic functions is recognized as having been made easier by the advent of: the combination of advances in understanding of human information processing and electronic information management that can enable adaptation and personalization in assessment, teaching and learning.

Technology as a Tool to Advance Assessment

The papers within this section (Hill, 2012; Chung, 2012; and Behrens & DiCerbo, 2012) highlight how developments in technology allow for the development of more advanced, more comprehensive assessment systems that can provide varying levels of data to inform the teaching, learning, and assessment process. Specifically, technology will allow for the collection and management of fine-grained data throughout the teaching, learning, and assessment process that can be used to monitor and inform student learning.

Hill's (2012) paper, *Assessment in the Service of Teaching and Learning*, identifies the use of digital technologies as a means for enhancing our current assessment system and those technologies can be used to help understand student performance more clearly. Hill recommends conducting periodic assessment activities throughout the school year to produce a rich archive of student development and learning. Hill also views digital technologies as a means for diminishing the lines between testing situations, instruction and learning. This is aligned with what Armour-Thomas and Gordon (2012) identify as dynamic pedagogy, a means for integrating teaching, learning, and assessment. Hill also recommends the involvement of teachers in the evaluation of student learning using scoring rubrics, as this may be a mechanism for heightening teachers' awareness of the learning goals and objectives.

Hill also sees digital technologies as a means for bringing assessment practices closer to real-world activities and subsequently real-world competencies. Hill identified challenges with implementing digital technologies, particularly authenticating students' work and the cost of paying evaluators to score the products of the digital technologies. To address these challenges, Hill identified solutions in terms of implementing a feedback process by which students' work would be viewed in stages and the use of automated scoring to alleviate some of the time and cost associated with human scoring of the work products.

In Chung's (2012) paper, *Toward a Relational Management of the Data of Educational Measurement*, he discusses the new opportunities and challenges that present themselves in the measurement of students' learning processes using technology-based tasks. Technology-based tasks provide an opportunity to develop more individualized instruction and learning experiences. This opportunity will present the challenge of how to effectively manage and use the huge quantities of learning-related data that will be generated as a result of these technological advances. The large body of fine-grain data on student behavior in digital environments will lead to new inferences about student-learning processes. Chung identified the key issues as: leveraging data to measure what student understand and can do, deriving meaningful measures of cognitive and affective processes, and developing capabilities for precise diagnosis and targeting of instruction.

In addition, Chung (2012) identified different levels of data aggregation that can be used to answer different kinds of questions (i.e., system-level data, individual-level data, and transactional-level data). System-level data can be used to answer questions about student retention rates, graduate rates, and time to degree. Individual-level data has been seen as the finest grain-size used in educational measurement. However, more recently there has been an interest in the use of data at an even finer grain level. This has been made practical through technology-based applications. These finer-grain data would fit into the category of transactional-level data. Transactional-level data reflect a student's interaction with a system (i.e., moment to moment choices on some task) where the interaction may be an end in itself or a means to an end.

Chung also discusses the higher use of learning analytics, which van Barneveld, Arnold and Campbell (2012) defined as the "use of analytic techniques to help target instructional, curricular, and support resources to support the achievement of specific learning goals." Chung (2012) made a number of recommendations related to the use of fine grained and transactional level data to develop capabilities for precise diagnosis and targeting of instruction. Chung also recommends developing adaptive educational systems that use these fine-grain data related to student learning to adapt instruction to individual needs in order to enhance learning processes and outcomes. Chung advises putting the data, and its representations, in the hands of educators so they can be better equipped to make decisions about how to adapt instruction to support student learning. Although Chung recommends the use of fine-grain transactional data to support student learning, Chung acknowledges that more empirical research is needed to fully understand the relationship between fine-grain data, learning processes and educational outcomes. Chung's recommendation to use fine-grained and transactional-level data in conjunction with existing data to advance the field of assessment and measurement is in line with Gorin's (2012) recommendation to consider various evidential sources.

The Behrens and DiCerbo (2012) paper, *Leverage Points for "Natural" Digital Activities in the Assessment of Human Attributes*, describes three core aspects of technological developments that can used for educational assessment: a) computers can be used to enhance human capabilities given computers' ability to store, process and mine large amounts of fine-grain data from multiple sources; b) the increased use of digital technologies makes it possible to gather new forms of data based on human interaction in digital environments; and c) digital technologies can be used to better visualize the fine-grain data so that observations, patterns and inferences can be made based on the data. These new technologies should allow new insights into student learning using computational methods of storing, analyzing and modeling student data. Behrens and DiCerbo (2012) recommend a reframing of assessment practices from identifying correctness of test questions to capturing a constellation of learning transactions using digital technologies to make inferences about student cognition and learning. This is supported by the recommendations of Gorin (2012) to have multiple sources of evidence to develop an evidentiary argument for the knowledge, skill, and abilities of learners. Behrens and DiCerbo also recommend shifting from an individual paradigm to a social paradigm.

A number of authors have identified the need to provide more opportunities for student learning within collaborative environments and into the teaching, learning, and assessment process (Behrens & DiCerbo, 2012; Bereiter & Scardamalia, 2012; Cauce & Gordon, 2012). Behrens and DiCerbo believe that this would bring assessment closer to the conditions where existing theory and empirical research suggest learning occurs (i.e., through collaboration). Lastly, Behrens and DiCerbo recommend a shift from assessment situations to assessment ecosystems, where this shift would help to counter the disruptive nature of the present assessment paradigm by building on the digital technology to extract data unobtrusively and provide ongoing rich data on student learning.

The findings of the Gordon Commission, which largely grew from the Knowledge Synthesis Project, can be found in the section: The Findings and Recommendations of the Gordon Commission.

References

Armour-Thomas, E. and E. W. Gordon (2012). Toward an Understanding of Assessment as a Dynamic Component of Pedagogy.

http://www.gordoncommission.org/rsc/pdf/armour_thomas_gordon_understanding_assessment.pdf

Baker, E. (2012) Testing in a Global Future.

http://www.gordoncommission.org/rsc/pdf/baker_testing_global_future.pdf

Behrens, J. and K. DiCerbo (2012) Technological Implications for Assessment Ecosystems- Opportunities for Digital Technology to Advance Assessment.

http://www.gordoncommission.org/rsc/pdf/behrens_dicerbo_technological_implications_assessment.pdf

Bennett, R. E. (2012) Preparing for the Future—What Educational Assessment Must Do.

http://www.gordoncommission.org/rsc/pdf/bennett_preparing_future_assessment.pdf

Bereiter, C. and Scardamalia, M. (2012) What Will It Mean To Be An Educated Person in Mid-21st-Century.

http://www.gordoncommission.org/rsc/pdf/bereiter_scardamalia_educated_person_mid21st_century.pdf

Boykin, A. W. (2012) Human Diversity, Assessment in Education and the Achievement of Excellence and Equity.

http://www.gordoncommission.org/rsc/pdf/boykin_human_diversity_assessment_education_achievement_excellence.pdf

Cauce, A. M. and E. W. Gordon (2012) Toward the Measurement of Human Agency and the Disposition to Express It.

http://www.gordoncommission.org/rsc/pdf/cauce_gordon_measurement_human_agency.pdf

Chung, G. (2012) Toward the Relational Management of Educational Measurement Data.

http://www.gordoncommission.org/rsc/pdf/chung_toward_relational_management_educational_measurement.pdf

Dixon-Roman, E. and K. Gergen (2012) Epistemology and Measurement, Paradigms and Practices, Part 1, A Critical Perspective on the Sciences of Measurement.

http://www.gordoncommission.org/rsc/pdf/dixonroman_gergen_epistemology_measurement_paradigms_practices_1.pdf

Gergen, K. and E. Dixon-Roman (2012) Epistemology and Measurement, Paradigms and Practices, Part II, Social Epistemology and the Pragmatics of Assessment.

http://www.gordoncommission.org/rsc/pdf/dixonroman_gergen_epistemology_measurement_paradigms_practices_2.pdf

Gordon, E. W. and E. B. Campbell (2012) Implications for Assessment in Education.

http://www.gordoncommission.org/rsc/pdf/gordon_campbell_implications_assessment_education.pdf

Gordon, E. W., E. W. Gordon, D. Berliner, and L. A. Aber (2012) Changing Paradigms for Education.

http://www.gordoncommission.org/rsc/pdf/gordon_gordon_berliner_aber_changing_paradigms_education.pdf

Gorin, J. (2012) Assessment as Evidential Reasoning.

http://www.gordoncommission.org/rsc/pdf/gorin_assessment_evidential_reasoning.pdf

Hakuta, K. (2012) Assessment of Content and Language in Light of the New Standards, Challenges and Opportunities for English Language Learners.

http://www.gordoncommission.org/rsc/pdf/hakuta_assessment_content_language_standards_challenges_opportunities.pdf

Hill, C. (2012) Assessment in the Service of Teaching and Learning.

http://www.gordoncommission.org/rsc/pdf/hill_assessment_service_teaching_learning.pdf

Ho, A. (2012) Variety and Drift in the Functions and Purposes of Assessment in K-12 Education.

http://www.gordoncommission.org/rsc/pdf/ho_variety_drift_functions_purposes_assessment_k12.pdf

Kaestle, K. (2012) Testing Policy in the United States- A Historical Perspective.

http://www.gordoncommission.org/rsc/pdf/kaestle_testing_policy_us_historical_perspective.pdf

Linn, R. (2012) Test-Based Accountability.

http://www.gordoncommission.org/rsc/pdf/linn_test_based_accountability.pdf

Mendoza-Denton, R. (2012) A Social Psychological Perspective on the Achievement Gap in Standardized Test Performance Between White and Minority Students, Implications for Assessment.

http://www.gordoncommission.org/rsc/pdf/mendoza_denton_social_psychological_perspective_achievement_gap.pdf

Meroe, A. S. (2012) Democracy, Meritocracy and the Uses of Education.

http://www.gordoncommission.org/rsc/pdf/meroe_democracy_meritocracy_uses_education.pdf

Mislevy, R. (2012) Four Metaphors We Need to Understand Assessment.

http://www.gordoncommission.org/rsc/pdf/mislevy_four_metaphors_understand_assessment.pdf

Thurlow, M. (2012) Accommodation for Challenge, Diversity, and Variance in Human Characteristics.

http://www.gordoncommission.org/rsc/pdf/thurlow_accommodation_challenge_diversity_variance.pdf

Torre, C. and Sampson (2012) Toward a Culture of Educational Assessment in Daily Life.

http://www.gordoncommission.org/rsc/pdf/torre_sampson_toward_culture_educational_assessment.pdf

Varenne, H. (2012)- Education Constraints and Possibilities in Imagining New Ways to Access Rights, Duties and Privileges.

http://www.gordoncommission.org/rsc/pdf/varenne_education_constraints_possibilities.pdf

2. A HISTORY OF THE ASSESSMENT OF EDUCATION AND A FUTURE HISTORY OF ASSESSMENT FOR EDUCATION[6]

The intended audience for this essay is education practitioners and policymakers who may not have a detailed knowledge of the history of American education. The essay's purpose is to reflect on the development of modern testing practices in a historical context. This can spur ideas on how to shape assessments to fit our 21st-century values.

Histories of testing often gloss over 19th-century written exams in the U.S., implying that testing only describes standardized exercises used for high-stakes decisions. However, everyday assessment by teachers has always occurred. The rote memorization and recitation prevalent in the 19th-century is not a model for today, but nor should we assume that the dominance of standardization is inevitable in an education system. Another form of assessment prevalent in the past was exhibitions. Teachers prepared their students to perform oratory, musical, and other skills for their communities. For parents and school committees it was an occasion to be entertained and assess the skills of the teacher. At the same time, exhibitions were *not* used to make promotional decisions about children.

Horace Mann led a reformist reaction against the reliance on exhibitions and teachers' autonomy, championing periodic written exams in Boston. The reformers also pushed for more humane teacher training and the creation of a superintendent to oversee schools. The tests became popular in cities nationwide to answer questions of comparability and high school admissions. While exhibitions persisted, educational tests and standardized textbooks facilitated the centralization of educational authority. There was no lack of opposition to tests in the mid-19th-century. Critics decried the focus on memorization over understanding; however, the traditional focus on rote learning aligned well with the competitive testing culture. When critics succeeded in having a test phased out, it was often merely replaced by another test. Criticism faced a hostile environment. High-stakes tests are not new, but they have not on their own led to a reduction in test use.

[6] Abstracted from Kaestle, C., 2012, *Testing Policy in the United States: A Historical Perspective* and amended by Edmund Gordon http://www.gordoncommission.org/rsc/pdf/kaestle_testing_policy_us_historical_perspective.pdf.

Intelligence Testing: The Early Years

Early professional psychologists engaged in research on explaining correlations as a way to provide comparable measures of accomplishment in a radically decentralized educational system. A focus on differences and inspiration from Darwinian genetics led many researchers to focus on finding g, a person's summary and presumably inherited level of intelligence. Working along a different line of inquiry, Alfred Binet developed scales of increasing item difficulty in order to assist in the classification of children along a timeline of "normal" development.

Synthesizing the hereditarian tradition and Binet's scales, Lewis Terman adapted and popularized the term "intelligence quotient," a figure derived by dividing the mental age score by the subject's chronological age. Terman's production of the Stanford-Binet intelligence test brought him fame as the president of the American Psychological Association. He used IQ tests to promote eugenicist ideas, including immigration restrictions and sterilization of low-IQ people.

IQ Testing in the Era of World War I

The use of IQ tests to classify army recruits during World War I boosted the tests' prestige. While the military shut down the use of the tests because of objections about the utility of English language intelligence tests for non-English speaking soldiers, public schools embraced them. IQ testing appealed to educators' regard for efficient, scientific decision-making about teachers and students in an era of larger systems and greater student diversity. The professionalization of bureaucratic management in education encouraged school superintendents to use tests to make placement decisions for students and teachers. The National Research Council produced a National Intelligence Test available in dozens of alternate forms. In 1920, its first year, the National Intelligence Test sold 200,000 copies.

Contemporary critics questioned the heritability and immutability of IQ. Walter Lippmann argued in 1922 that the tests only measured testers' guesses about what questions represented intelligence and could have negative consequences for children. However, Lippmann argued that if tests were understood to be measures only of the tasks they contained, and not as measures of innate ability, they would have valid purposes for placement. He rejected IQ testers' claims not because of their appeal to scientific management, but because of their unscientific nature.

The racist and sexist hereditarian positions commonplace in the early 20th-century proclaimed the mental inferiority of Blacks to Whites and women to men. While heredity is one of a constellation of factors related to mental ability, the IQ test advocates used hereditarian logic to argue against women's participation in education as students and teachers. Research by Helen Woolley, Leta Stetter Hollingworth, and others gradually debunked the pseudo-scientific evidence for female mental inferiority, although many barriers to women's progress remained.

Schools of education in the 1920s codified theories of psychology, curriculum, and administration for the efficient management of schools. The industrial metaphor was used to set standards for performance. Knowledge was conceptualized according to the tenets of early behaviorist psychology, as a matter of stimulus (question) and response (answers) developed through exercise (practice) and feedback (rewards). The schools of education that promoted these theories became training grounds for a generation of researchers and administrators.

The Expansion of Achievement Testing

Achievement testing in the early 20th-century developed into standardized multiple-choice measures. They surpassed intelligence tests because of their diverse subject matter—a different test was needed for different academic domains. Achievement testing was also more useful for evaluating instruction because, unlike IQ tests, they provided comparative data across classrooms and schools that were relevant to the schools' curricula. In addition, by making narrower claims about ability than intelligence tests,' they were less vulnerable to criticism than IQ tests that claimed to measure permanent attributes.

In the 1920s and following decades, hundreds of achievement tests were written to measure performance on a wide range of subjects. An industry grew up around these tests, scored by hand in the early years but eventually by machine. Testing research and development drove the discourse of professional educators. While some surveys of achievement noted the limitations of contemporary measurement, the number of cities using the tests to track progress grew steadily through the 1920s.

Colleges were drawn in to the testing regime, and the College Board was established to develop standards and examinations for college entrance. The Scholastic Aptitude Test was a direct descendant of the Army Alpha intelligence test from World War I and was immediately put to work—against the College Board's advice—in predicting later college performance. The test's developer, Carl Brigham, tried to distance the test from IQ testing and opposed its implementation as a major college admissions test. After his death, the College Board and the new Educational Testing Service promoted the SAT as a test independent from school learning and that would usher in a more meritocratic admissions regime, not anticipating the rise of test-preparation services used primarily by the affluent. Standing at the mid-ground between intelligence and achievement tests, the SAT became and has remained very influential in college admissions.

Prelude to Reform: The 1940s and 1950s

World War II left the United States as a superpower with an expanding education system featuring greater high school and college attendance and a growing research sector. The desire to institute a meritocracy within schools relied on an expansion of testing to provide evidence for sorting decisions. Students were chosen for more or less difficult courses based on ability but not necessarily on a single "track" in all subjects. Despite some critiques that testing was biased against the working class, the use of tests continued to expand in the late 50s.

Although there were no adequate measures to compare other countries' educational achievement in the 1950s, the launch of *Sputnik* was enough evidence for Congress to increase federal aid for elementary and secondary education. The National Defense Education Act of 1958 reinforced the central role of testing in U.S. schools by tasking states with using tests to identify the best potential scientists. Confidence in this meritocratic scheme was soon to be questioned, however.

Testing and Civil Rights

The activist 1960s phase of the civil rights movement brought a concern for equity to the testing endeavor. This took the form of both critiquing test bias and promoting targeted testing to ensure that schools were serving different groups well. Worries about the disparate impact of tests on diverse groups of students led to a decrease in the use of some tests for placement and in court decisions and efforts to eliminate cultural bias in other tests like the SAT.

The focus shifted from equal opportunities to equal outcomes. Judges prohibited the use of tests to re-segregate black students through tracking or placement for special education classes on the grounds that the tests weren't valid for that use. Congress required the scientific validation of tests used to sort children into special education programs. Restrictions on the use of tests were counterbalanced by an increased emphasis on the use of tests to identify and evaluate remedies for the highly publicized achievement gap, especially at the federal level. Reducing the achievement gap became a primary goal, in some cases a substitute for school integration. This would require holding teachers, schools, and districts accountable for their results.

The Federal Government and the Birth of National Assessment

Francis Keppel arrived in Washington as Kennedy's Commissioner of Education and was frustrated to learn that the Office of Education had no data about student learning. He worked with John Gardner of the Carnegie Corporation to come up with a system for testing a sample of children to gather benchmarks of learning. The result was the National Assessment of Educational Progress, which combined student sampling, matrix sampling (no individual answers every item), scoring related to expert-determined standards, and a mixture of multiple choice and short answer items. Word of NAEP's development prompted criticism that a national assessment would make schools compete with one another and drive good teachers away form teaching lower performing students, damaging the educational equity project. Further, national testing would narrow curricula and stifle local experimentation.

NAEP's designers answered the criticism by insisting that the results wouldn't be broken down by individual students, classrooms, or schools. The test's design attempted to assess progress without making unfair comparisons. The goal was to help the Office of Education report responsibly on the effectiveness of federal education programs. Keppel and others worked successfully to persuade Congress that evaluation of federal efforts had to go beyond describing how money was spent. In 1988 Congress authorized the release of NAEP scores by state. Assessment experts resisted any further disaggregation, arguing that any use of NAEP for individual scores would compromise its value as a monitoring device.

Accountability as a Major Goal of Assessment

Parallel to the efforts to establish the NAEP, other developments pushed the ideas of basing policy decisions on student-learning measurements. In 1965 Robert Kennedy raised the issue of whether ESEA would require districts to produce any evidence on effectiveness. He believed that many schools serving poor minorities were part of the problem with their performance. The result was an amendment to ESEA that required Title I districts to devise their own tests and report the results periodically to state education agencies.

ESEA was initially poorly funded and weakly targeted. The conclusions from the limited samples examined in early evaluations of Title I questioned the ability of so little money, spread so far out, to have an impact. The money was for the most part used as general assistance. Compounding the issue was that the Office of Education had no tradition or expertise for evaluation. There was little accountability for the failure to make progress towards the program's stated goals.

Nonetheless, the aspiration to use program outcomes to justify future government programs investment was boosted when President Johnson implemented the Pentagon's Planning-Programming-Budgeting System (PPBS) system for all government agencies in 1965, requiring independent performance analyses for their programs. The cost-benefit analyses required by PPBS challenged local and state school administrators' desire to use ESEA funding as general aid. The emphasis on outcomes would become a central focus of policy analysis.

The emphasis on outcomes was further reinforced by James Coleman's study, *Equality of Educational Opportunity*. The report's massive statistical study—a landmark both in methodology and in policy implications—concluded that schools' physical resources were less important to a student's achievement than the ability of their teachers, their family background and class, and their belief in their ability to control their fate. The report questioned the utility of spending resources on schools and implied that social class integration was more important than race *per se*. The report had little political impact due to obfuscation by the Office of Education and resistance by educators, but it had long-term influence on the educational research community because of its quantitative focus on outcomes and implication that schools cannot solve social problems alone.

The 1970s

Despite a Republican administration, civil rights movements continued to win new victories in establishing bilingual education, banning sex discrimination in education programs receiving federal funds, and encouraging participation of children with disabilities in regular classrooms. All increased the amount and significance of testing. Meanwhile, the minimum competency movement sought to solve unemployment via tougher high school graduation requirements enforced through testing, part of a growing accountability movement. The debates over IQ and race continued.

The 1980s: A Nation at Risk?

During the 1980s much of the impetus of school reform moved from the federal government to the states due to President Reagan's opposition to a federal role and the growing concern amongst governors that education was critical to their state's economic competition. The 1983 Department of Education report *A Nation At Risk* fueled these anxieties with a focus on international economic competition. Reformers aimed to raise standards for homework, attendance, graduation, and admission to teacher preparation programs.

The first President Bush attended a "summit" of governors calling for renewed reform in a partnership between the federal and state governments. In the 1988 reauthorization of ESEA, Congress required states to define the levels of achievement students should attain and identify schools that didn't meet the goal. State-level achievement testing had become one of the main instruments of reform. Proponents argued that high-quality tests would be worth "teaching to," while critics insisted that pervasive high-stakes testing inevitably led to test-savvy drilling in disconnected bits of knowledge. The debate turned on an anomaly in American testing: the ubiquitous multiple-choice question reflects a behaviorist theory of learning that hasn't been in the psychological mainstream since the 1950s. The ascent of cognitive psychology transformed the field and influenced curriculum development, but was not reflected in assessment practice.

The 1990s

The 1990s saw the development of two important frameworks for assessment policy: standards-based reform (SBR), which has become the mainstream policy, and performance assessment, which faced considerable challenges. SBR is based on content standards that define what students should know and performance standards that define by when students should be able to perform what tasks with that knowledge. SBR promises alignment with instruction and assessment and to serve the goals of both excellence and equity.

Assessment experts and school policy reformers also promoted performance assessments. Performance assessments more closely resemble real-world demonstrations of skills, and as such are necessary for assessing complex high-end capabilities. They promise to combat the practice of teaching to the low-level skills emphasized by multiple-choice tests. Performance assessment, however, is more costly and threatens existing standardization frameworks, raising questions about comparability. Because it is necessarily more integrated into teaching and learning, it would require large-scale teacher retraining. These barriers to implementation ultimately marginalized performance assessment as a policy choice. In Vermont, for example, early successes in implementing performance portfolios were quashed when the Vermont Supreme Court mandated an equalization process across districts based on “traditional” evidence of achievement.

On the federal level, President Clinton championed standards-based reform. The administration devised a bill called “Goals 2000” to require states to develop standards and submit them for federal approval, but a Republican resurgence in Congress in 1994 prevented its implementation. Clinton was left to press for SBR in an advisory mode, while using the reauthorization of ESEA to require more test-based evaluation of Title I schools.

Into the 21st Century

Like Clinton, President George W. Bush was determined to bring standards-based reform to the federal level. No Child Left Behind was a collaboration between the administration and liberal Senator Edward Kennedy, who saw an opportunity to insure Title I funding and implement the disaggregation of achievement scores by class, race, and ethnicity. NCLB required definitions of satisfactory yearly progress, included radical sanctions including the dismissal of staff, and set implausibly high goals for progress. Despite opposition, the civil rights aspect of SBA focused on provisioning more resources to underperforming schools and demographic groups earned the support not only of Bush loyalists but civil rights advocates.

The Obama administration declared the NCLB sanctions unproductive and gained massive influence over states and districts due to federal stimulus funds following the financial collapse of 2008. These provided supplementary money to states that complied with administration priorities for creating more charter schools and making teachers, salary raise dependent partially on student achievement test scores. Partisan differences prevented the reauthorization of ESEA during the first Obama administration; meanwhile, the administration forestalled the consequences of NCLB sanctions through waivers to the states, also tied to administration reform priorities.

Both the Bush and Obama administrations have promoted a framework of standards-based reform with strong federal oversight. It remains to be seen if the political will and organizational capacity exist for school systems to reduce achievement gaps without broader social and economic reforms. SBR is bolstered by considerable bipartisan consensus, but criticism of the narrowing of curricula under excessive testing is widespread among both academics and teachers. The concern is that teaching for understanding, familiarity with literature, and development of independent reasoning are being stifled with a focus on lower-level skills.

The Common Core

Among other reforms, the Obama administration has required that states join multi-state consortia to develop new standards and assessments. The largest group, supported by the National Governors' Association and Council of Chief State School Officers, proposed a set of aligned standards and assessment called the Common Core. The Common Core evades fears about federal control due to its *national but not federal* nature. Many administrators and other stakeholders are drawn to the economy of scale promised by many states working together to develop complicated standards and assessments.

The leaders of Common Core aspire to first-rate standards and assessments, but some of the most complicated of the envisioned assessments require more test time or the use of computers. Many states and districts are resisting these features. Common Core and its assessments are a work in progress. Judgments on their effectiveness will not be possible for several years.

Reflections: Lessons from History?

The history presented in this essay shows that assessment policy is necessarily a struggle between competing values. Standardized multiple-choice tests of basic skills have costs for teaching higher-order knowledge and representing diverse student capabilities, but they serve important equity and accountability goals by describing patterns of comparable performance across individuals and groups. Nonetheless, a new balance is needed between these different goals.

Present testing practices have powerful support because many people accept them as defining educational accomplishment. They are cheap and appeal to the popular priority placed on factual knowledge as a primary purpose of schooling. Those testing practices have over time been expanded to make judgments about different stakeholders, from individual students to teachers, schools, districts, states, and entire demographic groups. Test practices have changed over time—to a limited extent—to reflect evolving concepts of what intelligence is and what skills we wish to foster in our students. But an abiding continuity is the use of standardized, multiple choice or short-answer test items.

The Current Moment in Assessment Reform

Current reforms promise to better integrate assessment with teaching and learning, better represent recent thinking on learning, and to take advantage of new digital tools. Because they are more costly and difficult for lay people to understand, their advocates will have to develop concrete descriptions of what is possible and why it is urgent that such changes be made. Those descriptions must be as accessible to lay people and educators as they are to policymakers. At the same time, reforms must be strong and thorough enough to displace the negative features of the current testing regime: the obsession with accountability (the assessment *of* education) to the exclusion of a concern for improved teaching and learning (assessment *for* education). Today, we are focusing too much on low-level, fragmented knowledge at the cost of stifling creative teaching in a culturally sensitive context.

Kaestle concludes that the Gordon Commission comes at an auspicious moment to make the argument for fundamental reform. The work done so far in the 21st-century, on cognitive science, computer-assisted learning, and specific models of assessment reform, gives the author hope. To attain assessments integrated into a 21st-century education, however, we must apply not just our traditions but also our knowledge to the future.

Gordon's Perspective

In my view, the history of assessment in education is largely an account of long-time efforts of human societies to document the status, describe the characteristics, and measure the achievements of learning persons. This emphasis on the status of one's characteristics and developed abilities has tended to be used as measurements of the effect of, or the need for, education. This long history of human effort has resulted in the emergence of a highly developed system of science of the measurement of education. This science and its techniques have been embraced in the professionalization of prediction, selection, and certification allocation. However, this emphasis on assessment of education has not been as effective as instrumentation for informing and improving teaching and learning. The emphasis on measurement of status tends to neglect attention to teaching and learning processes, the potential capacities of the learner, and the process of becoming—which is at the heart of the teaching and learning transaction. The work of the Gordon Commission has maintained a bifocal emphasis, sensitive to the rich history of the assessment ***of*** education, and concerned with a future history of assessment ***for*** (to serve, inform and improve) education.

This concern with assessment ***for*** education is by no means new. Shortly after Binet produced his model for intelligence testing in an effort to assist in sorting out educable from uneducable students from the pool making demands on what was at the time a limited resource, Binet wrote his greatly neglected essay concerning the responsibility of a society capable of identifying those considered to be uneducable, for doing something to help those described as so limited. Western Europe and the United States, preoccupied with World War I and the selection of talent for their rapidly industrializing societies, embraced Binet's instrument for the assessment of, and ignored his recommendation of the need that we use assessment for education.

Some fifty years later, Else Haeussermann, Herbert Birch and I, challenged by the desire and need to plan education for children who had suffered damage to the central nervous system, developed an elaborate procedure for the assessment (evaluation) of educational potential in brain-damaged children (Haeussermann, 1957). We set out to design a set of procedures that could describe and document the processes by which children engaged academic learning, not so much what they could not do or what they knew how to do, but how they went about using or not using what they had. We were intent upon providing information that would be used by teachers to inform and improve the teaching and learning transactions for which they are responsible. Rather than measurement against standardized benchmarks, we sought to determine the conditions under which certain benchmarks could be reached, i.e., in what contexts could certain problems be recognized and engaged.

Teachers found the clinical reports from our assessment for informing and improving education to be enormously helpful, but we who developed these reports found the production to be excessively labor intensive. Our fellow psychologists objected to the absence of any metrics by which individual children could be compared to other children. Haeussermann retired, and subsequently went to her death regretting that she had not produced a standardized form of her assessment. I would memorialize her failure to do so as a monument to the future history of assessment – assessment in the service of education.

Working and writing at about the same time as Haeussermann was Mary Meeker, who, in despair at all the information available from standardized tests that was left unused, developed a set of templates that could be used to analyze the data from standardized tests to reveal the indications of mental activity that lay camouflaged in the data of some standardized achievement tests. When Messick and I revisited this work, we tried to unbundle selected test items to reveal the nature of the intellective demands behind the items. This information, if made available to teachers, we theorized, could be used to help students better understand the meaning of the test item, as well as the appropriateness and inappropriateness of the student's approach to the problem. Unfortunately, changes in circumstances and interest precluded continued work on this set of problems, but interest in the paradigm has persisted.

Readers will find that much of the attention of the Gordon Commission has maintained a focus on the assessment of education. It is difficult to avoid this focus on the assessment of education in a study that included analyses of the best of what we have done in assessment. Kaestle's history reveals assessment of education as glorious achievement of this field. However, the Gordon Commission is also charged with inquiry into the future of assessment in education. It is as we look into the crystal ball and even as we observe cutting-edge work and forerunners of the future of education and its assessment, that it becomes clear that while much has and can be learned from the continued assessment of education, rapidly emerging developments in education and its assessment will both demand and enable assessment that is in the service of informing and improving teaching and learning processes and outcomes. One paper written for the Gordon Commission has as its title, *Assessment in the Service of Education* (Hill, 2012).

3. THE CHANGING CONTEXT FOR EDUCATION AND ITS ASSESSMENT

Edmund W. Gordon

A host of contextual factors combine to influence the nature of education and its assessment. The persistent and dominant role played by families in the education and socialization of children has, perhaps, contributed to the emergence of schooling, relatively late in the history of human societies, as a central force. However, historical and contemporary analyses reveal that education is not and has never been co-terminus with schooling. Dewey, and later Cremin, and still later Gordon and Verenne, have emphasized the comprehensiveness and relational character of education, as less competing systems and more complementary systems for the facilitation of human development. Competition from religious institutions, political and social institutions, print media and now digital electronic information transfer (DEIT) could displace both families and schools as the principal sources of education (experiences and materials) in the 21st-century.

With these new media, teaching will become increasingly more self-directed and distant. Independent inquiry, thought, knowledge production and self-assessment will become more prevalent. The learning persons will gradually share space and role with the teaching persons as orchestrators of teaching and learning experiences. What is studied, and how, will more and more come under the control of learner choice and engagement. The knowledge and skill content of what will be learned and how it will be learned are more difficult to predict, since the paradigms that inform education will continue to change as will academic canons, and because learner choice and quality of engagement depend so heavily on the epistemological and political/economic contexts that shape both opportunity to learn as well as what is available to be learned.

Changing Paradigms

One of the three stated missions of the Gordon Commission on the Future of Assessment is to consider our best estimates of what education will become in the 21st-century and what will be required of the educational assessment enterprise by the middle of this century. In the pursuit of addressing that component of our mission, Commissioners and Consultants to the Commission considered a variety of anticipated and emerging changes in the paradigms by which the goals and processes of education are changing. Among these paradigms are such ideas as those that follow.

Led by such writers as William Butler Yeats, we see a shift from thinking about education as concerned with “filling buckets to lighting fires.” Increasingly the goals of education reflect the growing concern with encouraging and enabling students to learn how to learn and to learn to continue learning; to become enquiring persons who not only use knowledge but persons who produce and interpret knowledge. The pedagogical challenge will be less concerned with imparting factual knowledge and more concerned with turning learners on to learning and the use of their mental abilities to solve ordinary and novel problems.

The three Rs of Reading, wRiting and aRithmetic will continue to be essential skills, but thought leaders in education, Sir Kenneth Robinson is among them, increasingly point to varying combinations of Cs as essential processes in education. They are: Creativity and innovation, Conceptualization and problem solving, Communication and collaboration, and Computer literacy. These Cs are replacing the Rs as the modern ends toward which education is directed. Learning how to think critically and creatively, reason logically, interpret relationally, and to access and create knowledge will be more and more privileged in the 21st-century. The new century places high value on communication as reading and speaking, but also as listening and collaborating, and processing information from multiple perspectives. The capacity to recognize and even create relationships between novel and disparate inputs of information will be rewarded in this new century. The illiterate members of 21st-century societies will be those who cannot navigate the world of digital technology. Computer literacy will be a requirement of economic, educational and social intercourse, but it will mean far more than the ability to do word processing, social networking, and to play electronic games. Digitization will change the demands and opportunities of modern societies even more rapidly and radically than did industrialization, and, as a result, the processes of education and its assessment will change.

In the 18th, 19th- and 20th-centuries, we privileged decontextualization in the pursuit of precision in measurement and control in experimentation. When we turned to multivariate analysis to study complex phenomena, it was with a view to the sequential teasing out of the contribution made by each of several component variables, even while we were beginning to understand the notion of dynamic and dialectical interaction. The isolation of variables or components for the purpose of study may continue while the intent of such study is to know. However, as our purpose turns to *understanding* of the phenomena of the world and the relationships between these phenomena, experimenting, observing and measuring things out of the contexts in which they have developed and function will become more and more dysfunctional. Education and its assessment will have to become capable of capturing aspects of context, perspective and the attributions that come to be assigned to these conditional phenomena. The exactness and precision, which have been gained by decontextualization in the past, will be challenged by the situative and existential sensitivities required when contextualism and perspectivism are required for understanding as well as knowing.

In the interest of scientific validity, traditionally we have privileged “objective” knowledge over “subjective” information. We have been taught to try to control for or contain variance that is associated with affect and social/psychological situation. We have tended to examine cognitive functions independent of their contamination or being influenced by human biases and feelings. Yet, modern social and psychological sciences are pressing us to examine or assess human performance with greater respect for the influence of affective, emotional, situative and social processes. Evidence mounts in support of the fact that these processes influence the character and the quality of human performance, yet they are these instances of objectively documented human performance that are the source of the data of traditional assessments in education. However, assessment in education in the future will have to be more sensitive to subjective phenomena, i.e., to affect, attribution, existential state, emotion, identity, situation, etc., as will also the teaching and learning transactions in which learners are engaged.

Assessment of the outcomes of learning in the interest of accountability will be with us for a while, but the future is likely to bring increased concern for assessment for the purpose of informing and improving learning and the teaching processes that enable learning. Political pressure continues to support a preoccupation with the possibly inappropriate use of educational assessment data for accountability purposes, even though such practices are not supported by the empirical evidence, and some of us feel that such practices are actually counter-productive for the intended purposes. Pressure mounts from the profession and the practicalities of educational praxis for better information to inform intervention, prior to the search for better information by which to determine how well we are doing. We have known for more than a century that what we do in education is imprecise; that one model does not fit all; and that much of our intervention is under-analyzed trial and error. We believe that assessment in education can and should inform and improve teaching and learning processes and outcomes, without ignoring the importance of accountability. Whether the two purposes can be served concurrently and by the same assessment instruments and systems is one of the questions to be answered.

Humans will very likely continue to create technologies that make their work easier and that amplify and expand human abilities. Some of these, as with artificial intelligence inventiveness, could change the importance of some of the competencies for which we currently educate or, more likely, will exacerbate the need for other functions that we currently know less about enabling, i.e., agency, disposition, relational adjudication. The human ability-amplifying technologies may make some of our educational tasks easier, but they may also create monumental challenges and opportunities for the people who are responsible for assessing, teaching and learning in some well-orchestrated manner.

Just as human intellect is increasingly recognized to be a social phenomenon that is both experienced as and produced by social interaction and consensus, so also are teaching and learning. Even the learning we do “alone” benefits from the social transactions that have preceded it. Epistemic games and distance teaching and learning are examples of teaching and learning in isolation that depend on collective actions of others. Pedagogy of the future will need to reconcile the individual/social paradox of teaching and learning and the implications of this paradox for assessment. For assessment, such questions arise as:

- Isolated and collaborative performance as assessment contexts;
- Knowledge and skill retained in one’s mind and knowledge and technique accessed or generated in human social and human machine transactions;
- The limits of empiricism and contextualist/perspectivist dysconfirmation; and
- Systematized documentation of relationships between attribution, contexts, identity, and human performance.

Paradigms for education are constantly shifting in accordance with changing political/economic circumstances, demographic patterns, epistemological standards and technological advances. These drivers change the ways in which policy and test makers conceptualize and ultimately construct curriculum and assessment. As a result, inventory of the ways in which these paradigms are shifting and inquiry into the mechanisms that causes these shifts is central to conceptualizing the future of assessment and education. Current education practices need to consider the possible ways that future educational systems can meet, address, and re-envision the concept of education into the next 50 years. This paper attempts to outline the changing variables surrounding education, highlight the possible future conditions that these variables create, and conceptualize a future for education and testing that reconciles these two ideas.

Shifting Epistemologies

An important dimension of the work of the Gordon Commission on the Future of Assessment in Education has to do with the experience of paradigm changes in education and shifts in the epistemologies by which educational policies and practices are informed. A great deal of attention has been given to the exploration of changes in the ways that we teach and learn, changes in the processes of teaching and learning, and to shifts in the ways in which we think about the nature of the knowledge and techniques that we teach. Of perhaps even greater consequence is the concern with shifts in conceptions of the nature of what it means to know–that is, we, along with most who take education seriously, note the need to change how it is that we are educated, and how it is that we are educating within the United States.

The pedagogical troika–assessment, teaching and learning–gives reasonable entry to change with constructs that are relatively accessible, albeit not necessarily as straightforward as one might generally think. Epistemologies, of course, are frequently discussed in inquiries concerning learning about learning, and in figuring how best to situate the pedagogical troika. But too often underlying notions concerning epistemology, the study of the origins and meanings of knowledge, are isolated as philosophical jargon some distance removed from the context in which they should be applied. With this awareness, the Commission has engaged the task of looking to epistemology in meaningful, practical ways to help define challenges of futuristic projections concerning education and its assessment. In varying degrees, the members of the Gordon Commission have become involved in discussions and in the generation of papers concerning the challenges posed for education and assessment in education by the fact that our conceptions of knowledge and what it means to know continue to change. In our considerations of the future of assessment in education, we anticipate that shifts in the epistemologies that inform human thought will continue to occur, and that we are likely to experience these shifts as occurring more rapidly and in greater conflicting interaction than has been true in all of human history. These shifting epistemological perspectives are especially important in the work of the Gordon Commission because of the phenomena of focus for the Commission: education and its assessment. Both are so firmly grounded in conceptions of human behavior, traditions in observation and measurement, and even the nature of reality that were the consensus positions of the 19th- and 20th-centuries.

Many of these notions are being challenged or at least reconceptualized in contemporary thought and educational practice. Educational and psychological measurement theory has not stood still in the face of these changes. However, testing practice has remained relatively stable. Scholars of the measurement sciences are deeply involved in exploration of and reflection upon emerging challenges and contradictions. Some have ventured the consideration of alternatives. However, charged with the tasks of inquiry into the possible character of education in mid-21st-century and the demands likely to be made on assessment in education, the members of the Gordon Commission have been forced to examine the relationships between the ways in which we think about human behavior and its assessment, and the ways in which we practice assessment. In this context, anticipated and observed changes in the ways in which we think about relevant issues have forced serious examination of the theoretical assumptions that underlie what we do in assessment as well as those assumptions, and ideas that will form the conceptual context for what we can expect of assessment in future years.

I wish it were possible to report light at the end of the tunnel in which we are digging. There is, perhaps, no aspect of the work of the Gordon Commission about which there is less ambiguity. Most of us who are engaged with the Gordon Commission were born into and cut our teeth on modernist empiricist scientific thought. Our minds are programmed to privilege positivist thought. But our examination of human history and the emerging epistemologies are convincing of the tentativeness of what we know, and the limitations of the conceptual products of positivism and modernity. Despite the enormous technological progress that modern ways of thinking have enabled, and despite the power of the logical reasoning that has been framed by these established ways of knowing, we are being forced to consider that extant models reflect particular ways of knowing, and that these ways of knowing are socially determined, essentially subjective and subject to error.

The largest single group of scholars on the Gordon Commission identify themselves professionally with educational assessment and measurement. Our field of specialization rests on the empirical sciences and positivist thought, borrowed from the “natural sciences,” and applied to the behavioral and social sciences. However, our examination of the scholarship leads us to conclude that what we believe and know may not always be a sufficiently accurate reflection of reality to be used as the sole basis for thinking about the future of assessment in education. Some of us believe that we may not be able to get where we need to go with education and its assessment using the extant knowledge base and the conceptual frames upon which that knowledge and the related techniques rest. The Gordon Commission has included in its work three syntheses of relevant knowledge and thought concerning the shifting epistemologies that inform our work. The positions advanced do not reflect a consensus position endorsed by the members of the Commission. The papers do reflect a perspective to which the Chairperson of the Gordon Commission is sympathetic. The fully developed papers can be found on the Commission's website. Readers are encouraged to seek out these provocative discussions of limitations and the potentials for assessment in education that are reflected in the epistemologies that inform us.

These changes in epistemologies and paradigms have been accompanied by changes along the margin of the field of educational measurement. Without challenging the core positions of the field, we see alternatives proposed and experimented with that are likely to be seen with some variations in future approaches to assessment. Some of these efforts appear in the Advanced Placement Studio Art program. The advent of the use of portfolios in education failed to gain traction, but continues to hold appeal for those of us who want to capture more dynamic pictures and follow processes of development in learners. It seems that I have always had problems with the broad application and interpretation of Skinner's Behaviorism, but I continue to find considerable appeal in his notion concerning the qualitative analysis of behavior, which he claimed as a necessary precursor to the design of contingency management.

I found persuasive, his notion that meaningful reinforcement needed to be grounded in a deep understanding of the anticipated respondent. I continue to believe that a qualitative analysis of behavior (in context and from the perspective of the person whose behavior is to be understood) is essential to planned intervention in the life of another. I saw this in the work of Haeussermann as she tried to understand the adaptive capacities and tendencies of children with serious neurological insults. More than forty years ago, I proposed to the first Commission on Testing sponsored by the College Board an approach to the qualitative analysis of standardized test data rendering it more useful for understanding and informing the teaching and learning processes. The essence of that argument follows.

The Qualitative Analysis of Behavior

Much of the impetus for the development of a technology of assessment related to intellective function and achievement resulted from and has been maintained by a supply-and-demand approach to access to education and distribution of educational opportunities. Access to a limited supply of educational opportunities has been guarded by selection procedures that prior to the 20th-century were based on the prospective student's social status. In the pre-Reformation period, access to education was limited to the political and religious nobility and later to other privileged classes, while the 20th- and 21st-century selection procedures have come to be dominated by the student's demonstrated or predicted intellectual status. Where the supply of opportunities has been limited, great emphasis has been placed on the selection of students and the prediction of their performance when exposed to those opportunities. Binet's work in intelligence-test development was directed toward the creation of an instrument that could be used to identify those pupils who were likely to benefit from schooling.

His admonitions that education also turns to treatment of those exposed as not likely to succeed were generally ignored. In a period of scarce educational opportunities, Binet's concern for the educability of intelligence did not gain favor. Society found greater utility in the promise of the predictive and selective validity of his new test. This emphasis on selection and prediction has continued even though the social conditions that gave rise to it have changed. In recent years, we have seen in America a growing concern with universal access to secondary and higher education. The educational requirements of the nation are increasingly defined as post-high school educational opportunities for almost all youth and continued learning for most people. If this trend continues, selection and prediction can no longer be allowed to dominate in the technology of psycho-educational appraisal. Rather, the stage must be shared with an emphasis on description and prescription—that is, the qualitative description of intellective function leading not to the selection of those most likely to succeed, but to the prescription of the learning experiences required to more adequately ensure that academic success is possible.

Psychological testing obviously can be used to measure achieved development. From those achievement patterns, subsequent achievement in the same dimensions of behavior under similar learning experience conditions can be predicted with reasonable validity. Thus, people who have learned an average amount during one learning period (high school) may be expected to learn an average amount in the next learning period (college). However, adequate attention has not been given to the facts that psychological testing can be used to describe and qualitatively analyze behavioral function to better understand the processes by which achievement is developed, to describe nonstandard achievements that may be equally functional in subsequent situations requiring adaptation, or to specify those conditions in the interaction between learner and learning experience that may be necessary to change the quality of future achievements.

In the present situation, confronting those concerned with access to higher education for larger numbers of young people and for youth from more diverse backgrounds than those from which college students previously were chosen, it is not enough to simply identify the high-risk students. The tasks of assessment and appraisal in this situation are to identify atypical patterns of talent and to describe patterns of function in terms that lead to the planning of appropriate learning experiences. Accordingly, it is recommended that we:

1. Explore possibilities for adding to its quantitative reports on the performance of students, reports descriptive of the patterns of achievement and function derived from the qualitative analysis of existing tests.
2. Explore the development of test items and procedures that lend themselves to descriptive and qualitative analyses of cognitive and affective adaptive functions, in addition to wider specific achievements.
3. Explore the development of report procedures that convey the qualitative richness of these new tests and procedures to students and institutions in ways that encourage individualized prescriptive educational planning.
4. Explore the development of research that will add to understanding of the ways in which more traditional patterns of instruction will need to be modified to make appropriate use of wider ranges and varieties of human talent and adaptation in continuing education.

Curriculum Embedded Assessments

Exploration with a view to the generation of prescriptive information for the guidance of the development of educational and pedagogical intervention has dominated my concern for alternatives to educational assessments. It gained its best expression in the exploration of the notion we called ***dynamic pedagogy*** in which the assessment and instructional functions were combined—integrated with the facilitation of learning. Briefly, dynamic pedagogy describes the process of teaching and learning in which assessment, teaching and learning are inseparable processes in pedagogy (Gordon & Armour-Thomas, 2006). By using the term "dynamic" I mean to refer to demonstrated learner strengths and needs. I use dynamic assessment to describe an approach to measurement that is as much concerned with uncovering the mental processes that examinees use in their performance as it is with the product of their performance.

Dynamic assessment seeks to determine the status of examinees or learners and the processes of learning by which the status is achieved or manifested. It is dynamic in the sense that it is adaptive to the performance of the examinee/learner and it has no fixed entry or exit points. In this way, assessment begins where the learner is, follows the learners' lead, and it ends at the limit of the learner's demonstrated ability or willingness to try. Instead of standardized procedures, the assessments are tailored to the characteristics of the person being examined. Thus, the primary task is not simply to understand what the student or learner knows and can do but to elucidate the processes and conditions that the learner uses to demonstrate her/his status. In the dynamic approach, we seek to determine the conditions under which the examinee can demonstrate what she knows or the processes by which he draws from the zone of proximal development or new learning to demonstrate both intent and consolidated competence. I feel that the dynamic assessment perspective maybe more useful for providing guidance to teaching and learning than is standardized assessment which has the potential for misrepresenting the constantly changing nature of learning as the processes by which attitudes, knowledge and skill are acquired and utilized.

Additionally, I believe that the increasing concern for equity and fairness in testing requires that responsible approaches to educational assessment include attention to the quality of available teaching and learning transactions and to the sufficiency of learner's access to these experiences as part of the assessment process. If one is assessing ability to learn it may not be sufficient to simply *assume* that one has had or availed appropriate and sufficient opportunity to learn. Colleagues such as Chatterji, Koh, Everson, and Solomon (2008) describe a useful assessment technique that is designed to help students deconstruct learning tasks in any content (e.g., mathematics) area. Similar to dynamic assessment, their concept of "proximal assessment," is also embedded in the instruction, and it is used as a diagnostic and instructional process during student-teacher interactions. It is continuous and useful in the planning and conduct instruction. In their research, Chatterji and her co-researchers trained teachers to use proximal assessment by categorizing math problem solving (e.g., division) into specific targeted student skills. For example, they state that by arranging problems in order of difficulty, teachers can evaluate student learning and understanding at various steps and more precisely reveal where the student misunderstanding begins. In this way, the purpose of their assessment is intended to diagnose problems during the instructional process instead of at the end of the learning period. Whimbey and Lochhead have used a similar method as they debrief learners during the course of instruction seeking learner explanations for the learning tasks that they are executing. In all of these assessment exercises measurement, diagnosis, prescription, teaching and learning are integrated in the interest of student progression.

For our final example, we turn to work at Educational Testing Service. ETS has been conducting a long-term research and development initiative called Cognitively Based Assessment of, for, and as Learning (CBAL™). They are engaged in this complex initiative because of a belief that existing approaches to K–12 accountability assessments could be markedly improved by incorporating:

- Findings from learning-sciences research about what it means to be proficient in a domain (in addition to common core standards);
- Tasks that model effective teaching and learning practice;
- Mechanisms for returning information about student performance in a rapid-enough fashion to be of use to teachers and students; and
- Testing on multiple occasions so that highly consequential decisions have a stronger evidential basis.

In the CBAL Initiative, ETS's central goal is to create a future comprehensive system of assessment that:

- Documents what students have achieved ("of learning");
- Helps identify how to plan and adjust instruction ("for learning"); and
- Considered by students and teachers to be a worthwhile educational experience in and of itself ("as learning").

The system attempts to unify and create synergy among accountability testing, formative assessment, instruction and professional support. Envisioned is a system having the following key characteristics:

- Accountability tests, formative assessment and professional support will be derived from the same conceptual base. That base will be built upon cognitive-scientific research, Common Core or state standards and curricular considerations.
- The CBAL assessments will consist largely of engaging, extended, constructed-response tasks that are delivered primarily by computer and, to the extent feasible, automatically scored.
- Because of their nature, the CBAL tasks should be viewed by teachers and students as worthwhile learning experiences in-and-of themselves. Ideally, taking the test should be an educational experience, and preparing for it should have the effect of improving student domain competency, not just improving performance on the test.
- Accountability tests will be distributed over several administrations throughout the school year so that: (1) the importance of any one assessment and

occasion is diminished; (2) tasks can be more complex and more integrative because more time is available for assessment in the aggregate; and (3) the assessments provide prompt interim information to teachers while there is time to take instructional action.

- For accountability purposes, estimates of student competency will be aggregations of information collected over time. In addition to these competency estimates, the accountability tests will offer some formative information to teachers, students, parents and local policymakers.
- Results from the accountability tests will serve as initial information for more extensive formative or diagnostic assessment, indicating such things as the competency level, area(s) in which follow-up is suggested, and initial formative hypotheses. (However, the CBAL formative assessments will never be used for accountability purposes.)

The CBAL formative assessment will be designed to help teachers engage students in a structured process that reveals evidence about what students know and are able to do, helps teachers and students identify the characteristics of proficient performance, and moves students toward developing competency. The CBAL formative assessment will include classroom tasks and activities, resource materials and diagnostic tests. Most components of the CBAL formative assessments should be adaptable by the teacher for use when and how the teacher sees fit. The CBAL assessments should be designed to help students take an active role in their own learning and the evaluation of it. While the accountability testing, formative assessment and professional support components derived from the same conceptual base, they should be able to function independently of one another. So, for example, states should be able to implement the accountability system without also having to implement the formative system, should they so desire.

Personalized and Public Space[7]

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 any one 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 (Nov, 2012). As important, everything was changeable and under the control, read authority, of the user. And the users were urged to express their opinions.

[7] This passage is excerpted from the very insightful essay by Eva Baker, "Testing in a Global Future." The entire essay can be seen at: http://www.gordoncommission.org/rsc/pdfs/baker_testing_global_future.pdf

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.

Emerging Developments from the Sciences and Technology

As I contemplate my own aging and observe my children and grandchildren doing for me what I once did for them, and as I realize that if I live long enough my body and mind will have returned to the infant-like organism similar to its beginning, the existence of a self-regenerating mollusk is more fascinating than surprising. It seems that in nature, all things come and they go, and in the process they become transformed and/or transform themselves. As humans become more knowledgeable and more skilled at the manipulation of knowledge and the creation technologies by which knowledge is expressed and human abilities are extended, we become aware of the infinite possibilities of humans working together to exploit a universe that seems to have limitless possibilities. As the Gordon Commission on the Future of Assessment in Education, we did not get very far in the exploration of these domains, but a future commission will have to do so. The possibilities of nature, science, technology and scientific imagination await human exploitation. Assessment, teaching and learning will need to be influenced by the products of such exploitation. I am thinking of such possible developments as those associated with changes that we can comfortably anticipate such as:

1. Changes in the geographic centers of control of the political economies of the world, and changes in the relations of humans to production, resulting in sharper divisions between the peoples of the world who have and those who have not.
2. Increased dissonance born of diversity and variation in human characteristics, plural cultural identities and the existential proximity among the peoples of the world.
3. Shifting epistemologies, categories and conceptions of knowledge, and conceptions of what it means to know and understand.
4. Heightened relevance of and sensitivity to contextualist and perspectivist relativity in what is accepted as knowledge.
5. The ubiquitous presence of digital information mediated by electrochemical, electronic and mechanical amplifiers of human abilities.
6. Abundance of information and chaos concerning its meanings and processing, opening vast opportunities for data mining and relational data analytic management, coupled with the demand for the recasting of traditional taxonomies.
7. Changes in the paradigms that inform educational policy, practice and their assessment.

We may be forced to think differently about human capability as we move through the 21st-century. In one of the first of the consultative conversations convened by the Gordon Commission, Professor James Greeno reminded me that we may have an intractable problem in trying to speculate about the future, since the first thing that we know about the future is that it is in so many ways unpredictable. “How do we educate for unpredictability?” We concluded that learning how to accommodate ambiguity, change, and fluidity might help. Someone spoke of the capacity to gain and maintain focus while being able to entertain multiple perspectives. Another introduced the notion of adaptability. I left the conversation feeling the need to rethink my conception of intellective competence. Even Greeno’s and my efforts at combining his notion concerning intellective character with my notion concerning intellective competence seem inadequate preparation for unpredictability.

In thinking about education and its assessment for the world of the 21st-century, we are confronted with the need to go beyond notions of intelligence, knowledge and even abilities that are constrained by conceptions of fixity, stability, predictability, veridicality, some would even include truth. Most of our effort at understanding human adaptive capacity has privileged some conception of intelligence and most of those conceptions have included the mastery and manipulation of knowledge and technique. And, our efforts at assessment and measurement have focused on documenting the amount and status of what we know and know how to do. Lately, I have been thinking that more important than the focus on amount and status, understanding human adaptive capability may require a focus on process. Especially, as the complexities of the earth become more obvious – as knowledge and technique are recognized to be influenced more by context and situation, and as human abilities are amplified and extended by technology, knowledge and technique increase in liability.

In such a world, consider for a moment the utility of such processes as: Agency, Adaptation, Agility, Comprehension, Communication and Collaboration. Changing the course on one’s development may depend much more heavily of the teaching and learning person’s understanding and command of such processes as these, than solely on the amount and relative status of one’s knowledge and skills. Debates concerning these issues led the Gordon Commission to inquire into the question of what will it mean to be an educated person in the middle of the 21st-century.

4. TO BE AN EDUCATED PERSON IN THE 21ST CENTURY

Carl Bereiter and Marlene Scardamalia

What should distinguish an educated person of mid-21st century from the educated person of a century earlier? Unfortunately, the most straightforward answer consists of a number of added specifications with very little compensating elimination of older ones. New technology is downgrading certain technical skills such as penmanship, ability to do long division, and ability to thread a movie projector; but the academic content and competencies set out in the 1959 *Case for Basic Education* (Koerner, 1959) remain as important now as then, along with challenging new content and additional competencies that now demand attention. And some of the 1959 wisdom rings more tellingly now than it did back then, particularly Clifton Fadiman's words about "generative" subjects that enable future learning and about the value of education in saving students from feeling lost, in enabling them to feel "at home in the world" (Fadiman, 1959, p. 11). Rather than approaching the question with an additive mindset, however, we attempt in this paper to approach it in a way that is open to possibilities of transformation in educational ends and means.

The coming decades are likely to see the individual learner having to share space with the group as the unit of analysis in teaching and assessment. There are legitimate senses in which learning not only take places in groups but is a group phenomenon (Stahl, 2006): Group learning is something beyond the learning undergone by members of the group; it is something only definable and measurable at the group level. There are legitimate and important senses in which groups understand (or fail to understand), develop expertise, act, solve problems, and demonstrate creativity (Sawyer, 2003). While the title of this chapter indicates a focus on the individual, much of what we have to say is shaped by the larger question, "What will it mean to be an educated society in mid-21st century?"

A Different Kind of Person?

In speculating on what it will mean to be an educated person in the middle of the 21st century, the first question to consider is whether mid-21st-century people will be different kinds of persons from their 20th-century counterparts. There is much talk about brains being "rewired" by game playing and cell phone use. Without venturing into such speculation, we can note potentially far-reaching behavioral changes resulting from new kinds of social communication. There is the social website phenomenon of “friending,” which leads to vastly expanded circles of putative friends compared to the usual networks of direct contacts. Whether these constitute friendship in the normal sense may be questioned, but what is most evident is the extent to which communication in these social media is person-centered in contrast to being idea-centered. This shift is something of potentially major educational and perhaps cultural consequence, and we return to it briefly at the end of this chapter, in a section titled “Will Technology Facilitate Becoming an Educated Person?” Related to it, and also of potential profound consequence, is the trend toward short messages without the continuity of ordinary conversation. Short, mostly discontinuous messages also characterize text messaging and the commenting that pervades blogs and web news sites. As technology evolves enabling speech to play a larger role in online communication, the trend toward brevity may be reversed, but it could mean even farther distancing from the “essayist technique” that has been the medium of extended reflective thought (Olson, 1977). Extreme personalization and fragmentary communication would appear to be antithetical to what quality education has traditionally stood for. Are they really? And if they are, how should education respond to them?

The consequences of a shift toward greater person-centeredness are indefinite enough at this time that they may look favorable to some and dismal to others. A standing joke these days is Facebook denizens reporting what they (or sometimes their dog) are having for dinner. It does appear that much of the content appearing on social sites and personal blogs can only matter to people who have a personal interest in the author. A similar trend may be detected in contemporary poetry; whereas at one time you needed a classical education to understand the allusions in a poem, now you often need to know the poet. What is being lost here is the drive toward expansive meaning that characterizes the arts and scholarly disciplines. It may be that this is a good thing; it is consistent with postmodern skepticism about grand narratives. But it certainly gives a different meaning to “well-educated” from what it had a century ago.

The trend toward shorter, more fragmentary communication has more direct implications for ability to meet the intellectual challenges of this century. Can the increasingly complex problems of 21st-century societies be solved by sound bites? The answer is surely “no,” yet utterances of 15 seconds or less are already taking over political discourse, while, maddeningly, legislation keeps getting longer. Although we have not seen any systematic evidence on the matter, numerous Internet bloggers remark on the paradox of books and other media getting longer, while ability to sustain attention over long stretches is getting shorter. Quite possibly, these are not divergent trends but different manifestations of the same trend, which is a declining ability to do the sustained integrative thinking that can on one hand tighten prose around essential ideas, and on the other hand enable readers to process complex texts. The proof will come if speech-to-text becomes the preferred medium of asynchronous communication: Will it result in more extended thought, or will it tend to bury thought under transcribed babble?

Text is gradually being replaced by hypertext—coherent texts that contain abundant clickable links to related information sources. The virtues of hypertext are obvious to anyone researching a topic on the Internet, but it does pose a heightened challenge to focus. Following a link to a source that contains additional links, following one of those, and so on can quickly lead to loss of one's original purpose. Improved media design may make it easier to recover one's line of thought, but ultimately the challenge is an educational one: to heighten metacognitive awareness, to help students keep cognitive purposes in mind and to evaluate their current mental states against them. This is but one example of what promises to become a growing educational challenge: to promote *sustained work with ideas*. Society needs it, new media provide both tools and diversions from it, and schools have scarcely begun to recognize the challenge. Sustained work with ideas also poses a challenge for educational technology design, but one that has not yet come into clear focus for developers. Hopefully this will change. We are currently working on design of a new digital knowledge-building environment that has a person-oriented space for social interaction around ideas, but in addition an "idea level" where ideas abstracted from the social space become objects of inquiry, development, and improvement—where what goes on may be described as ideas interacting with ideas rather than people interacting with people.

Education's Two Faces

Being an educated person has traditionally had two aspects, one representing academic knowledge and skills and the other representing personal qualities—traits of character or intellect that the educational process is supposed to develop. Recent future-oriented literature has shown a definite tilt toward the second aspect, now described by such terms as "higher-order," "21st-century," or "soft" skills, "habits of mind," and "literacies." Reasons for the tilt toward personal qualities are not difficult to discern. There is the rapid growth of knowledge, which makes mastery of any subject increasingly beyond reach and renders knowledge increasingly vulnerable to obsolescence. There is the ready availability of factual information via web search engines, which reduces the need to store declarative knowledge in memory. And then there is the general uncertainty about what the future will demand of people, thus raising doubt about the value of specific knowledge and "hard" skills and favoring more broadly defined educational objectives such as "learning to learn," "critical thinking," "communication skills," and "creativity." These, it can be assumed, will always be useful. In practical educational terms, however, this is also a tilt away from things that teachers know how to teach with some degree of effectiveness to objectives of questionable teachability.

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. Real life requires that people not only have knowledge, but that they be willing and able to act upon it. This has multiple implications for the kinds of life experiences that constitute growing up into active citizenship, although it is not evident that times are changing in this respect. Many educators will argue that there is increasing need for students to eschew violence, honor diversity, and free their thinking of racism, sexism, homophobia, ethnocentrism and other prejudices. They will therefore want to include these in any description of an educated person. It must be recognized, however, that throughout history there have been well-educated people who demonstrated none of these qualities and were sometimes notable for their opposites. The standard rejoinder is that such people could not have been well educated; but we do not believe it is wise to burden the term "educated" with every desirable human quality. Better to acknowledge that there is more to being a good person than being well educated. One can go to virtually any poor village and find uneducated people who are paragons. Eliminating moral perfection from the definition of an educated person does not, however, mean eliminating emotions, beliefs, mindsets, and moral reasoning from consideration. On the contrary, it frees us to consider in a constructive way the role that these may play in cognitive processes, along with knowledge, skills, and aptitudes. A lot more is known about this interplay today than was known back when "higher-order skills" first came on the stage, and in the following discussion we attempt to draw on this recently developed knowledge.

Knowledge and Knowledgeability

The status of knowledge in what is frequently called the "Knowledge Age" is ambiguous. Everyone is of course in favor of knowledge but *knowledgeability*, the retention of knowledge in individual nervous systems, has come under scrutiny, for reasons already stated. A legitimate subquestion to *"What will it mean to be an educated person in mid-21st-century?"* is the question, *"What will it mean to be a knowledgeable person in mid-21st-century?"* An answer to this question divides into three parts, each of which poses both assessment and instructional problems.

21st-century subject matter. Over the course of educational history, new subjects have from time to time been adopted as essential, and more rarely a subject may be dropped. Science made its way into the curriculum against some resistance, and now is raised close to the top. The late 20th century saw ecology and cultural studies entering the list. Computer programming came and went as an element of general education—and may be on its way back again (cf. Resnick, et al., 2009). Probability, statistics, and graphical representation of data, which were largely absent in mid-20th-century schooling, are now essential for following the daily news. Not yet fully arrived in the curriculum are complex systems theory and mathematical modeling, although these are arguably essential for advanced work in virtually any discipline.

Identifying important new subject matter has been something curriculum planners have been doing energetically ever since the Sputnik era. Identifying what it will take for adequate knowledgeability in the present century calls, however, for more complex analysis. It is not enough to identify topics that are worthy of instruction. We need to identify where schooled knowledge is falling short of emerging needs. For instance, "financial literacy" is a need brought into the spotlight by current economic problems. However, proposals currently on the table are focused on personal finance. Important as this may be, people can be knowledgeable about their personal finances—knowing how to recognize and avoid high-interest traps, for instance—and still be financially illiterate when it comes to national economic policy. In fact, using one's personal financial wisdom as a paradigm for judging governmental policies is a serious and all too common mistake; it leads to a simple-minded "thrift" approach that ignores macroeconomic effects on currency, inflation, employment, and level of consumer spending. Economics, like practically everything else of societal importance, needs to be understood systemically—and that is what most strikingly distinguishes 21st-century knowledgeability from what could serve adequately in times past.

Depth of understanding. Teaching for understanding is widely advocated. Knowledge tests are being reshaped to test for it, with less emphasis on testing factual recall. But when it comes to assessing *depth* of understanding, educational assessment does not seem to have progressed significantly beyond Bloom's *Taxonomy* (1956). The *Taxonomy* cast the problem in behavioral terms: "Specifically, what does a student do who 'really understands' which he does not do when he does not understand?" (p.1) Accordingly, depth was operationalized, by a hierarchy of increasingly sophisticated things that students might do with their knowledge: applying, analyzing, synthesizing, and evaluating. This approach was further developed in a revision of the *Taxonomy* (Anderson

& Krathwhol, 2001) and in Perkins' "understanding performances" (Perkins, 1995; Perkins & Unger, 1999). While it is no doubt true that being able to do increasingly difficult things with knowledge requires increasing depth of understanding, this does not really get at what *depth* means, and the assessment tasks suffer from the fact that a student is liable to fail them for reasons other than a lack of understanding (Bereiter & Scardamalia, 1998).

There is another way of operationalizing depth: define it according to *what* is understood rather than *how well* the student can demonstrate understanding. Student understanding of evolution can be mapped in this way. At the lowest level, students understand that biological adaptation occurs but they treat it as something that just happens. Ohlsson (1991) found this to be a prevalent conception among university undergraduates. At a significantly deeper level, students have the idea of "survival of the fittest" and can explain the giraffe's long neck on the basis of longer-necked giraffes surviving while those with shorter necks died without reproducing. This is about as far as understanding evolution usually goes in school biology, but as advocates of Intelligent Design point out, it fails to explain the emergence of new species or the evolution of complex organs such as the eye. Explaining those things requires understanding several deeper concepts, and still deeper and more complex ones are required to explain other phenomena such as irregularities in the time course of evolution. All these understandings are testable and they form at least a partially ordered scale of depth of knowledge. Developing similar scales in other domains may require the kind of research that has been devoted to students' evolutionary concepts, but it is worth doing not only as a basis for testing but also as a basis for finer-grained learning objectives.

Defining progressions in depth of understanding is especially challenging for newer subject matter where there is not a long history of efforts to identify and teach essential concepts. Probability and statistics are being taught, but are they being taught in sufficient depth that they become useful tools for gaining insight into real-world problems? Huck (2009) has identified 52 misconceptions that indicate failures in the teaching of statistics and probability, but are the conceptual errors as miscellaneous as they appear, or are there deeper ideas of which the 52 misconceptions are a reflection—failure, for instance, to grasp and apply the idea of the set of equally likely events, which is foundational to most school-level work with probability? People's erroneous thinking about probability in real-world phenomena, however, seems to depend not so much on faulty knowledge of statistics and probability as on simplifying heuristics and biases that preempt formal knowledge (Kahneman, 2003). Another domain that cries out for a mapping of concepts according to depth is systems theory. First graders are being introduced to the concept of *system* and are being schooled in a reasonable definition of it. But where does instruction go from there? How many students, or teachers for that matter, can distinguish a systemic explanation from a mere multivariate explanation? Where does understanding of ecosystems fall short or go wrong?

Quantity of knowledge. Despite its being frequently disparaged in the education literature, sheer quantity of knowledge still matters. It matters because it increases the likelihood of making fruitful connections and analogies; it increases one's resources for performing successful web searches; and it provides entry to informative texts that can convey deeper information (Hirsch, 1987). We have heard informally about an experimental test tried out by the College Entrance Examination Board sometime before 1970. It tested miscellaneous world knowledge of the kind represented in mass-media news magazines, and found that scores on such a test were as good or better at predicting college grades than the familiar aptitude tests. The experiment was abandoned, reportedly, because it conveyed such a negative impression of the nature of college education. In reality, of course, schools and colleges do not teach or test miscellaneous knowledge. It is picked up informally by living an active intellectual life in an information-rich environment. The 21st-century ought to be better for this than preceding centuries. Having a rich store of miscellaneous knowledge is accordingly one reasonable marker of successful mid-21st-century educational growth—not the most important, but one deserving of support. Assessing quantity of miscellaneous knowledge presents a problem that occurs whenever students are not expected all to learn the same things. In a later section of this paper, we will consider this problem, which takes on increasing importance as education moves increasingly toward individualization, self-directed learning, and knowledge building.

Education for Change?

The phrase "education for change" has begun to appear around the Web; several charter schools have it as their name. It is not clear what "education for change" means, and the websites we have visited do not offer explanations. One obvious implication of the accelerating rate of change is that students should be prepared to undertake substantial learning efforts later in life. Project-based learning is supposed to provide such preparation, but does it? Evidence, of course, is lacking—because of the time gap. The commonest projects consist of gathering and presenting information on some topic (Moursund, 1999). We find education graduate students who are so practiced at this that they resist any major course assignment that calls for anything different. Yet this kind of project does not bear very directly on lifelong learning needs. Seldom does practical life call for researching a topic. In adult life, studying topics of interest is avocational, pursued for its intrinsic rewards: a worthy activity, but not an especially timely one. Instead, life's exigencies call for gaining new competencies and for obtaining information to solve problems. Some school projects involve the latter, although often the problems are given and the procedures for gathering the needed information pre-arranged. Even explicitly "problem-based" learning generally offers little experience in dealing with ill-defined, emergent problems, where it is uncertain what information will prove helpful; yet those are the kinds of problems that "education for change" ought to be preparing students for.

Acquiring new competencies is something not only "knowledge workers," but people in all lines of work may anticipate. Present-day schooling provides practically no experience in doing this independently; to the extent that students gain such experience they gain it outside the curriculum—in sports, hobbies, and community or entrepreneurial work. Schooling could do more, however, and address kinds of competence that draw more directly on disciplinary knowledge. In mathematics, for instance, there are countless special applications in which students could gain some proficiency—far more than can be taught in the regular curriculum. There is mathematics of heating and air conditioning, structural engineering, the tuning of racing cars, navigation, grilling and smoking meat, finance, cosmetology, musical composition, and on and on. Working independently or in self-selected teams, students could undertake to gain proficiency in some mathematical application of interest to them and be called on to present evidence of proficiency. Replacing some topic-centered projects with competency-centered projects could probably enliven school experience considerably and be of more direct lifetime value. "Competency-based education" has been a recognized movement since the 1960s, but it emphasizes pre-specified objectives systematically taught and evaluated (Burke, et al., 1975). In competency-centered projects, by contrast, students would be responsible for all of this and the emphasis would be on gaining experience in self-directed acquisition of new competencies. The rapidly growing number of websites that award "badges"—including "expertise badges"—seems to reflect a similar interest in self-directed acquisition of competencies, although at present the earning of badges appears to be an individual rather than a collaborative effort, and in most cases the badges are awarded on the basis of activity (visiting websites, commenting, and so on) rather than on the basis of demonstrated competence.

Arguably, the young have more to teach us than we have to teach them about adapting to the famous "accelerating pace of change." Of far greater concern in the big picture is not the ability of individuals, but the ability of institutions to adapt to changing conditions. Companies that fail to adapt or to adapt as rapidly and effectively as their competitors are a recurrent feature of the business news. Of course, institutions are composed of individuals, but adaptability at the individual level does not ensure adaptability at higher systemic levels. A striking case in point is the infamous “mile-wide, inch-deep” curriculum. Everyone is opposed to it, yet it persists despite vigorous reform efforts. But perhaps the most dramatic example of failure to adapt to changing conditions is the U.S. Congress, which at this writing appears to be almost completely dysfunctional, unable to cope with the most pressing problems of unemployment, economic stagnation, widening income inequality, access to medical care, climate change, deteriorating infrastructure, civil liberties, and more, while mired in dogma and mindsets of the past. If education can make a difference, it will not be through fussing about change per se but through equipping people with what they need to form rational judgments on issues such as those just listed. This, of course, is uncontroversial. Controversy arises—or should arise!—when it comes to defining what people need in order to form and act upon such rational judgments. Is it training in critical thinking or is it better understanding of economics, ecology, and so forth—or can education manage to provide both? And, realistically, how far can schools go in opening closed minds when there are such strong social forces opposing it?

Cosmopolitism

To be an educated person in today’s world is to be a cosmopolitan (Rizvi, 2008)—someone who is a citizen of the world, at home in and able to navigate among its variety of cultures, ideas, and life styles; someone who may cherish his or her own background traditions and world view but is not bound by them. We use the term “cosmopolitism” instead of the much more common “cosmopolitanism” to refer to the state of being cosmopolitan, because “cosmopolitanism” has acquired too much ideological baggage. The current *Wikipedia* article of that name in fact defines “cosmopolitanism” as an ideology, a one-world ideology. Although this ideology may appeal to liberal thinkers, a goodly portion of the American populace is liable to reject it as an attack on American exceptionalism.

In current curricula the goal of broadening students’ outlook on the world is subsumed by “multicultural education” (Banks, 1994), knowledge goals such as “global awareness” (Partnership for 21st-Century Skills, n.d.), foreign-language learning, and various sorts of cross-cultural student activities. But, as with other human development objectives, breaking cosmopolitism down into knowledge, skills, attitudes, and activities seems to miss the essence, which we would define as “feeling at home in the world.” This is a politically sensitive topic. Unless you are very careful about how you express yourself, cosmopolitism may come out sounding like something you might read in a luxury-travel magazine or, worse yet, like cultural imperialism. Yet there can be little doubt that the kind of worldliness traditionally promoted by the “junior-year abroad” needs a much fuller and richer development in the coming decades.

Media Literacy

There is no need to detail the remarkable proliferation of ways to express oneself and to represent knowledge and ideas. These are bound to become increasingly diverse and powerful, not only in terms of what people are exposed to, but in terms of what they can exploit as means of communication. If information media continue to develop at the rate they did in the past 40 years, there is no telling what the situation will be like by mid-century. This suggests that media literacy, if there is anything that warrants that name, needs to be grounded in fundamentals and to rise above trendiness.

There was a time when information technology literacy meant trotting students off to a computer laboratory where they did exercises in word processing, computer graphics, and Internet searching. Although vestiges of that practice can still be found, information technology is now too various and multi-purpose for that to be helpful or even feasible. It now makes more sense simply to carry out educational activities in which various kinds of technology play natural and useful roles. Equally obsolete, however, are the old media literacy activities that involved recognizing propaganda and persuasion techniques, using artificial or relatively trivial examples while avoiding controversial examples of major social significance.

Media literacy in this century is going to have to take a higher path. Students need to recognize the media as (a) causing social change, (b) being shaped by social change, and (c) evolving painlessly through the interaction of many factors. The consolidation of media in conglomerates is a rapidly developing fact. Former CBS news journalist Dan Rather has voiced his alarm at “the corporatization of the news, the politicization of the news and the trivialization of the news” (Rather & Diehl, 2012, p. 289). Rather traces the descent from the great independent newspapers that featured investigative journalism and worldwide news coverage to television networks that treated news as a public service, finally to arrive at the state where “we now have four talking heads in a studio shouting at one another, instead of four overseas bureaus covering real news” (p. 289).

A recent survey by Fairleigh Dickinson's PublicMind is ominous with regard to the present state of news media literacy. It indicates that people who watch Fox News know less about world events than people who do not frequent any news sources at all (Cassino & Wooley, 2011). How could this be? It can hardly be supposed that Fox News failed to report the overthrow of President Mubarak in Egypt, for instance. One explanation is that people who frequent this channel, with its heavy emphasis on right-wing domestic politics, are sufficiently imbued with American exceptionalism that they tend to disregard events in other countries. (They may be the same people who favored legislation instructing the courts not to pay attention to foreign court decisions.) What seems to be happening (and new technology is unlikely to alter the trend) is that news reporting and commentary get distributed over countless sites. More information is becoming available than in times past, along with more varied interpretations of events, but aggressive, sustained inquiry into what is behind the news (investigative journalism on socially significant issues) is being replaced by professional and amateur punditry (the "four talking heads shouting at one another").

Since it is out of the question for people to become their own investigative journalists, the best education can do is help people become their own pundits. This means thinking critically and reflectively about information received. However, a more proactive stance than this is needed in order to deal with information overload. It means the educated citizen functioning as a theory builder rather than merely an opinionator. It means applying the hypothetico-deductive method (also known as abduction): producing a conjecture that explains the facts and then searching for information to test whether the conjecture is true. The conjecture in the preceding paragraph about why Fox News viewers would know little about world events is hypothetico-deductive. It explains the reported fact, but is it true that Fox News viewers are exceptionally imbued with belief in American exceptionalism? We have found evidence that Fox News vehemently upholds American exceptionalism and that conservatives are significantly more likely to endorse it than moderates or liberals, which is in accord with the conjecture, but we have not yet found direct evidence that Fox News viewers tend to be especially committed to exceptionalism and especially disdainful of foreign influences. So the conjecture stands, although it could be replaced by a more convincing explanation or defeated by counter-evidence. We use this example to suggest that an educational response to what might be the major "new media" issue of coming decades calls for something much more substantive than what generally passes for "media literacy." It is not just teaching critical thinking or media skills or familiarizing students with new media; it is engaging students in real theory development about real issues, using new media as a resource.

Moral Reasoning

The educated person of the mid-20th century will need to be a capable and dedicated moral reasoner. But what is new about this? Moral education by one name or other has long held a respected place in education, and pedagogy for moral reasoning was well worked out by Lawrence Kohlberg and others in the 1960s. What is new is the globalization of moral issues and the increased complexity that goes with it. It is becoming increasingly difficult to ignore injustices in distant places or in cultures different from one's own, but reasoning about them becomes complicated by issues of cultural hegemony and differences in worldviews. Three moral issues prominent in the news at this writing are honor killings, female circumcision, and abortion. Honor killings and female circumcision are established folkways in some societies. As long as they were confined to distant and poor communities, Westerners were free to disapprove but do nothing about them. With globalization, however, what was once remote has become closer and sometimes internal to Western societies, and there is a growing universalism that on one hand respects cultural diversity and eschews cultural imperialism and on the other hand espouses universal human rights and tends to bind women the world over in insistence upon the same rights. So, people of a modern liberal disposition find themselves caught between one stricture that says don't intervene in the cultural practices of other societies, and another that says gross violations of human rights cannot be tolerated anywhere. At this time it may be feminist thinkers who agonize most over this dilemma, but it ought to be of concern to everyone. The moral dilemma goes far beyond such Kohlbergian moral dilemmas as whether Heinz should steal to get a life-saving drug for his wife.

A moral dilemma that engages a much larger swath of citizenry in Western nations is abortion. Unlike the previous two issues, abortion has organized groups and even political parties lined up on opposing sides. Appeal to emotions is a standard technique, but its role in abortion controversy dramatizes the new level of challenge it poses to moral reasoning. A few decades ago the emotional appeals available to opposing sides in abortion debates may have been about equal in persuasive power. But now anti-abortion websites can produce vivid displays of the most gruesome, bloody sights, set in contrast to the charming smiles of babies. Some states are now carrying the matter even further by passing laws that require women seeking abortions to view ultrasound video of the fetus they are carrying. Nothing that pro-choice groups can present comes anywhere near this in emotional arm-twisting. This example suggests that moral reasoning faces a more uphill struggle than it did in the past, not because of some epidemic of irrationality (although that may also be occurring) but because modern communications have greatly increased the persuasive power of visceral appeals to emotion.

Rational Thought and Emotionality

Two points about the abortion example have implications that extend beyond moral reasoning to reasoning in general. One is that new media are elevating means of expression that are alternative to words, even though words (and in particular written words) remain the principal medium of rational thought and discourse. This is probably an irreversible trend and not necessarily detrimental to rational thought. One need only consider how graphical representation of data, now pervasive in news media, has put quantitative information within the reach of people who could not have grasped it in verbal or numerical form. But if this shift is accompanied by a reduced ability to follow an extended exposition in text, we have a serious educational problem. It is bad enough that political messages are put out as sound bites, but if people's thinking about important issues is also carried out in parcels of thought equivalent to sound bites, we have a cultural problem that 21st-century education must do something about.

The other point draws upon an important advance in understanding reasoning, known among cognitive and brain scientists as "dual process theory." Dual process theory posits separate systems of response, both of which may be activated by an event or a message but that differ not only in how they work but in the speed at which they work. System 1 (Kahneman, 2011; Stanovich & West, 2000) is a system of rapid response, which works on the basis of associations (similarities, co-occurrences, etc.) and may trigger emotions of any sort, such as fear, disgust, or anger, and may precipitate immediate actions such as flight or attack. System 2 is a slower-acting system that carries out a sequential thought process. The important practical implication is that by the time System 2 kicks in, System 1 has already acted and left the thinker with an immediate judgment, impression, or action response. There is reason to believe that System 1 typically dominates moral judgment; we respond with immediate gut reactions of disapproval or approval, revulsion or admiration, and System 2 serves, when it can, to provide us with justifications for those reactions (Haidt, 2001). This seems transparent in people's opposition to same-sex marriage and to gay rights in general. Even if Biblical justifications are invoked, they have a distinctly ad hoc flavor because of the many Biblical injunctions ignored (as currently satirized by the website godhatesshrimp.com) and because of the absurdity of the worldly arguments they bring forth to buttress their case (threat to marriage, and so on). This example illustrates that the dominance of System 1 over System 2 is not just a matter affecting personal morality, but is something that can have far-reaching social consequences when the issue is one on which gut reactions are strong and widely held. New media are providing both means to provoke massive System 1 reactions and ready-made System 2 justifications for them. Of course, they can also provide stimuli for countervailing System 1 reactions as well as information and food for impartial System 2 thought. But arguably the balance of power is shifting toward System 1 arousal rather than System 2 rationality.

Affect-laden responses, triggered by associations and correlations, are as much a part of being human as deliberative thought. System 1 is educable, as in the educating of tastes and empathy. While it can be a source of stereotyped or habit-bound reactions it can also produce novel and unpredictable cognitive turns (Thagard, 2006). Being able to recognize the action of both systems in our judgments and to evaluate and possibly revise System 1 judgments with System 2 processes is a new take on what it means to be rational (Stanovich, 2004).

Thinking and Learning Skills

"Teach them to think" is an educational objective that can be traced as far back as Socrates. However, the idea of treating thinking as a skill (or set of skills) seems to have been a mid-20th-century innovation, and a questionable one. Before that "teach them to think" was treated more as a kind of character development and all-around intellectual development. "Teach them to think" might have been glossed as "teach them to be thinkers." For good or ill, meeting the 21st-century's need for good thinkers is being treated by education systems around the world as a skill-learning problem rather than a human development problem.

Several groups are developing thinking skill tests with the express purpose of driving schools to teach thinking skills. Because this will predictably and perhaps intentionally lead to teaching for the test, serious attention ought to be given to issues of teachability, transfer, and fairness. There has been little such attention. A point may be put on this skepticism by examining what has been claimed as "one of educational psychology's greatest successes" (Mayer & Wittrock, 2006, p. 298), the teaching of problem-solving strategies. What is being referred to more specifically, however, is the teaching of strategies for solving mathematical word problems. The evidence is quite clear that teaching such strategies as drawing diagrams, working backward, making a plan, and paraphrasing instructions leads to improved performance—on mathematical word problems. So teachability has been demonstrated; but what is being taught? There is no evidence that students actually use the taught strategies, and so what is being taught may only be a habit of mindfulness—a habit of thinking a bit before plunging ahead with numerical operations. As for transfer, the evidence is that the effects are rather strictly limited to word problems of the kinds used in instruction. But might this not be a valuable skill in its own right? Hardly. Where in real life does one encounter fully stated problems with all the necessary information for solution provided? On the puzzle pages of newspapers, but scarcely anywhere else. What is being taught is, thus, essentially a limited puzzle-solving skill. Evidence has been accumulating ever since Luria and Vygotsky's original research, showing that schooled people are better disposed to deal with hypotheticals than less-schooled people, who are more inclined to base conclusions on real-world knowledge (Scribner, 1979). Mathematical and logical word problems are hypotheticals par excellence and solving them requires adhering to the explicit (often unrealistic) terms of the problem and not allowing common sense and world knowledge to intrude. Testing a skill that requires suppressing common sense and world knowledge thus raises questions of fairness as well as questions of relevance.

Questions may be raised about other purported thinking skills as well. How does the kind of creativity that can be demonstrated on a test relate to the kind of sustained and cumulative creative work that is valued in the real world? If one fully understands the opposing sides in a controversy, is there any need for something additional that constitutes "critical thinking" skill? More generally, what is there in the various purported thinking skills that cannot better be treated as "habits of mind" (Costa & Kallick, 2000)? Are thinking skills in the aggregate the same as fluid intelligence? Factor analytic evidence indicates that they are (Kline, 1998). If so, testing thinking skills as educational objectives means using intelligence tests as achievement tests, something bound to cause an uproar leading to eventual abandonment of the tests. There is not even adequate empirical and theoretical basis for calling cognitive traits such as creativity, critical thinking, and problem-solving ability skills at all. The fact that people demonstrably differ in them is not sufficient proof. Learnability, teachability, and wide-ranging transfer have to be demonstrated, and evidence to support such claims turns out to be little more than evidence that teaching for the test improves test scores (Detterman, 1993). Although in the present climate it is heretical to suggest it, schools might be better off dropping thinking skills objectives altogether and turning instead to the time-honored goal of helping students develop as thinking persons.

Real 21st-Century Competencies

The Organization for Economic Co-operation and Development (2010) has begun referring to its member nations as "innovation-driven." This implies a feed-forward process characterized not only by acceleration but also by unpredictability. Is there not therefore something rather ludicrous about educationists and business representatives sitting around a table and pretending to define the skills this uncertain future will require? The likelihood ought to be acknowledged that essential skill needs have yet to come into view and that a closer look at emerging capabilities and challenges might give a foretaste of what they will be (Scardamalia, et al., 2010). There are, however, cultural changes already in motion that bespeak competency needs schools are failing to address adequately. The following are five that everyone can see but that get little recognition in "21st-century" skill lists:

1. **Knowledge creation.** Except in a few areas—politics, religion, and education being the principal ones—21st-century societies recognize that the route to betterment lies through creation of new knowledge. As the health sciences advance, it becomes less clear what constitutes healthy diet, and some people throw up their hands and opt for a simplistic solution. Society's collective answer, however, is to pursue further research. And so it is with environmental problems, energy problems, homeland security problems, infectious disease problems, and all the grave problems that threaten to bring on societal collapse (cf. Diamond, 2005; Homer-Dixon, 2006). Producing the necessary knowledge requires not only an increase in the number of people capable of significant knowledge creation, but also a citizenry appreciative of and willing to support knowledge creation. A step toward meeting both needs is promoting a better understanding of the nature of knowledge and the nature of science—particularly the positive role that theory plays in the advance of knowledge. There is evidence that teachers by and large do not understand this, that they view theories as mere embodiments of the uncertainties of empirical knowledge (Windschitl, 2004). Developing students as knowledge creators involves a more radical transformation, however, one that authentically engages students as participants in a knowledge-creating culture (Bereiter & Scardamalia, 2006, 2010).

2. **Working with Abstractions.** Whether it is a doctor evaluating your state of health or a mechanic diagnosing a problem with your car, judgments that used to call for interacting with the object in question are now likely to be based on interaction with computerized data. This is true in an increasing range of manufacturing and service occupations. Even if technology processes the data into a realistic simulation so that you can apply skills of the old hands-on type, those data and their representation have been transformed by a theory that stands between the real phenomena and their presentation. A modern worker needs to be able to move flexibly and rationally between concrete reality and abstractions from it. Yet applying disciplinary knowledge to practical life has always involved abstraction. Information technology has only

made the need to negotiate between the concrete and the abstract ubiquitous. Every time you formulate a real-world problem as a mathematical problem and then do the math, you are performing such negotiation. Schooling, however, preserves a wariness of abstractness that was explicit in Dewey (1916, 183-185) and, in the mistaken view of many educators, given a theoretical basis by Piaget. Converting the abstract into the concrete remains an honored part of the art of teaching. "Mathematical modeling" is a fancy name for an effort to go the other way, converting the concrete to the abstract. There needs to be much more of this, extending outside mathematics to other kinds of modeling that facilitate practical action.

3. **Systems Thinking.** Dating from Herbert Simon's original work on "bounded rationality" (1957), it has been evident that most human predicaments are too complex for our limited information-processing capacity. And problems are getting more complex (Homer-Dixon, 2006). This is partly because more is known about the problems and partly because modern life is introducing more variables. The supermarket check-out question, "Paper or plastic?" is relatively easy to answer in terms of environmental impact, as long as one considers only the environmental consequences of the grocery bag ending up in a garbage dump. But if one traces the whole path from natural resources to manufacture and on through the life history of a grocery bag, the environmental impacts become so complex that even experts find it difficult to settle on a choice. Substantial theory has developed about complexity and how it evolves, self-organization being a central concept (Kaufmann, 1995). An educated person in mid-21st century will need to understand complexity scientifically, because of its pervasive significance throughout the natural and social world, but beyond that the educated person needs ability to live with increasing complexity and turn it to advantage wherever this is possible. Most of the detrimental ideologies that block progress on societal problems involve retreats from complexity, simplistic economic ideologies being perhaps the most widespread but by no means the only examples.

4. **Cognitive Persistence.** Cognitive persistence includes sustained study and pursuit of understanding, comprehending long texts, following extended lines of thought, and sustained creative effort turning promising initial ideas into fully developed designs, theories, problem solutions, and so on. The point is not that requirements for this kind of competence are increasing, although this may be true if increasing numbers of jobs involve work with complex ideas. The point, rather, is that obstacles and distractions from cognitive persistence may be increasing. We have

already noted concerns about the bite-sizing of discourse in modern media. Life on the Internet is full of distractions, which are causing employers to be concerned about work time being lost. In schools and colleges, the heavy emphasis on examinations may be encouraging spasmodic cramming rather than cumulative intellectual work. Both motivational and cognitive issues are involved. A Canada-wide study showed a progressive decline across the school years in "intellectual engagement," which the authors distinguished from social and academic engagement (Willms, Friesen, & Milton, 2009). Conservative social critics see the problem as a more pervasive decline in work ethic, with schools being parties to this decline (cf. Malanga, 2010; Murray, 2012). Whether things are actually getting worse is something that will not be evident without research extending over years. But that cognitive persistence is something deserving serious educational attention seems clear.

5. **Collective Cognitive Responsibility.** Collective responsibility characterizes expert teams of all kinds. It goes beyond the current buzzword, "collaboration," in that it means not only everyone working productively together but also everyone taking responsibility for success of the whole enterprise. Collective *cognitive* responsibility adds "collective responsibility for understanding what is happening, for staying cognitively on top of events as they unfold…. For knowing what needs to be known and for insuring that others know what needs to be known" (Scardamalia, 2002, pp. 68-69). Whereas collective responsibility for getting a job done may be as old as the species (think of hunting down a mastodon), collective cognitive responsibility has a distinct 21st-century flavor. Coming decades are likely to see the spread of "massively collaborative" problem solving and idea development (Greene, Thomsen, & Michelucci, 2011, October), where cognitive responsibility will be very widely distributed and leaders may be able to facilitate but will no longer be able to manage it. Collective cognitive responsibility is already essential for design teams, research teams, planning teams, and the like, especially when there is a leveling of status hierarchies. Schools can be good places for developing the ethos and the competencies for it, although this requires teachers turning some of their traditional cognitive responsibility over to the students while ensuring that collaborative activities are rich in cognitively challenging possibilities.

Implications for Measurement and Assessment

This chapter has contained more than the usual number of question marks, which may be justified, considering that our task has been to look ahead 40 years and speculate about the effects of changes only beginning to take shape. The question marks signal research needs, but how the needed research is to be grounded itself raises additional and in some cases deeper questions. Throughout the discussion we have emphasized a developmental view of what it means to become an educated person in mid-21st century in contrast to a piecemeal skills-and-knowledge view, and this view raises a host of questions. 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. At present there are, for instance, well-established tests of so-called critical thinking, and one could imagine making a test of this sort part of a program to periodically assess progress in students' development of critical thinking. But between doing well on a test of reasoning and other abilities believed essential for critical thinking and actually being a critical thinker, there is a wide gap. In a statement of "expert consensus," the crafters of specifications for the original California Critical Thinking Skills Test recognized that critical thinking involves more than skills:

> The ideal critical thinker is habitually inquisitive, well-informed, trustful of reason, open-minded, flexible, fair-minded in evaluation, honest in facing personal biases, prudent in making judgments, willing to reconsider, clear about issues, orderly in complex matters, diligent in seeking relevant information, reasonable in the selection of criteria, focused in inquiry, and persistent in seeking results which are as precise as the circumstances of inquiry permit. (Facione, 1990, p. 2).

This could serve as a succinct description of critical thinking as a developmental goal. But how then rationalize assessing critical thinking by what is mainly a test of logical evaluation of arguments? According to Facione (1991), the reasoning was that scores on the test would reflect presence or absence of the dispositions suggested in the above description. Thus, presumably the 34-item test could act as a surrogate for a much more extensive developmental assessment. This is a huge leap of faith. Assessment of 21st-century educational outcomes should not have to depend on such leaps of faith.

What would be indicators of actually being a critical thinker? One would look for evidences of the cognitive virtues itemized in the "expert consensus," with special credit for turning critical scrutiny upon one's own cherished or conventional beliefs. Relevant observations might be few and far between, but to the extent that classroom discourse took place through online writing or through speech that could be converted to text, sufficient evidence could be available as a basis for assessment. Automated discourse analysis could even make such assessment relatively effortless (cf. Dönmez, et al., 2005). This presupposes, of course, that the classroom discourse is of such a nature that critical thinking has a chance to reveal itself. Teachers, thus, are essential to the assessment process not only as observers or evaluators, but as enablers of the kind of activity that is intended to be assessed.

What we have just said with regard to critical thinking could be extended to many other developmental objectives, whether formulated as skills, habits of mind, intellectual virtues, attitudes, or dispositions. Creativity, a favored 21st-century objective, needs to be assessed as more than a cognitive ability or a personality trait (Renzulli, 2002). Among other things, it is a lifestyle and a career choice (Sternberg, 2003), and it may characterize not only an individual career, but also the lifestyle of a community (Florida, 2002). Hence, there is value in treating creativity as a developmental goal rather than a skill goal; for treating it as the goal of developing into a person who creates. Portfolio assessment might serve better than tests as both a measure and a motivator of creativity development.

The assessment of knowledge appears to be on much solider ground than the assessment of intellectual skills (which in some cases may have no ground at all and may only represent the mislabeling of abilities as skills). Knowledge assessment faces two mounting challenges, however, which we have already noted: the problem of assessing depth and the problem of assessing quantity of unprescribed knowledge acquisition. Assessing depth becomes a tractable problem to the extent that an ascending order of things-to-be-understood, can be specified. This is a curriculum objectives problem, first and foremost, and only subsequently becomes a testing problem. The other challenge, however, is an assessment problem from the beginning. Stated most generally, how do we measure knowledge growth when students are not all expected to learn the same things? As self-directed and inquiry learning grow, this problem becomes more urgent. It is hard to get educators wholeheartedly committed to greater student agency and to questions that really matter to the students when in the end everyone is going to be evaluated on the same set of prescribed learnings. However, there is no dilemma if the list of required knowledge is really short and consists of concepts that are really important and really powerful. It is only when the required objectives suck up the whole curriculum that student agency becomes trivialized. Student agency is not stifled if students know what the big ideas are and why they are important. That becomes information even young students can factor into their knowledge building plans.

But what is to be done about learning that is free to vary, and where sheer quantity may count for something? We have a model of this kind of estimation problem from studies of vocabulary size. Estimates of a person's total vocabulary size may be based on sampling from a large universe, such as that contained in a dictionary of between 100,000 and 200,000 words. People's scores will fall far short of the 80 percent conventionally taken to constitute mastery of a knowledge domain, but they will be sufficient to yield an estimated vocabulary size in the tens of thousands. A universe of incidental knowledge items might be generated from semantic analysis of websites, or vocabulary size itself could be a surrogate for miscellaneous knowledge. So there is methodology available for estimating knowledge “size ”; whatever form it eventually takes, it will be something different from the familiar kinds of knowledge testing based on specifications of particular things that are supposed to be learned.

Finally, let us turn to the five “real 21st-century competencies” discussed previously: knowledge creation, working with abstractions, systems thinking, cognitive persistence, and collective cognitive responsibility. There is nothing particularly original or controversial about listing these as core competencies, yet they remain on the margins of current educational goal setting, treated as mere adjuncts to such A-list competencies as problem solving, critical thinking, and creativity. Perhaps one reason for this sidelining is that there is no obvious way to measure them. But recent research and development provides important new possibilities:

- Assessing knowledge creation. The idea of young students actually working at knowledge creation may seem outlandish, but there are well-recognized varieties of knowledge creation that fall within demonstrable capabilities of the young (Bereiter & Scardamalia, 2010). There is testable knowledge associated with knowledge creation—especially knowledge of how knowledge progresses. This is receiving attention through work on NOS—“nature of science” (e.g., Working Group on Teaching Evolution, 1998). There are measurable dimensions of knowledge-creating dialogue, such as “scientificness” (Zhang, et al., 2007), but so far the most promising means of assessment comes from having students describe group knowledge advances and their own and other students’ contributions to them (van Aalst & Chan, 2007).
- Assessing work with abstractions. Ability to work with abstractions seems like something that could be tested directly, although research is needed to find out whether it is a teachable generic skill or merely a convenient label. Cross-cultural research suggests that ability to work with abstractions may have considerable generality and is dependent on formal education (Scribner, 1979). While people of all kinds handle reasoning tasks better when they are represented concretely than when they are represented abstractly (Johnson-Laird, 1983), the size of the performance gap between the two conditions might prove to be an individual difference variable useful in assessing ability to work with abstractions.

- Assessing systems thinking. There is a knowledge base in systems theory. The problem is how to teach it, not how to test it. Arguably this is the outstanding challenge facing instructional designers, given that systems theory is becoming essential not only for scientific literacy but also for literacy in any theoretical domain. Beyond that, there are a number of experimental efforts to foster systems thinking. Results to date are more limited than one would like, considering the importance of the objective, but we can predict that by 2050 there will be effective ways to teach and test widely generalizable systems thinking abilities. Unfortunately, we may also predict that many school curricula will be unaffected by this research and will continue to treat systems thinking as synonymous with open-mindedness and flexibility.
- Assessing cognitive persistence. Regardless of what research may eventually show about alleged declines in attention span and regardless of the educational benefits new media may provide, it would help both teaching and assessment if at least once a year students were required to read a long and comparatively complex text (or set of texts relevant to some knowledge objective) and assessed in some depth as to their understanding. These could also be used to assess knowledge building at the group level, which, as we noted at the beginning of this chapter, warrants attention in its own right and not merely as a reflection of individual learning.
- Assessing collective cognitive responsibility. This calls for assessment at both the individual and the group level. It is hard to imagine doing this without supportive technology. Elaborations of automated social network analysis are already proving valuable in enabling both teachers and students to monitor types and levels of participation and discourse in collaborative knowledge building (Oshima, Oshima, & Matsuzawa, in press; Zhang, et al., 2007). On the horizon is automated assessment of idea evolution and of the social interactions that foster it.

Will Technology Facilitate Becoming an Educated Person?

That technology will continue to change at a dazzling rate has now become part of conventional wisdom, including conventional educational wisdom. Traditionally, educational practice has been said to lag 10 to 20 years behind the advance of knowledge. Surely this is no longer true with regard to the uptake of new technology, especially new hardware. However, there is little reason to believe that the lag in uptake of advances in pedagogy has diminished. The result is that new technology gets harnessed to old pedagogy

(Reimann & Markauskaite, 2010). Contributing to the cultural lag in education is the fact that most new information technology products are developed for business and then are adopted, often without modification, by educational institutions. The problem here is not that the products are inappropriate; technology for text production, image processing, and the like, once developed to the point of being usable in ordinary office work, can equally well be used in schools. The problem is that businesses do not generally use productivity software to educate their workers, whereas schools do, or at least aspire to do so. But the technology provides little support for educational missions, having never been designed for that purpose. Consequently we have, for instance, learning management systems with attached discussion boards that do not work very well in supporting educationally productive discussion yet have changed hardly at all in 20 years (Hewitt, 2005).

There is nothing inherently wrong with using new technology to perform old tasks, and there is truth in the cliché, “It isn’t the technology, it’s how you use it.” What is needed in order to advance on the educational challenges discussed in this chapter, however, is a dynamic relation between technological invention and social/cognitive/pedagogical invention, with each helping to ratchet up the other. We will mention two possibilities for such a dynamic relation, which are ones we happen to be working on. One seeks to support sustained work with ideas by bringing inputs from varied media into a common digital workspace where supports are provided for explanation-building, extended problem solving, and so on. The other aims to provide feedback to students and teachers relevant to their knowledge-building efforts, for instance by mapping rates of different kinds of dialogue contributions and growth in domain vocabulary and by alerting discussants to possible misconceptions or overlooked concepts.

The kind of change that would make technology truly supportive of educating the mid-21st-century person is the same change we have argued for throughout this chapter. It is a change that places human development goals as central and knowledge, skill, attitude, and other goals as subservient. There is impressive creative work going on in designing technology that puts difficult concepts and operations within reach of more students. At the same time, project-based, problem-based, and knowledge-building pedagogy are flourishing, all with the intent of giving students higher levels of agency in the learning process. But more technology is needed for bringing these strands together, especially by supporting the kind of dialogue that converts experiences acquired through games, experiments, projects, and arguments into coherent understanding (Scardamalia & Bereiter, 2006). The principal obstacles from our standpoint are not technological but reside in the absence of sophisticated consumer demand.

Conclusion

This has not been so much a forward-looking essay as one that examines the current state of our culture and its institutions and asks what education would be like if it actually addressed current needs. School reform is a thriving business for consultants. But it is proceeding on the basis of folk theories of learning, cognition, and action, largely oblivious to the past 35 years of relevant scientific and pedagogical advances. We have singled out thinking skills as an especially retrograde folk notion that seems to be deflecting educational reform from human development goals and knowledge goals that could be the focus of systemic change. Fortunately, the Partnership for 21st-Century Skills (www.p21.org), which started out with its main focus being on test-driven skill objectives, has evolved toward designing intellectually enhanced curricula in school subjects. This is a definite step toward broader human development objectives derived from a conception of what it will mean to be an educated person in the mid-21st century. We have tried in this chapter to identify critical intellectual aspects of this educated personhood. We have seen how traditional broad concerns such as depth of understanding, literacy, and moral reasoning take on new meanings and face new challenges. More specifically, however, we have identified five competencies that call for determined educational effort, yet remain on the sidelines of assessment and teaching: knowledge creation, working with abstractions, systems thinking, cognitive persistence, and collective cognitive responsibility. We would argue that these are the real "21st-century skills," the competencies needed for productive and satisfying life in an "innovation-driven society."

References

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching and assessing: A revision of Bloom's Taxonomy of educational objectives: Complete edition.* New York: Longman.

Banks, J. A. (1994). *An introduction to multicultural education.* Needham Heights, MA: Allyn and Bacon.

Bereiter, C., &Scardamalia, M. (1998). Beyond Bloom's taxonomy: Rethinking knowledge for the knowledge age. In A. Hargreaves, A. Lieberman, M. Fullan, & D. Hopkins (Eds.), *International handbook of educational change* (pp. 675-692). Dordrecht: Kluwer.

Bereiter, C., & Scardamalia, M. (2006). Education for the knowledge age: Design-centered models of teaching and instruction. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology* (2nd ed., pp. 695-713). Mahwah, NJ: Lawrence Erlbaum Associates.

Bereiter, C., & Scardamalia, M. (2010). Can children really create knowledge? *Canadian Journal of Learning and Technology,* 36(1). Web: http://www.cjlt.ca/index.php/cjlt/article/view/585.

Bloom, B. S. (Ed.). (1956). *Taxonomy of educational objectives: Handbook 1. Cognitive domain*. New York: David McKay Company, Inc.

Burke, J. B., Hansen, J. H., Houston, W. R., and Johnson, C. (1975). *Criteria for describing and assessing competency programs.* Syracuse: National Consortium of Competency-based Education Centers.

Byrne, D. (1998). *Complexity theory and the social sciences: An introduction.* London: Routledge.

Cassino, D., & Woolley, P. (2011, Nov. 21). *Some news leaves people knowing less*. Web document: http://publicmind.fdu.edu/2011/knowless/

Costa, A. L., and Kallick, B. (Eds.) (2000). *Discovering and exploring the habits of mind.* Alexandria, VA: ASCD.

Detterman, D. K. (1993). The case for prosecution: Transfer as an epiphenomenon. In D. K.

Detterman & R. J. Sternberg (Eds.), *Transfer on trial: Intelligence, cognition, and instruction* (pp. 39–67). Stamford, CT: Ablex Publishing Corp.

Dewey, J. (1916). *Democracy and education*. New York: Macmillan.

Diamond, J. (2005). *Collapse: How societies choose to fail or succeed*. New York: Viking Books.

Dönmez, P., Rose C.P., Stegmann, K., Weinberger, A., Fischer, F. (2005). *Supporting CSCL with automatic corpus analysis technology*. Proceedings of Computer Supported Collaborative Learning, Taipei, Taiwan.

Duncan, R. G., & Hmelo-Silver, C. E. (2009). Learning progressions: Aligning curriculum, instruction, and assessment. *Journal of Research in Science Teaching*, 46, (pp. 606-609).

Facione, P. A. (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction*. Milbrae, CA: California Academic Press. (ERIC ED 315 423).

Facione, P. A. (1991). *Using the California Critical Thinking Skills Test in research, evaluation, and assessment*. Milbrae, CA: California Academic Press. (ERIC ED 337 498).

Fadiman, C. (1959). The case for basic education. In J. D. Koerner (Ed.), *The Case for Basic Education* (pp. 3-14). Boston: Little, Brown.

Florida, R. L. (2002). *The rise of the creative class: and how it's transforming work, leisure, community and everyday life*. New York: Basic Books.

Greene, K., Thomsen, D., Michelucci, P. (2011, October). *Explorations in massively collaborative problem solving*, Paper presented at the Third IEEE Conference on Social Computing 2011, Boston, MA.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108, 814-834.

Hewitt, J. (2005). Toward an understanding of how threads die in asynchronous computer conferences. *Journal of the Learning Sciences*, 14(4), 567-589.

Hirsch, E. D., Jr. (1987). *Cultural literacy: What every American needs to know.* Boston, MA: Houghton Mifflin

Homer-Dixon, T. (2000). *The ingenuity gap: Facing the economic, environmental, and other challenges of an increasingly complex and unpredictable world*. New York: Knopf.

Homer-Dixon, T. (2006). *The upside of down: Catastrophe, creativity, and the renewal of civilization*. Washington, DC: Island Press.

Huck, S. W. (2009). *Statistical misconceptions*. New York: Psychology Press, Taylor & Francis group.

Johnson-Laird, P. N. (1983). Mental models: Towards a cognitive science of language, inference and consciousness. Cambridge, UK: Cambridge University Press.

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *American Economic Review*, 96, 1449-1475.

Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Farrar, Straus and Giroux.

Kauffman, S. (1995). *At home in the universe: The search for laws of self-organization and complexity*. New York: Oxford University Press.

Koerner, J. D. (1959). The case for basic education. Boston: Little, Brown.

Kline, P. (1998). *The new psychometrics: Science, psychology and measurement.* London: Routledge.

Malanga, S. (2010). *Shakedown: The continuing conspiracy against the American taxpayer*. Chicago: Rowan & Littlefield.

Mayer, R.E., & Wittrock, M. C. (2006). Problem solving. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology* (2nd ed., pp. 285-303). Mahwah, NJ: Lawrence Erlbaum Associates.

Moursund, D. (1999). *Project-based learning using information technology*. Eugene, OR: International Society for Technology in Education.

Murray, C. (2012). *Coming apart: the state of white America*, 1960-2010. New York: Crown Forum.

Organization for Economic Co-operation and Development. (2010). *The OECD Innovation Strategy: Getting a head start on tomorrow*. Paris, OECD.

Ohlsson, S. (1991). *Young adults' understanding of evolutionary explanations: Preliminary observations* (Tech. Rep. to OERI). Pittsburgh: University of Pittsburgh, Learning Research and Development Center.

Olson D. R. (1977) From utterance to text: The bias of language in speech and writing. *Harvard Educational Review*, 47(3), 257-281.

Oshima, J., Oshima, R., & Matsuzawa, Y. (in press). Knowledge Building Discourse Explorer: A social network analysis application for knowledge building discourse. *Educational Technology Research and Development*.

Partnership for 21st Century Skills. (n.d.). Global awareness.
Web document http://www.p21.org/overview/skills-framework/256

Perkins, D. (1995). *Outsmarting IQ: The emerging science of learnable intelligence*. New York: Free Press.

Perkins, D. N., & Unger, C. (1999). Teaching and learning for understanding. In C. M. Reigeluth, (Ed), *Instructional-design theories and models: A new paradigm of instructional theory*, Volume II. Pp. 91-114. Mahwah, NJ: Lawrence Erlbaum Associates.

Rather, D., & Diehl, D. (2012). Rather outspoken: My life in the news.

Renzulli, J. S. (2002). Emerging conceptions of giftedness: Building a bridge to the new century. Exceptionality, 10(2), 67-75.

Resnick, M., Maloney, J., Monroy-Hernandez, A., Rusk, N., Eastmond, E., Brennan, K., Millner, A., Rosenbaum, E., Silver, J., Silverman, B., & Kafai, Y. (2009). Scratch: Programming for all. *Communications of the ACM*, vol. 52, no. 11, pp. 60-67 (Nov. 2009).

Rizvi, F. (2008). Epistemic virtues and cosmopolitan learning. *Australian Educational Researcher*, 35(1), 17-35.

Sawyer, R. K. (2003). *Group creativity: Music, theater, collaboration*. Mahwah, NJ: Lawrence Erlbaum Associates.

Scardamalia, M. (2002). Collective cognitive responsibility for the advancement of knowledge. In B. Smith (Ed.) *Liberal education in a knowledge society* (pp. 67-98). Chicago: Open Court.

Scardamalia, M., & Bereiter, C. (2006). Knowledge building: Theory, pedagogy, and technology. In K. Sawyer (Ed.), *Cambridge Handbook of the Learning Sciences* (pp. 97-118). New York: Cambridge University Press.

Scardamalia, M., Bransford, J., Kozma, R., & Quellmalz, E. (2010). *New assessments and environments for knowledge building*. White Paper for the Assessment and Learning of 21st Century Skills Project. Paper posted to http://www.atc21s.org/home/

Scribner, S. (1979). Modes of thinking and modes of speaking: Culture and logic reconsidered. In R. O. Freedle (Eds.), *New directions in discourse processing* (pp. 223-243). Norwood, NJ: Ablex.

Simon, H. A. (1957). *Models of man: Social and rational: Mathematical essays*. New York: Wiley.

Stahl, G. (2006). *Group cognition*. Boston, MA: MIT Press.

Stanovich, K. E. (2004). *The robot's rebellion: Finding meaning in the age of Darwin*. Chicago: University of Chicago Press.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate. *Behavioral and Brain Sciences*, 23, 645–665.

Sternberg, R. J. (2003). The development of creativity as a decision-making process. In R. K. Sawyer, V. John-Steiner, S. Moran, R. J. Sternberg, J. Nakamura, & M. Csikszentmihalyi (Eds.), *Creativity and development* (pp. 91-138). New York: Oxford University Press.

Thagard, P. (2006). *Hot thought: Mechanisms and applications of emotional cognition*. Cambridge, MA: MIT Press.

Van Aalst, J., & Chan, C. K. K. (2007). Student-directed assessment of knowledge building using electronic portfolios in Knowledge Forum. *The Journal of the Learning Sciences*, 16, 175-220.

Willms, J. D, Friesen, S., & Milton, P. (2009). *What did you do in school today? Transforming classrooms through social, academic and intellectual engagement*. (First National Report). Toronto: Canadian Education Association. Retrieved from http://www.cea-ace.ca/publication/what-did-you-do-school-today-transforming-classrooms-through-social-academic-and-intellectual engagement

Wilson E.O. (1996). *In Search of Nature*. Washington, DC: Island Press.

Windschitl, M. (2004). Folk theories of “inquiry:” How preservice teachers reproduce the discourse and practices of an atheoretical scientific method. *Journal of Research in Science Teaching*, 41(5), 481-512.

Working Group on Teaching Evolution, National Academy of Sciences. (1998). *Teaching about evolution and the nature of science*. Washington, DC: National Academies Press.

Zhang, J., Scardamalia, M., Lamon, M., Messina, R., & Reeve, R. (2007). Socio-cognitive dynamics of knowledge building in the work of 9- and 10-year-olds. *Educational Technology Research and Development*, 55, 117-145.

5. POSTMODERN TEST THEORY

Robert Mislevy

(Reprinted with permission from ***Transitions in Work and Learning: Implications for Assessment, 1997***, by the National Academy of Sciences, Courtesy of the National Academies Press, Washington, D.C.)

Good heavens! For more than forty years I have been speaking prose without knowing it.

Molière, *Le Bourgeois Gentilhomme*

INTRODUCTION

Molière's Monsieur Jourdan was astonished to learn that he had been speaking prose all his life. I know how he felt. For years I have just been doing my job—trying to improve educational assessment by applying ideas from statistics and psychology. Come to find out, I've been advancing "neopragmatic postmodernist test theory" without ever intending to do so. This paper tries to convey some sense of what this rather unwieldy phrase means and offers some thoughts about what it implies for educational assessment, present and future. The good news is that we can foresee some real improvements: assessments that are more open and flexible, better connected with students' learning, and more educationally useful. The bad news is that we must stop expecting drop-in-from-the-sky assessment to tell us, in 2 hours and for $10, the truth, plus or minus two standard errors.

Gary Minda's (1995) *Postmodern Legal Movements* inspired the structure of what follows. Almost every page of his book evokes parallels between issues and new directions in jurisprudence on the one hand and the debates and new developments in educational assessment on the other. Excerpts from Minda's book frame the sections of this paper. They sketch out central ideas in postmodernism—neopragmatic postmodernism, in particular—and how they are transforming the theory and practice of law. Their counterparts in assessment are discussed in turn.

Modernism and Postmodernism

This section introduces the terms "modernism," "postmodernism," and "neopragmatic postmodernism" as I will use them. It is necessarily incomplete, and adherents of each position will find much to disagree with.

Modernism

Modern legal theorists believe that they can discover the “right answers” or “correct interpretation” by applying a distinctive legal method based on deduction, analogy, precedent, interpretation, social policy, institutional analysis, history, sociology, economics, and scientific method . . . . Legal moderns . . . express the intellectual and artistic quest for perfection through the process of *uncovering* and *unmasking* the secrets of the world by transcending contexts that limit human understanding . . .. Legal modernism also . . . is based on an understanding of language that assumes that words and conceptual ideas are capable of objectively capturing the meaning of events the law seeks to describe and control. (Minda, 1995:5-6)

To Plato the nature and intelligibility of the world of appearance could be accounted for only by recognizing it as an “image” of the truly intelligible structure of being itself. These “forms” are the essence of being in the world, although we experience only images or imperfect instances of this or that. He likened our condition to that of dwellers in a cave, who see shadows on cave walls but not the objects that cast them. The struggle, by means of logic and the scientific method, to infer the universe’s “true” forms and to explicate their invariant relationships to experience characterizes what we may call the modern approach.

Modernism in physics, for example, can be illustrated by the prevailing belief, up through the beginning of the twentieth century, that objects exist in a fixed Euclidean space and interact in strict accordance with Newton’s laws. Measurement was a matter of characterizing properties of objects such as their mass and velocity—with uncertainty to be sure but only from the imperfections of our measuring devices. The variables were the universe’s; the distance between our knowledge and the truth was quantified by a standard error of measurement and shrank toward zero as we fine-tuned our models and improved our instruments.

In law the essence of modernism is the idea that “there is a ‘real’ world of legal system ‘out there,’ perfected, formed, complete and coherent, waiting to be discovered by theory” (Minda, 1995:224). The source was debated, to be sure: Dean Christopher Columbus Langdell (in 1871) maintained that careful study of cases should reveal the underlying axioms of justice, from which “the law” in its entirety follows logically. Oliver Wendell Holmes argued pragmatically that “law and its institutions evolved from views of public policy, social context, history, and experience” (Minda, 1995:18) and that its application always relies on judgments about its role in society.

In educational and psychological testing, modernism corresponds to the pursuit of models and methods that characterize people through common variables, as evidenced by common observations—under the conceit that there are objectively correct ways of doing so. The source of these models and variables has been debated over the years along logical versus pragmatic lines analogous to the Langdell versus Holmes stances. Witness on the one hand, factor analytic research programs that seek to "discover" the nature of intelligence and personality (e.g., Spearman, 1927; Thurstone, 1947) and, on the other, painstaking consensus-building procedures for assembling item pools to "measure achievement" in subject domains (e.g., Lindquist, 1951). These distinct branches within modern test theory correspond to the trait and behaviorist psychological perspectives. Under both perspectives, care is taken (1) to define, from the assessor's point of view, contexts in which to observe students; (2) to specify, from the assessor's perspective, the ways in which students' behavior will be summarized; and (3) to delineate the operations through which the assessor can draw inferences, within the assessor's frame of reference.

Postmodernism

> In jurisprudence, postmodernism signals the movement away from "Rule of Law" thinking based on the belief in one true "Rule of Law," one fixed "pattern," set of "patterns," or generalized theory of jurisprudence . . .. As developed in linguistics, literary theory, art, and architecture, postmodernism is also a style that signals the end of an era, the passing of the modern age . . . describing what happens when one rejects the epistemological foundations of modernity. (Minda, 1995:224)
>
> Wittgenstein's view of language is that all of our language has meaning only within the language games and "forms of life" in which they are embedded. One must understand the use, the context, the activity, the purpose, the game which is being played . . . . (Minda, 1995:239)

The notion of discourse plays a central role in postmodernism. Language generates our "universe of discourse:" the kinds of things we can talk about and the particular things we can say; what we construe as problems, how we attempt to solve them, and how we evaluate our success. But what is the source of words and concepts? Postmoderns claim that the commonsense idea that meanings of words reside "in" language is fundamentally misguided. For them, language constructs, rather than reflects, the meaning of things and events in the world (Minda, 1995:239).

Relativity and the quantum revolution shattered the belief that Newtonian and Euclidean models were the correct ultimate description of the universe. Ironically, improved instrumentation devised to finalize the modern research program revealed that its fundamental models were not in fact the universe's. Mathematical descriptions of observations departed increasingly from such intuitive notions as simultaneity and definitive locations of persistent entities. Just as ironically, while we obtain better accuracy in modeling phenomena and more power to solve applied problems than the "modern" physicists of the nineteenth century dreamed, we feel farther away from ultimate understanding. The universe is not only stranger than we imagine, mused the mathematician J.B.S. Haldane, it is stranger than we *can* imagine!

Just as relativity and quantum mechanics gave rise to postmodern physics, Minda noted several diverse movements that provoked a postmodern era in law in the 1980s: law and economics, critical legal studies, feminist legal theory, law and literature, and critical race theory. Cognitive psychology was the analogous shock to educational assessment—in particular, recognition of the crucial roles of students' perspectives in learning and of the social settings in which learning takes place. Snow and Lohman (1989:317) put it this way:

Summary test scores, and factors based on them, have often been thought of as "signs" indicating the presence of underlying, latent traits. . . . An alternative interpretation of test scores as samples of cognitive processes and contents, and of correlations as indicating the similarity or overlap of this sampling, is equally justifiable and could be theoretically more useful. . . . Whatever their practical value as summaries, for selection, classification, certification, or program evaluation, the cognitive psychological view is that such interpretations no longer suffice as scientific explanations of aptitude and achievement constructs.

Neopragmatic Postmodernism

> Some postmodernists have adopted a *neopragmatic* outlook as an antidote to the postmodern condition. These postmodern critics are skeptical of the truth claims of modern theory, but they have not given up on theory. On the contrary, they believe that theory can have utility when used as a tool for the empirical investigation of problems. . . . Its practitioners accept the postmodern view that truth and knowledge are culturally and linguistically conditioned. On the other hand, neopragmatist practice is unlike . . . what some theorists call *poststructuralist* criticism because it is less concerned with exposing the contradictions of modern conceptual and normative thought than revealing instrumental, empirical, and epidemiological solutions for the problem at hand. (Minda, 1995:229-230)

Minda distinguishes "neopragmatic" postmodernists from "ironic" postmodernists, the latter of which "embrace the predicaments and paradoxes of the current intellectual condition in order to better understand the world of legal, social, and philosophical thought, and they attempt to bring out the irony of the experience of living in a postmodern world" (1995:4-5). In legal theory "the ironists attempt to facilitate the crisis and fragmentation of modern theory by employing postmodern criticism to 'displace, decenter, and weaken' central concepts of modern legal Western thought" (1995:230). In the fine arts, ironic postmodernism is rather de rigueur. Physics and, by extension, engineering demand a neopragmatic stance. Models and variables may indeed be our creations rather than nature's, and they are ever subject to alternatives and revisions—but we must in some way accommodate the constraints nature imposes upon us as we struggle with the challenges we confront. And if there is a job to do, languages, models, and conceptual frameworks are what we have to work with.

Like law, educational assessment lies somewhere between literature and physics. Cognitive research reveals recurrent patterns in the ways people learn and solve problems, yet what is important to learn and the conditions under which it will be learned are largely socially determined. "Neopragmatic postmodern test theory" explores the potential of using methodological and inferential tools that originated in a modern perspective to support learning in ways conceived in a postmodern perspective.

MODERN TEST THEORY

Technical Considerations

> "Legal modernism also . . . is based on an understanding of language that assumes that words and conceptual ideas are capable of objectively capturing the meaning of events the law seeks to describe and control" (Minda, 1995:6).

Most familiar practices of educational assessment can be traced to the first third of the twentieth century. Their forms were shaped by constraints on gathering and handling data in that era and by purposes conceived under then-current beliefs about learning and schooling. A paradigm of mental measurement analogous to classical (read "modern") physical measurement developed, and the tools of test theory evolved to guide applied work within this setting—designing tests, characterizing their evidential value, and evaluating how well they achieved their intended purposes. The targets of inference are aspects of students' learning, characterized as numbers on a continuum, upon which evaluations and decisions would be based if they were known with certainty.

In his 1961 article "Measurement of learning and mental abilities," Harold Gulliksen (1961:9) characterized the central problem of test theory as "the relation between the *ability* of the individual and his *observed score*." Referring explicitly to Plato's cave, he said "the problem is how to make the most effective use of these shadows (the observed test scores) in order to determine the nature of reality (ability) which we can only know through these shadows." The purposes of test theory, in this view, are to guide the construction of assessment elements and events (i.e., domains of test items and test conditions) and to structure inference from students' behavior in the resulting situations. The modernist underpinnings of the enterprise are reflected in a quotation from Gulliksen's (1961:101-102) review of test theory on the occasion of the twenty-fifth anniversary of the Psychometric Society, concerning the search for the "right" item-response-theory models:

> An attempt to develop a consistent theory tying test scores to the abilities measured is typified by Lord's (1952) recent work . . . in which he formulated at least five different theories of the relationship between test scores and abilities, and showed how it was possible to test certain ones of these. It is to be hoped that during the next 10 or 20 years a number of these tests will be carried out so that we will have not five different theories of the relationship between ability and test score and various possible trace lines, but we will be able to say that, for certain specified tests constructed in this way, here is the relationship between the score and the ability measured, and this is the appropriate trace line to use.

Social Considerations

> "Neopragmatists believe that theory merely establishes the rules for playing a particular language game" (Minda, 1995:236).

Although the physical measurement analogue connotes a certain objectivity and detachment, assessment based on the modernist approach nevertheless shapes, and is shaped by, social considerations. It structures conversations about learning in several ways:

- *Communication of expectations.* In and of themselves, domains of tasks and modes of testing convey, to students, teachers, and the public at large, what is important for students to learn and to accomplish.
- *Communication of results.* Once a domain of tasks and conditions of observation have been specified, a score and an accompanying measure of precision give a parsimonious summary of a student's behavior in the prescribed contexts that is easily transmitted across time and place.
- *Credibility of results.* Test scores earn credibility beyond the immediate circumstances of the assessment if the data have been verifiably gathered under prescribed conditions.

That traditional assessment procedures serve these purposes is quite independent of the fact that they evolved under the mental measurement paradigm. Any procedures that might rise in their stead to assess and communicate students' learning would, in some way, need to address the same functions.

PROGENITORS OF CHANGE

> The transition from the old to the 'new' jurisprudence began with the breakdown of the core beliefs and theories that served to define modern jurisprudence. The breakdown is partly a manifestation of the proliferation of new jurisprudential discourses and new movements in legal thought. (Minda, 1995:243)

I claimed earlier that developments in the psychology of learning and cognition brought about a postmodern era in assessment, and I shall say more about that later. These developments do indeed lay the groundwork for new developments in assessment, but I do not believe they were sufficient in and of themselves to change the field. Had modern testing seen satisfactory progress in its research agenda, there would have been less impetus for change. But in assessment, as in physics, improved methodology and inferential methods led away from, rather than toward, the anticipated solutions.

Developments in Methodology

> "There is a rising sentiment in the legal academy that modern legal theory has failed to sustain the modernists' hopes for social progress" (Minda, 1995:248).

Twenty-five years after Gulliksen's article, Charles Lewis observed that "much of the recent progress in test theory has been made by treating the study of the relationship between responses to a set of test items and a hypothesized trait (or traits) of an individual as a problem of statistical inference" (Lewis, 1986). New modeling and inferential techniques included item response theory, generalizability theory (Cronbach et al., 1972), structural equations modeling (e.g., Jöreskog and Sörbom, 1979), and the application of more powerful estimation methods from the statistical literature (e.g., Bock and Aitkin, 1981). They provided solutions to previously intractable problems such as tailoring tests to individual examinees and sorting out relationships in patterns of achievement test scores in hierarchical schooling systems.

These developments make for more efficient gathering of evidence and more powerful forms of argumentation for addressing questions that could be framed within the universe of discourse of modern test theory. But by requiring analysts to more clearly explicate their targets of inference and how observations provided evidence about them, these advances in modern test theory began to reveal important problems that lie beyond the paradigm's reach. The following two examples illustrate the point:

- How can we measure change, or can we? Through the use of standard test theory, evidence can be characterized and brought to bear on inferences about students' overall proficiency in behavioral domains, for determining students' levels of proficiency, comparing them with others or with a standard, or gauging changes from one point in time to another. Cronbach and Furby (1970:76) cautioned that characterizations about the nature of this proficiency or how it develops fall largely outside the paradigm's universe of discourse:

- Even when [test scores] X and Y are determined by the same operation [e.g., a true-score or item-response-theory model for a specified domain of tasks], they often do not represent the same psychological processes. At different stages of practice or development different processes contribute to the performance of a task. Nor is this merely a matter of increased complexity; some processes drop out, some remain but contribute nothing to individual differences within an age group, some are replaced by qualitatively different processes.
- *Differential item functioning (DIF)*. Classical test theory took test scores at face value, treating all response patterns with the same total score as identical. Item response theory explicated the conditions that would have to hold among patterns of item responses for total scores to capture all nonrandom variations among students. Essentially, the same expectation of success on each given task would have to hold for all students at a given true-score level, regardless of item content or students' background characteristics. Differential-item-functioning techniques devised to check these conditions often found that they failed in achievement tests—most importantly, in ways that related to curriculum (e.g., Miller and Linn, 1988) and solution strategies (e.g., Birenbaum and Tatsuoka, 1983; French, 1965). Because what is hard and what is easy is not universal—they depend, not surprisingly, on what and how students have been studying—summary scores inevitably fail to characterize some aspects of students' knowledge and progress.

Developments in Psychology

Cronbach and Furby's comments on measuring change presaged a growing awareness that domain-referenced assessment methodologies, including item response theory, were simply not rich enough to support discourse about the nature and progress of students' learning. In assessment, as in physics, however, merely recognizing inadequacies in a paradigm is not sufficient for change. Newton and Huygens debated the contradictory wave- and particle-like properties of light as early as the seventeenth century. Paradigms are not displaced by data, the saying goes; paradigms are displaced only by other paradigms. Conceptions of learning that ground a broader universe of discourse for assessment are emerging from cognitive and educational psychology. The following paragraphs review some key insights into the ways people acquire and use knowledge and skills. Each, it will be noted, accents the uniquely personal and socially conditioned nature of learning.

- *Mental models/schema theory.* A “mental model” or “schema” is a pattern of recurring relationships—anything from what happens at birthday parties to how to figure out unit prices to how to carry out conversations—with variables that correspond to particular ways the pattern can occur. Some schemas are informal and intuitive; others we learn in part formally and explicitly. David Rumelhart (1980:55) claims that schemas play a central role in all our reasoning processes. Once we can “understand” the situation by encoding it in terms of a relatively rich set of schemata, the conceptual constraints of the schemata can be brought into play and the problem readily solved.

No cognition is purely passive or data driven; we always construct meaning in terms of knowledge structures. Learning sometimes means adding bits to existing structures; sometimes it involves generalizing or connecting schemas; other times it involves abandoning important parts of schemas and replacing them by qualitatively different structures.

- *How expertise develops.* While experts in various fields of learning generally command more facts and concepts than novices, the real distinction lies in their ways of viewing phenomena and representing and approaching problems (e.g., Chi et al., 1981). Experts learn to work from what Greeno (1989) calls the “generative principles of the domain,” and they automatize recurring procedures (they “compile knowledge”) so that they can devote their attention to novel aspects of problems. Increasing “metacognitive skills” also mark developing expertise: self-awareness in using models and skill and flexibility in how to construct them, modify them, and adapt them to problems.
- *Situated learning.* Assessment has focused on aspects of learning that are characterized insofar as possible as properties of individual students. Yet the nature of the knowledge we construct is conditioned and constrained by technologies, information resources, and social situations as we learn about tools, physical and conceptual, and how and when to use them. For example, reading comprehension depends on one’s competence in recognizing words and parsing syntactic structures, but it also depends as much on an understanding of the context and substance of what the message is about. Students who have similar competencies with structural aspects of language can take vastly

> different meanings away from the same text, depending on their experience with the phenomena in question. These findings, along with those discussed above, argue that learning is more richly characterized in terms of the student's breadth and configurations of connections across social and substantive contexts than by success in a given domain of tasks—even though such success occurs only by virtue of those connections.

These crosscutting generalizations should not obscure the fact that cognitive psychology is a fractured, often fractious, field. Competing claims of rival researchers differ from one another as much as all differ from the trait and behaviorist perspectives. This is largely because different researchers are exploring different ranges of behavior, acquired and used under different circumstances. Birnbaum (1991:65) suggests:

Problem-solving depends on the manipulation of relatively fragmented and mutually inconsistent *micro theories*—each perhaps internally consistent, and each constituting a valid way of looking at a problem: "This will allow us to say, for example, that some [set of beliefs] is more appropriate than some [other set of beliefs] when confronted with problems of diagnosing bacterial infections. Scientists are used to having different—even contradictory—theories to explain reality. . . . Each is useful in certain circumstances." (Nilsson, 1991:45)

In assessment, as in law, the neopragmatic postmodernist welcomes all these lines of research as potentially useful tools for solving different practical problems; that is to say:

For postmodern legal scholars, choosing the "best" answer for legal problems requires "tactical" judgments and questions regarding the values of the decision maker much more than a quest for a so-called "best" argument. One consequence of this has been the realization that there exists a multiplicity of answers for law's many problems. (Minda, 1995:252)

Rapprochement

Good teachers have always relied on a wider array of means to learn about how the students in their classes are doing and to help plan further learning. Alongside the tests and quizzes they design and score under the mental measurement paradigm, they also use evidence from projects, work in class, conversations with and among students, and the like—all combined with additional information about the students, the schooling context, and what the students are working on. Teachers call these “informal” assessments, in contrast with the “formal” assessments typified by large-scale standardized tests.

The stark contrast between formal and informal assessment arises because to understand students’ learning and further guide it, teachers need information intimately connected with what their students are working on, and they interpret this evidence in light of everything else they know about their students and their instruction. The power of informal assessment resides in these connections. Good teachers implicitly exploit the principles of cognitive psychology, broadening the universe of discourse to encompass local information and address the local problem at hand. Yet precisely because informal assessments are thus individuated, neither their rationale nor their results are easily communicated beyond the classroom. Standardized tests do communicate efficiently across time and place—but by so constraining the universe of discourse that the messages often have little direct utility in the classroom.

The challenge now facing neopragmatic postmodern test theory is to devise assessments that, in various ways, incorporate and balance the strengths of formal and informal assessments by capitalizing on an array of methodological, technological, and conceptual developments.

POSTMODERN TEST THEORY

> “Postmodern legal critics employ local, small-scale problem-solving strategies to raise new questions about the relation of law, politics and culture. They offer a new interpretive aesthetic for reconceptualizing the practice of legal interpretation” (Minda, 1995:3).

Cognitive psychology challenges the adequacy of the “one-size-fits-all” presumption of standard assessment, which defines the target of inference in terms of an assessor-specified domain of tasks, to be administered, scored, and interpreted in the same way for all students. The door has been opened to alternative ways to characterize students’ proficiency and acquire evidence about it—ways that may involve observing students in different situations, interpreting their actions in light of additional information about them, or triangulating across context and situation, as may be required for one’s purpose (Moss, 1996).

Moss (1994) and Delandshere and Petrosky (1994) offer postmodern insights into assessment from a less structural perspective than mine, criticizing test theory as it is conceived from a modernist point of view. I am interested in the utility of model-based inference in assessment, as reconceived from a postmodernist point of view. I submit that concepts from psychology and inferential tools from model-based reasoning can support assessment practice as more broadly conceived—just as Newton's laws still guide bridge design, quantum mechanics and relativity theory notwithstanding. The essential elements of the approach are (1) understanding the important concepts and relationships in the learning area in order to know the aspects of students we need to talk about in the universe of discourse our assessment will generate and (2) determining what one needs to observe and how it depends on students' understandings, so as to structure assessment settings and tasks that will provide evidence about the above-mentioned aspects of students' understandings. Here is an example from a project I have been working on recently, concerning advanced placement (AP) Studio Art portfolios.

Viewed only as measurement, the AP Studio Art portfolio program would be a disaster. Students spend hundreds of hours creating the portfolios they submit for scoring at the end of the year, and raters who are art educators and teachers spend hundreds of hours evaluating the work—all to produce reliability coefficients about the same as those of 90-minute multiple-choice tests. The situation brightens when the program is viewed as a framework for evidence about skills and knowledge, around which teachers build art courses with wide latitude for topics, media, and projects. A common understanding of what is valued and how it is evaluated in the central scoring emerges through teacher workshops, talked-through examples with actual portfolios, and continual discussions about how to cast and apply rating rubrics to diverse submissions. Meaning emerges through countless conversations across hundreds of classrooms, each individual, but with some common concepts and shared examples of their use—each enriched and individuated locally in a way that grounds instruction and local evaluations, but with a common core that grounds more abbreviated program-wide evaluations. This is, at heart, a social phenomenon, not a measurement phenomenon. Carol Myford and I have found an item-response-theory measurement model for ratings valuable, nevertheless, to illuminate how raters use evaluative criteria and to characterize uncertainty about students' scores (Myford and Mislevy, 1995). We do not use the model to "gauge the accuracy of a measuring instrument." We use it to survey patterns of similarity and variation, of agreement and disagreement, among tens of thousands of virtual dialogues among students, raters, and teachers, through their portfolios—to the end of discovering sources of misunderstanding and cross talk that can frustrate the conversations.

Model-based reasoning is useful not so much for characterizing the unique essence of a phenomenon but as a tool of discourse—for organizing our thinking, for marshaling and interpreting evidence, and for communicating our inferences and their grounding to others. The discipline that model-based reasoning demands even benefits us when we don't believe the models are true: it is easier to notice phenomena that don't accord with the patterns we expect to see and, therefore, to revise our thinking. A skeptical attitude about models in assessment makes our uses of them more flexible, more powerful, and, ultimately, more effective at meeting and fulfilling the aims of education than they would be if we believed that they accurately captured the totality of the phenomenon.

From a modernist perspective, Lord Kelvin declared at the turn of the century "when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind." Measurement, in his eyes, was a one-off representation of truth. From a postmodern perspective, even if you can measure, your knowledge is still meager in a fundamental sense, but at least you have a practicable framework for discourse—for structuring action, for communicating your observations and your reasoning, for struggling with practical problems, for surprising yourself in ways that lead to further understanding. Lord Kelvin's quote is a modernist scientific version of Yogi Berra's "it ain't over 'til it's over." The postmodern response is Jesse Jackson's "and even then it ain't over."

SOME IMPLICATIONS OF A POSTMODERN PERSPECTIVE

> Neopragmatists thus attempt to explain how one can do theoretical work without rejecting all pretenses of foundational knowledge. Neopragmatists argue that the theorists must take a situated stance in their scholarship and adopt an instrumental approach to theory. Whatever works in context becomes the standard for their theoretical investigation and judgment . . .. When applied to legal studies, neopragmatism forms the academic perspective of scholars who reject all foundational claims of legal theory but remain committed to the view that legal theory can be useful for solving legal problems . . .. Neopragmatists thus believe in and are committed to the Enlightenment idea of progress, even while they resist using the modernist's framework. (Minda, 1995:230)

In the remainder of the main body of the paper, I offer comments from a neopragmatic postmodern perspective on enhancing familiar kinds of assessment, even while moving our interpretational perspective beyond its modernist roots. As an example, I address the question of the degree to which "adult literacy," an essential element of workplace skill, can be defined and gauged across literacy training programs.

Progress Within Modern Test Theory

Familiar forms of assessment were shaped by constraints on how data could be gathered, stored, transmitted, and analyzed. Logistical and economic pressures limited the large-scale use of essays and interviews that required human interpretation, thus favoring objective-response tasks over more constructive and sustained tasks. It was not possible to store or share ephemeral performances in order to develop common standards or to verify that ratings were fair. These constraints are being eased by technological developments—computers, video-taping and audio-taping, electronic communication, mass storage, and access to resources. Some new possibilities appear even within the traditional mental measurement paradigm. I will mention some briefly but then argue that technology alone will not break through the essential inferential barriers of the modernist test theory perspective:

- *New kinds of tasks and scoring.* Computers can present students with tasks that are interactive (e.g., simulated experiments), dynamic (e.g., medical treatment problems in which simulated patients' conditions change over time), constructive (e.g., a stimulated construction site onto which elements must be moved to meet client's needs), and less tightly structured (e.g., a word problem that is to be approached in many ways). Some scoring can be also done automatically, including for these examples.
- *Distributed testing and scoring.* Students' responses to computerized tasks can now be captured and electronically transmitted. Performances can be videotaped and audio-taped. Constructed paper-and-pencil responses and artwork can be scanned. Students can thus be assessed in remote places and at different times, and raters can evaluate their performances in remote places and at different times. Students in school consortiums can share work on a common project, interacting with, and receiving feedback from, teachers and students across the nation.
- *"Replayability" (Frederiksen and Sheingold, 1994).* Beside easing constraints on time and location, capturing performances helps address the rater-agreement problems that troubled Horace Mann more than a century ago. Performances can now be seen, discussed, and evaluated by as many people, in as many times and places, as desired. Now that we are no longer limited to the evaluations of raters present at the original performance, we can have broader and more interconnected scoring of individual students and use exemplars to establish shared expectations and standards of evaluation, over time and across distance, among raters, teachers, and students.

Despite technology and efficient statistical models, the objective of characterizing students' proficiency must remain poorly met if one is constrained to one-size-fits-all data and to ignorance of contextual and educational background factors. The more examinees differ as to relevant contextual and experiential factors, the more likely it is that each task in a complex and context-rich domain will consume considerable time and costs without providing much information about how students would fare on other tasks—the Shavelson et al. (1992) "low generalizability" problem (also see Linn, 1993). Each individual task may provide copious information for some inferences—but not for inferences about the usual target, domain-true score. The same complex task can be invaluable in an assessment linked with instruction and grounded in context yet worthless in a broadly cast survey because it is trivial, unapproachable, or incomprehensible to most students.

Beyond Modern Test Theory

> Postmodernism thus challenges legal thinkers to reconsider their most basic understanding of the nature of law and politics—their belief in an objective and autonomous law. Postmoderns argue that decision making according to rule is not possible, because rules are dependent upon language, and language is socially and culturally constructed and hence incapable of directing decision makers to make consistent and objective choices. Objectivity is possible only if agreement or consensus about different interpretive practices can be reached. (Minda, 1995:245)

Standards of content and performance are a topic of intense current interest. I have argued that limited sets of common assessment tasks, scored and interpreted in common ways that ignore context, cannot by their nature tell us all we would like to know about students' learning. They may tell us something worth knowing, especially if the inferences and actions based on them do take context into account (Messick, 1989). As noted in the preceding section, technology is broadening the span and efficiency of such assessment. And with such assessment it is possible to gauge the levels of performance of individual students and groups of students. The real challenge, it seems, is to extend the notion of standards beyond the confines of the modernist perspective: Is it possible to retain the relevance and connectedness traditionally associated with informal assessment yet simultaneously serve the communicative and credibility-based functions traditionally associated with formal assessment?

The AP Studio Art experience suggests that the answer is yes. Learning there is individuated, but a shared conception of the nature of intended learning, developed through examples and feedback, makes it possible to interpret work in a common framework. Such a structure appears necessary if assessments with constructive and individuated data, such as portfolios and exhibitions (e.g., Wiggins, 1989), are to span time and distance.

Common meaning is necessary for credibility, but it is not sufficient. Why should anyone trust an interpreted evaluation of a performance from a distant time and place? Standardized test results gain a measure of credibility from their prescribed procedures; these are established “rules of the game,” which, if followed, circumscribe the interpretation of the results. Even though the results don’t tell about everything that is important, parents and boards of education can ask questions and verify procedures in order to spot invalidating practices. But the more individuated an assessment is, the more difficult it becomes to establish credibility.

For example, in some ways teachers are in the best position to evaluate students’ work, by virtue of their knowledge about context and situation. Their contextualized evaluations are unquestionably basic for guiding classroom learning. Can their evaluations be used for high-stakes purposes beyond the classroom, in light of their vested interest in their students’ success and the typically wide variation in their interpretations of performance? As noted above, a common framework for interpretation is required first. The validity of mappings of performances into that framework can be addressed by mechanisms such as audits, cross evaluation across schools, and triangulation of types of evidence (Resnick, in press). Technology can play an enabling role, through replayability, mass storage, and electronic communication. Statistical modeling can play a quality assurance role, through the analysis of ratings of multiply scored work, as discussed above in connection with AP Studio Art.

An Example: Adult Literacy Assessment

As defined by the National Literacy Act of 1991, literacy involves “an individual’s ability to read, write, and speak in English, compute, and solve problems at levels of proficiency necessary to function on the job and in society, to achieve one’s goals, and to develop one’s knowledge and potential.” The act requires state education agencies to “gather and analyze data—including standardized test data—on the effectiveness of State-administered adult education programs, services, and activities, to determine the extent to which the State’s adult education programs are achieving the goals in the state plan” [to enhance levels of adult literacy and improve the quality of adult education services] (*Federal Register,* 1992). These federal evaluation requirements have prompted interest in identifying standardized tests and methodologies that are appropriate for assessing the effectiveness of adult education programs and for determining the feasibility of linking such tests in order to provide national trend data on program effectiveness (Pelavin Associates, 1994).

But the diversity of both the objectives and the participants served by adult education programs reflects a broad and multidimensional definition of literacy. Accordingly, adult education programs vary considerably with respect to the nature and level of skills they emphasize and with respect to the kinds of students with whom they work. Some strongly resemble and largely replace the academic reading experience that high schools supply, in order to help dropouts obtain General Education Development certificates. Others help immigrants and others who are literate in languages other than English to speak, read, and write in English. Still others work with adults who are literate, if not skilled, from the perspective of traditional schooling, in order to develop more specific skills for use in the workplace. Moreover, these diverse programs use tests for a broad variety of diagnostic, instructional, and evaluative purposes. Both the nature of the instruction and the purpose of testing determine the kinds of tests that will be appropriate.

Is it possible to link results from these varied tests, across the diverse programs, to secure a common metric for evaluating program effects and tracking trends over time? Writing on the prospect of calibrating disparate tests to common national standards, Andrew Porter (1991:35) wrote,

If this practice of separate assessments continues, can the results be somehow equated so that results on one can also be stated in terms of results on the other? There are those who place great faith in the ability of statisticians to equate tests, but that faith is largely unjustified. Equating can be done only when tests measure the same thing.

Professor Porter's skepticism is justified. We are perhaps too familiar with correspondence tables that give exchangeable scores for alternate forms of standardized tests. But they work only because the alternate forms were constructed to meet the same tight specifications; equating studies and statistical formulas merely put into usable form the evidentiary relationships that were built into the tests (see Linn, 1993, and Mislevy, 1993, for definitions, concepts, and approaches that have been developed to link educational tests for various purposes).

Statistical procedures neither create nor determine relationships among test scores. Rather, the way that tests are constructed and administered and the ways that the skills they tap relate to the people to whom they are administered determine the nature of the potential relationships that exist in evidence that scores from the various tests convey. Much progress has been made recently with statistical machinery for this purpose, with power beyond the expectations of educational measurement researchers a generation or two ago. However, we now recognize the objective of building once-and-for-all correspondence tables as a chimera—it is not simply because we lack the tools to answer the question but because the question itself is vacuous. Statistical procedures, properly employed, can be used to explicate the relationships that do exist in various times and places and harness the information they do convey for various purposes. Perhaps more importantly, they help us understand what information different tests do not, indeed cannot, convey for those purposes.

Thus, the first two conclusions listed below, about what can be expected from applying statistical linking procedures to adult literacy tests, are negative. They repudiate a naive modernist goal of rectifying various indicators of a common true variable, when those indicators have evolved to serve different purposes in different contexts, gathering qualitatively different kinds of information.

- *No single score can give a full picture of the range of skills that are important to all the different students in different adult literacy programs.*
- *No statistical machinery can translate the results of any two arbitrarily selected adult literacy tests so that they provide interchangeable information about all relevant questions about student competencies and program effectiveness.*

What *is* possible? Three less ambitious, but more realistic affirmative contingencies, each employing modernist statistical techniques from a neopragmatic postmodernist perspective. All require the prerequisite realization that no test scores can capture the full range of evidence about students' developing proficiencies within their courses, nor can they convey all that is needed to determine how well students are progressing toward their own objectives. This understood, here are some options for dealing with such information as there is in literacy test scores, when different literacy programs must use different tests in accordance with their differing goals and instruction:

- *Comparing directly the levels of performance across literacy programs in terms of common indicators of performance on a market basket of consensually defined tasks in standard conditions.* Some aspects of competence, and assessment contexts for gathering evidence about them, will be considered useful by a wide range of programs, and components of an assessment system can solicit information about them in much the same way for all. However, these "universal" assessments—and in particular pre-post comparisons with such assessments—provide seriously incomplete information to evaluate the effectiveness of programs, to the extent that their focus does not match the programs' objectives (to say nothing of the students' objectives!).

- *Estimating levels of performance of groups or individuals within clusters of literacy programs with similar objectives—possibly in quite different ways in different clusters—at the levels of accuracy demanded by purposes within clusters, with shared assessments focused on those objectives.* These components of programs' assessments might gather evidence for different purposes, types of students, or levels of proficiency, to complement information gathered by "universal" components.
- *Making projections about how students from one program might have performed on the assessment of another.* When students can be administered portions of different clusters' assessments under conditions similar to those in which they are used operationally, the joint distribution of results on those assessments can be estimated. These studies are restricted as to time, place, program, and population, however. The more the assessments differ as to their form, content, and context, the more uncertainty is associated with the projections, the more they can be expected to vary with students' background and educational characteristics, the more they can shift over time, and the more comparisons of program effects become untrustworthy.

CONCLUSION

It is a critical time for jurisprudential studies in America. It is a time for self-reflection and reevaluation of methodological and theoretical legacies in the law. At stake is not only the status of modern jurisprudence, but also the validity of the Rule of Law itself. In the current era of academic diversity and disagreement, the time has come to seriously consider the transformative changes now unfolding in American legal thought. The challenge for the next century will certainly involve new ways of understanding how the legal system can preserve the authority of the Rule of Law while responding to the different perspectives and interests of multicultural communities. It is without a doubt an anxious and exciting time for jurisprudence.

> What was once understood as the mainstream of modern view has broken into a diverse body of jurisprudential theories and perspectives. . . . No matter how troubling it may be, the landscape of the postmodern now surrounds us. It simultaneously delimits us and opens our horizons. It's our problem and our hope. (Minda, 1995:256-257)

Ironist critics of educational assessment reject the modernist notion that the "truth" about a student's understanding or a program's effect lies but a simple step away from our ken, to be spanned by observations with standard, context-free measuring instruments and unambiguous statistical analysis of the results. But to further reject any use of these models and information-gathering tools just because they arose under the discarded epistemology is to forgo decades of experience about some ways to structure and communicate observations about students' learning. Educators fear that wholesale abandonment of familiar assessment methodology strips away tools that help them address these facets of their task. Believing these ways of structuring discourse hold *no* value is as wrong as believing that they *alone* hold value. I hear parents and teachers say that we "should not throw the baby out with the bath water." But how to tell which is which?

My answer (a neopragmatic postmodernist answer, as it turns out) is this: Models, principles, and conceptual frameworks are practicable tools—not for discovering a singular truth but for structuring our discourse about students, so that we may better support their learning, and for learning about expected and unexpected outcomes of our efforts, so that we may continually improve them. Understandings of students' learning and programs' effects are enriched by multiple perspectives and diverse sources of evidence, some new or previously neglected, but others with familiar (albeit reconceived) forms. Postmodern architects play with ironies in design, advancing alternative sensibilities and forgotten voices—but they had better design buildings that are livable and safe. Fundamental constraints and fundamental responsibilities persist. And as long as we in education purport to help other people's children learn, at other people's expense, we bear the duty of gaining and using as broad an understanding as we can to guide our actions and of conveying our reasoning and results as clearly as we can to those to whom we are responsible.

ACKNOWLEDGMENTS

I am indebted to Bob Linn and Barbara Storms for their thoughtful comments on an earlier version.

REFERENCES

Birenbaum, M., and K.K. Tatsuoka. (1983). The effect of a scoring system based on the algorithm underlying the students' response patterns on the dimensionality of achievement test data of the problem solving type. *Journal of Educational Measurement* 20:17-26.

Birnbaum, L. (1991). Rigor mortis: A response to Nilsson's "Logic and artificial intelligence." *Artificial Intelligence* 47:57-77.

Bock, R.D., and M. Aitkin (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM-algorithm. *Psychometrika* 46:443-459. Chi, M.T.H., P.

Feltovich, and R. Glaser (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science* 5:121-152.

Cronbach, L.J., and L. Furby (1970). How should we measure “change”—Or should we? *Psychological Bulletin* 74:68-80.

Cronbach, L.J., G.C. Gleser, H. Nanda, and N. Rajaratnam (1972).*The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. New York: Wiley.

Delandshere, G., and A.R. Petrosky (1994). Capturing teachers’ knowledge: Performance assessment (a) and poststructural epistemology, (b) from a post-structuralist perspective, (c) and post-structuralism, (d) one of the above. *Educational Researcher* 23(5):11-18.

Frederiksen, J.R., and K. Sheingold (1994). *Linking Assessment with Reform: Technologies that Support Conversations About Student Work Princeton*, NJ: Center for Performance Assessment, Educational Testing Service.

French, J.W. (1965). The relationship of problem-solving styles to the factor composition of tests. *Educational and Psychological Measurement* 25:9-28.

Greeno, J.G. (1989). A perspective on thinking. *American Psychologist* 44:134-141.

Gulliksen, H. (1961). Measurement of learning and mental abilities. *Psychometrika*26:93-107.

Jöreskog, K.G., and D. Sörbom (1979). *Advances in Factor Analysis and Structural Equation Models*. Cambridge, MA: Abt Books.

Lewis, C. (1986). Test theory and Psychometrika: The past twenty-five years. *Psychometrika*51:11-22.

Lindquist, E.F. (1951). Preliminary considerations in objective test construction. Pp. 119-185 in *Educational Measurement*, E.F. Lindquist, ed. Washington, DC: American Council on Education.

Linn, R.L. (1993) Educational assessment: Expanded expectations and challenges. *Educational Evaluation and Policy Analysis* 15:1-16.

Lord, F.M. (1952). A theory of test scores. *Psychometrika Monograph* 17 (4, Part 2): 1-80.

Messick, S. (1989). Validity. Pp. 13-103 in *Educational Measurement*, 3rd ed, R.L. Linn, ed. New York: American Council on Education/Macmillan.

Miller, M.D., and R.L. Linn (1988). Invariance of item characteristic functions with variations in instructional coverage. *Journal of Educational Measurement* 25:205-219.

Minda, G. (1995). *Postmodern Legal Movements.* New York: New York University Press.

Mislevy, R.J. (1993). *Linking Educational Assessments: Concepts, Issues, Methods, and Prospects*. Princeton, NJ: Educational Testing Service.

Myford, C.M., and R.J. Mislevy (1995). *Monitoring and Improving a Portfolio Assessment System*. Princeton, NJ: Center for Performance Assessment, Educational Testing Service.

Moss, P. (1994). Can there be validity without reliability? *Educational Researcher* 23(2):5-12.
(1996) Enlarging the dialogue in educational measurement: Voices from interpretive research traditions. *Educational Researcher* 25(1):20-28.
Nilsson, N.J.
(1991) Logic and artificial intelligence. *Artificial Intelligence* 47:31-56.
Pelavin Associates
(1994) *Comparing Adult Education Tests: A Meeting of Experts.*
Washington, DC: Pelavin Associates.

Porter, A.C. (1991) Assessing national goals: Some measurement dilemmas. Pp. 21-42 in The Assessment of National Goals. *Proceedings of the 1990 ETS Invitational Conference*, T. Wardell, ed. Princeton, NJ: Educational Testing Service.

Resnick, L. (1994). Performance puzzles. *American Journal of Education* 102(4): 511-526.

Rumelhart, D.A. (1980) Schemata: The building blocks of cognition. Pp. 33-58 in *Theoretical Issues in Reading Comprehension*, R. Spiro, B. Bruce, and W. Brewer, eds. Hillsdale, NJ: Lawrence Erlbaum Associates.

Shavelson, R.J., G.P. Baxter, and J. Pine (1992). Performance assessments: Political rhetoric and measurement reality. *Educational Researcher* 21(4):22-27.

Snow, R.E., and D.F. Lohman (1989). Implications of cognitive psychology for educational measurement. Pp. 263-331 in *Educational Measurement*, 3rd ed, R.L. Linn, ed. New York: American Council on Education/Macmillan.

Spearman, C. (1927). *The Abilities of Man: Their Nature and Measurement*. New York: Macmillan.

Thurstone, L.L. (1947). *Multiple-Factor Analysis*. Chicago: University of Chicago Press.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan* 70:703-713.

6. TECHNOLOGICAL IMPLICATIONS FOR ASSESSMENT ECOSYSTEMS

John Behrens and Kristen DiCerbo

Consider:

> "The role of the problem-posing educator is to create, together with the students, the conditions under which knowledge at the level of the doxa is superseded by true knowledge at the level of the logos. Whereas banking education anesthetizes and inhibits creative power, problem-posing education involves a constant unveiling of reality. The former attempts to maintain the submersion of consciousness; the latter strives for the emergence of consciousness and critical intervention in reality."
>
> Paulo Freire: Pedagogy of the Oppressed, p. 68.

We start with this quote from Freire (1993) because it reflects the consternation many people currently have with commonly available instruction and assessment, while pointing the way toward what new genres for instruction, education and assessment might look like. Typically, we are not concerned with what individuals know and can do in situations that are uniquely created for assessment itself, but rather are concerned with their "emerging consciousness and critical intervention in reality." In this paper we discuss the intersection of assessment theory and evolving practices and conceptualization in the use of digital technologies to understand and advance learning, instruction and assessment.

A core motivation of this work is a focus on directly improving student learning by giving students and instructors increasingly detailed feedback regarding student knowledge, skills, and attributes, in contexts that reflect those in which they will need to apply the information outside the classroom. Key to this is developing holistic understandings of activity as a core input to assessment practice and its interplay with data and the use of data. The persistence of electronic data, its combination, and the variety of its origins provide a new opportunity to move conceptualizations of assessment from discrete disconnected testing events to a larger lens of activity, data collection, and assessment inference ecosystems.

Why the Digital Revolution is Different

Since Kuhn (1962), it is common to see paradigm shifts and scientific revolutions in every evolving concept. It would be easy to think the current technological shifts we are seeing are simple incremental progressions in societal advancement as we have seen in the past with the invention of the steam engine, telephone, or radio. We believe, however, that the current changes we are experiencing are qualitatively different because of the nature of digital technology. First, digital tools allow the extension of human ability by providing symbol manipulation tools that function at the core of human meaning and activity. Second, digital devices can have hardware and software aspects that collect, store and transmit data ubiquitously and unobtrusively. This opens new discussions regarding the nature of data and its role in human self-awareness. Third, digital technologies not only provide interaction with the physical world, as earlier era machines did, but their flexible symbol manipulation allows mapping back into key representations at the core of human communication, including: visual display, auditory communication, and even haptic recording. We discuss each of these privileged attributes in turn.

Extension of human ability through computation

Modern digital computers are noted for their speed and functioning as general symbol processors. Digital computing allows the translation of physical aspects of the world (e.g., pages of printed text) into electronic representation that can be acted upon by general computational machinery. This allows the application of computer programs of logic to search, sort, and combine information in ways that create new rules that act as additional intelligence and insight. The memory capabilities of computers allow the storage and organization of information that supplements the memory capabilities of our own mental capacity. The ability to automate computation allows the repeated application of simple steps to solve complex problems through brute force repetition of evaluations or simulation of processes. For example, resurgence in addressing many previously intractable statistical problems is being led with the use of methods for simulating complex statistical distributions (Brooks et al 2011). Daily activities are transformed through such ubiquitous simulations as the word processor, which simulates the appearance of the physical page, though many “documents” will never achieve a physical presence.

An understanding of the role of digital computers as extensions of human ability can, in some ways, be best illustrated by observing the roles in which humans are being replaced by computers. Behrens, Mislevy, DiCerbo and Levy (2012) note the disappearance of the “typing pool” in modern corporate America is a result of technological change that have led to changes in roles and expectations regarding the production of knowledge and its documentation. In some cases, computers replace vocations where the primary unique function of the individual was unique information or information synthesis (e.g. travel agents), while in other cases the replacement is a function of speed, or automation of simple tasks (e.g., parking lot attendants).

Perhaps most notable in the computational aspects of the digital revolution is that not only do computers often replicate what humans do well (though not universally), but also that they serve as tools that allow us to accomplish things we had not considered in a pre-digital era. Consider, for example, the dramatic advances in biology brought about by the sequencing of the human genome (Kent et al., 2002) that depends largely on the computer-based statistical analysis of biological materials, that are handled by automated physical processes overseen by computerized devices. In such a scenario, entire new understandings of aspects of human nature are driven largely by methods nearly completely dependent on digital computing machines.

Data collection, recording and transmission

When the general symbol manipulation ability of digital computers is combined with sensors for automated input, with networks for the transfer of information to remote storage and computing, data collection moves toward becoming ubiquitous and unobtrusive. This is a dramatic shift from previous eras in which physical collection of data was often obtrusive and likely to cause reactive effects when inserted into daily activity. For example, Krathwohl (2009) enumerates the variety of methods that needed to be considered in introducing bulky video recording devices into classroom settings, and citing the ingenious efforts of Kounin (1970) to avoid reactivity in the classroom. These efforts stand in contrast to the current commonplace use of unobtrusive cell phones and automatic uploading to social media sites, which have changed social norms for the collection and use of data in daily, as well as professional life.

Of particular interest are the rapid changes in the availability of mobile devices, most notably the cell phone. These extremely compact personal computing devices allow the ongoing collection of data through user input that can be combined with access to database information or historical data on an individual. For example, the spatial positioning and accelerometer information of cell phones can be combined with databases about traffic patterns, to direct drivers to unclogged commuter routes. This combination of access to historical records and the collection of ongoing streams of data has led to the notion of “data exhaust” (Olsen, 2000), which consists of the digital data discharged by the use of digital devices through the course of a day or lifetime. Indeed, digital devices of all kinds are typically enabled to collect data in ubiquitous and unobtrusive ways. Keyboards on computers, for example, are instrumented to collected data regarding the pushing down of the keys. When combined with the ongoing time-stamp available in most systems, a picture can be created of the typing pattern of an individual, and thereby a digital pattern of an individual may be created unobtrusively for use in person identification and authorization (Peacock, Ke, & Wilkerson, 2004).

The emergence of these technologies in everyday life changes the location and cost of data collection, thereby changing our individual and social relationship with data. As data about learning and human activity becomes increasingly ubiquitous and inexpensive, we change the ways in which data need to be collected. DiCerbo and Behrens (2012), for example, argued that the nature of testing will change, as the need for isolated testing occasions fades out in favor of ongoing and unobtrusive data capture.

Representation

A third transformational aspect of digital technologies is their ability to translate data of different types into various representational forms. For example, the beautiful “pictures” of space communicated from the Hubble telescope are not really pictures but rather artistic renderings of raw data that, in some instances, have no human visual analog because the data are collected in wavelengths imperceptible to the human eye. The data collected from such devices are actually sensor readings that are transformed and modified, to be communicated as interpretable visual analogies.

Other transformations are commonplace as well. The striking of computer keys in a video game may communicate the need for movement of characters around a virtual space, while the pressing of those keys during the running of a word processing application leads to the appearance of shapes on the screens corresponding to letters of an alphabet. There is no inherent isomorphic relationship between hitting the keys and observing changes in the word processing display, except insofar as the computer program has been designed to mimic the conventions of previous devices (e.g., the typewriter).

This fluidity of representation allows the digital capture of video images in one part of the earth, the transmission to other devices, and the re-display in other places. In such situations the visual impression of the observer at the input would match that of the visual impression of the observer upon re-display. When combined with centralized storage and social media-based contributions, we see the storage and accumulation of large libraries of images in such applications as Flickr. When these visual images are combined with image recognition software such as Google Goggles, we complete an end-to-end loop of digital activity around images, which far exceeds the flexibility and scope previously available for mass photography and image use. Indeed, one may often observe digital natives (Prensky, 2001) using cameras to make images of specific pieces of information such as notes on a board or information for short-term use. This illustrates the use of the device as a short-term recording tool to aid in short-term memory, thereby augmenting the original uses which were centered on the pre-digital concepts of the camera as a device for artistic expression and personal memory development. Changes in cost and flexibility have led to changes in use and conceptualization.

Resulting Shifts in Assessment

The combination of these digital properties opens new possibilities for understanding, exploring, simulating, and recording activity in the world, and this thereby opens possibilities for rethinking assessment and learning activities. We see these opportunities in three distinct areas that require an unpacking and reconceptualization of traditional notions of assessment in light of the new digital situation. These opportunities can be summarized as shifting from:

1. an item paradigm to an activity paradigm
2. an individual paradigm to a social paradigm
3. assessment isolation to educational unification

Shifting from an item paradigm to an activity paradigm

We will use the term “item paradigm” to represent our impression of common assumptions that assessment designers and policymakers (and ourselves at different times) have, or did have, about the fundamental aspects of assessment. “Item” is a vernacular term that refers to a discrete piece of assessment interaction and data collection, typically in the form of a question, or combination of a question and possible answers (as we see in the multiple choice format). Most items on standardized tests fall into a class of formats called “fixed-response items” because the set of possible responses are fixed by the assessment designer as the “options” from which one may choose.

DiCerbo and Behrens (2012) argue that the multiple choice format was designed to address the most difficult part of the assessment delivery process, which is alternately called response processing or evidence identification (Almond, Mislevy & Steinberg, 2002). They suggest that to simplify the process, assessment designers started their conceptualization by simplifying the scoring process around fixed-response automation (reading of hand-marked bubbles), and simplified the presentation process in front of it, matching the psychometric models after it in the overall design process. This was necessary when computing capabilities were limited. The capabilities required to search a complex work product, extract particular features, and apply scoring rules in an automated fashion were beyond the scope of technology at the time.

With the digital revolution, it is now conceivable that we can extract evidence from a variety of work products resulting from a range of activity, including writing essays (Dikli, 2006), configuring computers (Rupp et al., 2012), and diagnosing patients (Margolis & Clauser, 2006). Williamson (2012) provides a conceptual base for this process, detailing how particular pieces of evidence can be extracted from a complex work product, even in cases where there are unconstrained result possibilities. Although Williamson's chapter title refers to scoring of items, in fact all of this advancement in automated scoring allows us to stop thinking at an item level. We can now write activities that require complex performances parallel to those learners would complete in the real-world. When we begin assessment design, we can begin with a consideration, not of what we want to report, but what real world activities we want students to be able to perform following instruction. Table 1 presents a contrast between the item paradigm and the activity paradigm.

Table 1. *Differences between the item paradigm and the activity paradigm*.

	Item Paradigm	Activity Paradigm
Problem Formulation	Items pose questions	Activities request action
Output	Items have answers	Activities have features
Interpretation	Items indicate correctness	Activities provide attributes
Information	Items provide focused information	Activities provide multi-dimensional information

Attributes, not correctness

A common side effect of an item-centric view of assessment is that the assessment may be conceptualized and designed in terms of the matching algorithm of scoring as the primary conceptual lever in the assessment process. Two dangers may occur from this. The first danger is that the test is conceptualized in terms of overall goodness of response based on average correctness. A common pattern for assessment design is, (1) identify a domain, (2) sample ideas or activities from the domain, (3) make questions about those ideas or activities, (4) score them correct or incorrect. The difficulty is that this pattern can be undertaken with very little specification of the domain or discussion of the precise type of evidence or inference desired. The correctness paradigm can drive the construction with very little acknowledgment of the relationship between the role of individual items and the overall inference being sought. It begs the question "if the item is measuring correctness, I need to know 'correctness' of what."

A second concern with the correctness paradigm is that it fails to account for the many situations in which we are interested in assessing specific attributes of an individual and not only overall goodness. We may want to identify specific strategies used, the presence of a specific belief or action, or place someone in a cluster of similar individuals not because of correctness, but because of work features that are relevant to diagnosis or instruction. This is a generalized feature-centric view of response scoring. This is an important concept as work products become more ubiquitous and available for diagnostic purposes.

Technology allows us to expand our thinking about evidence. Digital systems allow us to capture stream or trace data from students' interactions. This data has the potential to provide insight into the processes that students use to arrive at the final product (traditionally the only graded portion). These data are log files of student action sequences that offer the possibility of thinking about the features of a performance and the evidence that they provide. For example, Rupp et al., (2012) analyzed log files consisting of time-stamped commands that students entered to configure computer networking devices on a simulation-based assessment. They identified features including the number of commands used, the total time taken, and the number of times in the log that students switched between devices, as evidence that could be combined into a measure of efficiency. Note that none of these features was scored "correct" or "incorrect," and their combination allows us to make an inference about the students, apart from the overall correctness of their performance.

Similarly, Shute, Ventura, Bauer, & Zapata-Rivera (2009) leveraged game data to make inferences about 21st-century skills. They were interested in making inferences about students' problem-solving ability, which they modeled as having two indicators: efficiency and novelty. Actions in the game were then identified and scored to provide evidence about efficiency and novelty. For instance, if a student came to a river in the game and dove in to swim across it, the system would recognize this as a common (not novel) action and automatically score it accordingly (e.g., low on novelty). Another person who came to the same river but chose to use a spell to freeze the river and slide across would be evidencing more novel (and perhaps more efficient) actions, and the model would be updated accordingly. Again, the emphasis is on identifying features that provide evidence for a particular construct, not on "correct" or "incorrect," and technology allows us to capture the student actions in the game to expand our thinking about what constitutes evidence.

Multi-dimensional information

The correctness paradigm described above works in many assessments because an item is correct for a specific attribute in a specific dimension or scale. This is a correct answer for a question about "fill in scale name here." We know, however, that the dimensionality of interpretation of a task is related to the structure of the task as well as the overall conceptualization of the task. Question-based tasks that ask a question in written language necessarily require competencies in reading (or hearing) of words of the language, the interpretation of the sentence structure and the response to the request based on relevant domain knowledge. However, in many contexts the linguistic aspects of the task are considered "construct irrelevant" or "noise" around the single construct related to the domain being measured.

This is problematic insofar as it requires the assessment designer to increasingly decontextualize tasks so that a "pure" item is written that allows for inference most specifically to the relevant construct. Indeed this is a requirement in a one-dimensional "correctness paradigm" delivery system. An alternate approach would be to conceptualize an activity as multi-dimensional from the start and work to understand the data as meaningful that was previously thrown away as "construct irrelevant."

Technology does not just assist in presentation of activities and evidence identification, but also in evidence accumulation. Evidence accumulation refers to the synthesis of the evidence generated from activities; it relies heavily on the ability of statistical models to combine disparate information to make inferences about students. Bayesian networks (Jensen, 1996; Pearl, 1988) represent a flexible approach to latent variable modeling of complex activities (Almond, DiBello, Moulder, & Zapata-Rivera, 2007; Levy &Mislevy, 2004). They represent one example of the types of statistical techniques that might be applied to the problem of accumulating evidence from multidimensional activities.

In many complex activities, a variety of observations can be made from the activity, which may relate to different skills that we wish to make inferences about. However, each of these observations was also made from the same general activity, and may therefore share some commonality from being part of that activity. Williamson, Almond, Mislevy, and Levy (2006) walk through an example like this using Bayesian Networks. They were estimating one overall construct of networking disciplinary knowledge with two subskills: troubleshooting and network modeling. From one activity there were four observables related to troubleshooting and one related to network modeling. Since they all came from the same activity, however, an intermediate context variable was created to account for the relationship among the five observables. Once the probability values throughout the table are set, evidence gathered from one observable propagates through the network to update the probability that a student has mastered a particular skill and also their probabilities of succeeding at other observables.

Link activity, data and inferences

The previous discussion links features and evidence, but a larger concern is linking activity to data to inferences. Traditional assessment has a clear start and stop (often traditionally marked with the phrase “pencils down”). This results in a clearly defined experience from which to extract data. However, this experience is neither contextualized nor representative of the real world experiences about which we would like to make inferences. However, without this rigid, defined experience, how does one link experiences, data, and inference?

We have found the principles of Evidence Centered Design (ECD), (Mislevy, Steinberg & Almond, 2003) and it’s logical bases (Mislevy, 1994) extremely useful in our work in automated classroom assessment (Behrens, Mislevy, Bauer, Williamson, & Levy, 2004). First, ECD emphasizes the logical form of the assessment argument and suggests careful consideration of the train of reasoning in assessment design and development. When combined with a new implementation and systems approach to understanding educational assessment, this led us to think in a forward manner regarding what computer technologies could offer for the conceptualization and delivery of assessment activities (Behrens, Mislevy, DiCerbo& Levy, 2012). Second, while many discussions of ECD emphasize this important evidentiary aspect of assessment (and their operational consequences) we equally found benefit from ECD’s detailing the elements of assessment delivery in a way that is sufficiently abstract as to include human language (Mislevy, Steinberg & Almond, 2002), a broad range of classroom activities (Mislevy, Behrens, DiCerbo& Levy, in press), games (Behrens, Frezzo, Mislevy, Kroopnick, & Wise 2008) and simulation in general (Frezzo, Behrens, Mislevy, West, &DiCerbo, 2009).

The ECD assessment framework provides us with a guide about how to better make a link between experiences, data and inferences (Mislevy, Steinberg & Almond, 2002). Assessment design activity can be thought of as a series of three questions: "What are we measuring?", "How do we want to organize the world to collect evidence for the measurement?", and "What are the conceptual linkages between observable evidence and abstract inferences?" In ECD, the Conceptual Assessment Framework (CAF) expresses answers in terms of models about the student, tasks, and evidence. We first consider what inferences to make about students; we then consider behaviors we could observe that would tell us about those things and activities that would allow us to observe them; and finally we determine how to identify important elements of the experience and combine them together.

It is important to note that in some new assessment delivery tools, assessment designers have broad and integrated claims they wish to seek, but they dissect the scene to elicit highly structured subsections that limit the range of possible actions open to the learner. We think this reflects a pre-digital conceptualization of the problem as one of constraining the presentation interactivity to align with simplified scoring. However, we believe by using a flexible scoring system behind an open-ended activity presentation system, the user flow and evidence identification goals of an assessment activity can both be met. Technology shifts the burden of inference from presentation to evidence identification. That is, in the world of multiple-choice exams, the work is in creating useful items in the constrained space. Identifying evidence from these items is simple. With technology, the space for presentation of tasks is much larger and the difficulty is shifted to how to identify appropriate evidence from the bounty of responses in that environment.

Shifting from an individual paradigm to a social paradigm

The advent of the Internet has brought about a revolution in social communication that has reinforced the concept of the social nature of human activity. What we have long known from our emotional experience, we now see in the data of our daily interactions: emails, media posts, tweets, and collaborative work spaces such as wikis. Collaboration in the digital world is helping to solve difficult problems. For example, the Foldit game (Khatib et al., 2011) is an effort to solve protein structure problems through game play. Players attempt to discover the ways that real proteins fold, and have in fact uncovered structures of actual proteins that have eluded scientists (and computers), by building on the results of others and competing to get the best optimization scores. However, very few assessments allow for collaboration; the prototypical test situation consists of one examinee seated at a desk being told to "keep your eyes on your own paper."

Many digital environments are specifically designed to be collaborative. Commercial online massively multiplayer games like World of Warcraft rely on collaboration among game players as a major driver of action. In the world of digital environments in education, River City, a multiuser virtual environment, asks teams of middle-school students to collaboratively solve a simulated 19th-century city's problems with illness (Dede, Nelson, Ketelhut, Clarke & Bowman, 2004). Players use a group-chat feature to communicate with each other about their findings. They also communicate with characters in the game via chat. Although no published work was found on this, these chat logs could be mined for evidence of collaboration knowledge, skills, and attributes using natural language-processing techniques to assist in the eventual automation of this scoring.

Shaffer and his colleagues conduct research in the context of an epistemic game called Urban Science that mimics the professional practicum experiences of urban planners (Rupp, Gushta, Mislevy, & Shaffer, 2010). Rupp et al., (2010) make use of transcripts of interactions between individual learners and between learners and mentors to identify evidence of players' skills, knowledge, identity, values, and understanding of evidence in planning (epistemology). Bagley & Shaffer (2010) used both discourse and network analysis to analyze transcripts of interactions between players and mentors, comparing a virtual-chat condition to a face-to-face chat and found that discourse, outcomes, and engagement levels were similar between the two groups. These studies suggest both methodologies for working with chat logs and confirm that these artifacts of digital interactions can help us assess 21st-century skills.

Consider the ecosystem

If we take building a digital environment seriously, we are led to deeply consider the purpose of our assessment activity and how different goals for feedback may lead to different and multiple forms of interaction with the learners. For example, in a traditional classroom the teacher and student both have access to data regarding performance on homework assignments, quizzes, in-class practice, mid-term and final exams, and conclusions from formal and informal dialog. Behrens et al (2005) for example, described six different types of assessment activities that were undertaken in the Networking Academies at that time (Quiz, Module Exam, Practice Final, Final Exam, Voucher Exam, Practice Certification Exam) in terms of six different feature types (organizational purpose, instructional purpose, grain-size of feedback, grain-size of claims and tasks, level of task complexity, level of task security). Variation in assessment activity goal led to different patterns of design features afforded those activities. For example, quizzes were designed to provide small grain size feedback with low security while final exams are designed to assess higher grain claims with larger grain size feedback and higher security.

Among the "Seven C's of Comprehensive Assessment" discussed in that paper (Claims, Collaboration, Complexity, Contextualization, Computing, Communication, Coordination), the final point on Coordination emphasizes this notion of an information and experience ecosystem. Behrens et al., illustrated this by linking and equating the practice certification exam delivered in the schools with the professional certification exam given to examinees under third-party certification conditions. This strengthened the validity of inferences individuals and organizations would make about future professional certification performance based on school-based performance. Even when assessment activities are not mathematically linked and equated, we believe the conceptualization of the total learning lifecycle needs to be considered.

While in some ways such an enumeration seems commonsensical, it is a departure from many assessment formulations that use "the test" as the unit of analysis and assume logical independence between assessment activities. In such an approach, purposes aligned with specific assessment goals can be missed and assessment activities (items or tests) may be developed with one purpose in mind, which are then inappropriately applied in other contexts. One is reminded of the relativity of validity for the purposes to which a given assessment activity is oriented. The ecosystem approach attempts to understand the broad range of needs and tailor individual assessment activities to the specific needs, but also create a design across assessment activities and events to ensure all needs are met appropriately.

Educational assessment focuses on activities of humans which must be understood in the human context of social and physical environments with goals, norms, etc. Activity theory (Engeström, 1987) teaches us that because assessment is a human event in a social context, we need to have a framework for understanding what we are paying attention to and what we are not. Often, educational assessment appears to ignore many aspects of the assessment activity without consideration. In fact, this is essentially ignoring important dimensions of variation.

Activity theory provides a framework against which to consider a broad range of dimensions that affect human activity. It outlines some of the more commonly thought-of aspects of assessment including: the subject, the tools, the object, and the outcome. In a narrow view of assessment this would translate into the learner, the test, determining whether the student can multiply two numbers, and finishing the test and obtaining a final score. A richer view would suggest that in the ecosystem the subjects are the student, classmates, and the teacher; the tools might include web resources, books, characters in a game, calculators, and manipulatives; the object is solving a math problem, and the outcome includes both achievement and motivational measures.

In addition, activity theory includes a layer of less commonly thought of pieces of the ecosystem consisting of rules, community, and division of labor. Rules include the norms around the activity, such as whether collaboration with a peer is permissible. Community refers to the group of people engaged in a practice, so it might be a classroom or an online discussion board for a particular game. The division of labor defines who does what in the activity, including whether work is distributed at all. Consideration of all this context can lead to the uncovering of tensions (Frezzo, Behrens, & Mislevy, 2010), such as whether there is familiarity with the tools to be used, or how much choice a student has within and between activities.

In traditional assessment, much of the context is already in place, so it is easy for it to remain unexplored. In the creation of simulations and digital environments, each of these elements requires consideration. Interactions with characters in a game can be scripted, tools available at any particular time can be defined, rules for interaction are outlined, and the community around the experience is built. When a team within the Cisco Networking Academy created a computer networking game called Aspire, they used the simulation tool Packet Tracer as the game engine. Packet Tracer is embedded throughout the curricula of the Networking Academy, so by using this tool, it was ensured that students would have familiarity with the interface and interactions. When the team then sought to expand the game beyond the Networking Academy, they realized it would then be used by people unfamiliar with Packet Tracer, so a new initial level was added to familiarize players with the tool. Thus, the usability of the same tool changed, based on the context in which it was to be used.

The design of the Quest Atlantis game involved much of this thinking about context. Barab, Dodge, Thomas, Jackson, and Tuzun (2007) write, “Instead of simply building an artifact to help individuals accomplish a particular task, or to meet a specific standard, the focus of critical design work is to develop sociotechnical structures that facilitate individuals in critiquing and improving themselves and the societies in which they function…" (p. 264). They describe how, although they could have focused the Quest Atlantis virtual environment solely on particular science standards about erosion, they became concerned with highlighting attitudes toward environmental awareness and social responsibility. For example, one issue in game design is how to develop levels which represent expertise and usually create new opportunities and resources for interaction. Barab et al., decided to make a structure connected to social commitments, creating a story about collecting pieces of crystal, with each representing a social commitment the designers wanted to enforce, like environmental awareness. They instilled in the community around the game a value of these commitments through the design of the ecosystem. This larger perspective of the ecosystem communicated by the interactions we design is often completely ignored, but critical design of digital experiences can bring it to the forefront.

Build Inviting and Non-coercive Environments for Data Collection

An argument has been put forward that standardized tests coerce teachers and, by extension, students into the coverage of particular topics in the curriculum (Noddings, 2001). Rowland (2001) argues that a culture of compliance has overtaken a culture that promotes intellectual struggle with difficult concepts. He writes, "…completed tick boxes of generic skills undermines the enthusiasm and passion of intellectual work" (para. 4). According to this way of thinking, much of current assessment practice focuses on narrow discrete skills and encourages students to march lockstep through them without questioning and inquiry in order to arrive at the single correct answer. In the quote by Freire at the beginning of the paper, this is "banking education."

Others may quibble at the extremeness of this viewpoint, but it is difficult to argue that most current assessments invite students in, or that students would choose to interact with tests on their own, as currently construed. Ideally, systems should honor the student by creating environments that engage students and invite them to participate, rather than coerce them into participation. We seek to create active and engaging learning and assessment environments, both because it honors the students' desire to be self-defining and intervening in reality, and because such an environment can be most aligned with the kinds of real activities about which we most want to make inferences. What pleases and engages the student and brings them most fully to the activity is also what we most naturally want to understand, encourage and describe.

Judging from the usage statistics suggesting 97% of teens play computer, web, portable, or console games (Lenhart, Kahne, Middaugh, Macgill, Evans, & Vitak, 2008), digital experiences can be very engaging and inviting. Creating an artificial environment to allow action consistent with living "in the wild" is an important aspect of modern measurement of 21st-century skills (Behrens, Mislevy, DiCerbo& Levy, 2012). An open, simulated environment allows for a full range of knowledge, skills, and attributes over time including: recognition of cues regarding problem situations, formulation of problems, recovery from mistakes, understanding and responding to environmental feedback, and other complex emotional and information processing skills. For example, networking professionals may describe part of their job satisfaction in terms of "Eureka!" Moments when they fix a network; students have been observed experiencing just such moments when simulation based learning environments are used (Frezzo, 2009). The ability to digitally create environments for authentic experiences gets to the heart of engaging students in the process.

Researchers have found that the features likely to create immersion include elements of challenge, control, and fantasy (Lepper & Malone, 1987). Engagement or "flow" is most likely to occur when the cognitive challenge of a problem closely matches the student's knowledge and skills (Csikszentmihalyi, 1990; Gee, 2003). Game designers do an excellent job of making a match between players' skill and the challenge level, and digital environments in general facilitate making this match with adaptive systems. Digital environments also allow for a balancing of the rules and constraints of the activity versus the agency or freedom of the participants (Bartle, 2005; Chin, Dukes, & Gamson, 2009). Finally, digital experiences allow for the creation of interesting fantasy environments (for example, River City (Ketelhut, Dede, Clarke, Nelson & Bowman, 2007) and the various locations in Quest Atlantis (Barab et al., 2009).

Some may argue that this complex environment merely serves to introduce construct-irrelevant variance. The notion of construct irrelevant variance needs to be deconstructed. It essentially means variation in performance because of task demands that were undesired or unanticipated. The relevant issue is not construct relevance, but rather inference relevance. Some changes in the activity or activity structure may indeed be inferentially irrelevant. Adding the feature does not affect the inference being made. At other times, we may add features that are construct irrelevant but inferentially relevant. For example, if we add an aspect of the simulation that degrades performance interpretation in the core areas of the activity, that would be a feature to remove. The key here is that not all variance is bad, but we need to be aware of how it affects inference.

Shifting from Assessment Isolation to Educational Unification

Tests are artificial situations, developed to elicit specific actions from learners, to give them the opportunity to demonstrate their competencies. As such, a test is an assessment. Assessment, as a general class of action, however, may or may not include testing. The careful eye of a teacher undertakes student assessment consistently throughout the day perhaps on many dimensions that are never formally accounted for: student is tired, student is hungry, student may not be living at home, student forgot homework, student at high risk for drop out and so forth. These types of inference may be combined together by an ongoing series of informal observation or reports from others. The student may never come to take a test, while the teacher nevertheless creates a mental model of the student and updates it over the course of the semester. Even in the realm of achievement, teachers base their models on interactions, observations, informal questioning, and classroom work products before a test is ever given. Tests are assessments, but assessments do not require tests.

Technology has the potential to further break down the barrier between assessment and instruction. When students interact with a digital environment during an "instructional" activity, and information from that interaction is captured and used to update models of the students' proficiency, is that instruction or assessment? Shute, Hansen, & Almond (2008) demonstrated that elaborated feedback in a diagnostic assessment system did not impact the validity or reliability of the assessment, but did result in greater learning of content. They termed this an "assessment for learning system." In many game and simulation environments, the environment is both a learning and assessment environment in which the system is naturally instrumented and the play is not interrupted for assessment purposes.

The Whole World is the Classroom

In the 20th-century, we created artificial environments and sets of tasks that increase (or force) the likelihood of being able to observe a set of activities and sample them for inferential purposes. We call these tests. They require the interruption of normal instruction and are sometimes called "disruptive" or "drop in from the sky" testing (Hunt & Pellegrino, 2002). Technological limitations on interacting with students, providing experience and capturing relevant data, especially in the classroom, often lead to dramatic truncation in the goals and aspirations of assessment designers. Sometimes the truncation makes its way back to the original conceptual frame of the problem so that the assessment designers do not even consider the target activity to which we wish to infer, but stop at distal and common formulations that may have severe inferential weaknesses for claims of generalization or transfer. To counter this, we encourage specification of the claims we want to make about activity “in the wild.” That is, we try to understand the claims as contextualized in practice outside of the assessment environment. Here again, most practitioners would argue that all good assessment conceptualizations should do this, but likewise many experienced practitioners will confide that one’s ability to think beyond the constraints of their authoring environment is often quite difficult.

In the 21st-century, activities, records of activities, data extracted from patterns of those records, and the analysis of that data, are all increasingly digital. The day-to-day records of our activities are seamlessly recorded in a growing ocean of digital data: who we talk to (cell phone and Facebook records), where we are (Google Latitude), what we say (gmail and gvoice), the games we play online, what we do with our money (bank records), and where we look online. This emerging reality we refer to as the “Digital Ocean.”

As the activities, and contexts of our activities, become increasingly digital, the need for separate assessment activities should be brought increasingly into question. Further, the physical world and the digital world continue to merge. Salen (2012) describes how even the division between the digital world and the physical world is disappearing. She describes a lesson in the Quest2Learn school on the Work = Force x Distance equation. Some students were able to understand this with observations in a digital environment. Other students, however, were still struggling. The students then had the opportunity to use Wii-like paddles to push virtual objects up virtual inclines with haptic feedback about the amount of effort required to get the object up different inclines. As students became physically aware of the force they were using and the amount of work required, they began to understand the relationships between work, force, and distance. In this case, the students were getting real physical feedback about a virtual activity; the barrier between the two was removed.

The experience above was still in a school environment, but others are working on identification and accumulation of evidence from digital interactions that occur across a range of environments. In the Shute et al., (2009) research on problem solving in games described above, Oblivion, a commercial game not often seen in schools, was the platform used for the investigation. Shute has also done work with World of Goo (Shute & Kim, 2011), while Wainess, Koenig, & Kerr (2011) document how commercial video games contain specific design features that facilitate learning and assessment. Beyond games, we can also consider ways to make use of all the sensors that daily gather information about our actions and states. An extreme example of the probabilities here is offered by Stephen Wolfram (a founder of Wolfram Alpha and Mathmatica) who has analyzed everything from the time of emails sent, number of meetings, and hours spent on the phone daily, based on his archive and logs of activity dating back over 20 years (http://blog.stephenwolfram.com/2012/03/the-personal-analytics-of-my-life/). While he is not focused on making inferences about learning, his is a good illustration of the data that is available from digital sensors solely through our everyday interactions in the digital world.

Conclusions

The digital revolution has brought about sweeping changes in the ways we engage in work, entertain ourselves, and interact with each other. Three main affordances of digital technologies suggest they will create a paradigm shift in assessment: 1. digital tools allow the extension of human ability, 2. digital devices can collect, store and transmit data ubiquitously and unobtrusively, and 3. digital technologies allow mapping back into key representations at the core of human communication. The combination of these digital properties opens new possibilities for understanding, exploring, simulating and recording activity in the world and this thereby opens possibilities for rethinking assessment and learning activities.

The emerging universality of digital tasks and contexts in the home, workplace and educational environments will drive changes in assessment. We can think about natural, integrated activities rather than decontextualized items, connected social people rather than isolated individuals, and the integration of information gathering into the process of teaching and learning, rather than as a separate isolated event. As the digital instrumentation needed for educational assessment increasingly becomes part of our natural educational, occupational and social activity, the need for intrusive assessment practices that conflict with learning activities diminishes.

References

Almond, R. G., DiBello, L. V., Moulder, B., & Zapata-Rivera, J. (2007). Modeling diagnostic assessment with Bayesian networks. *Journal of Educational Measurement*, 44(4), 341-359.

Almond, R. G., Steinberg, L. S., & Mislevy, R. J. (2002). Enhancing the design and delivery of assessment systems: A four-process architecture. *Journal of Technology, Learning, and Assessment*, 5. Retrieved from http://escholarship.bc.edu/ojs/index.php/jtla/article/viewFile/1671/1509

Bagely, E & Shaffer, D.W (2010) Stop Talking and Type: Mentoring in a Virtual and Face-to-Face Environment. *International Journal of Computer-Supported Collaborative Learning*. Retrieved from http://epistemicgames.org/eg/wp-content/uploads/bagely.pdf

Barab, S. A., Dodge, T., Thomas, M. K., Jackson, C., & Tuzun, H. (2007). Our designs and the social agendas they carry. *The Journal of the Learning Sciences, 16*, 263-305.

Barab, S. A., Gresalfi, M., Ingram-Goble, A., Jameson, E., Hickey, D., Akram, S., & Kizer, S. (2009). Transformational play and Virtual worlds: Worked examples from the Quest Atlantis project. *International Journal of Learning and Media*, *1*(2), Retrieved from http://ijlm.net/knowinganddoing/10.1162/ijlm.2009.0023

Bartle, R. (2005). Virtual Worlds: Why People Play. In T. Alexander (Ed.) *Massively Multiplayer Game Development 2*. Hingham, MA: Charles River Media.

Behrens, J. T., Mislevy, R. J., DiCerbo, K. E., & Levy, R. (2012). Evidence centered design for learning and assessment in the digital world. In M. Mayrath, J. Clarke-Midura, & D. H. Robinson (Eds.). *Technology-based assessments for 21st-Century skills: Theoretical and practical implications from modern research* (pp. 13-54). Charlotte, NC: Information Age Publishing.

Behrens, J. T., Mislevy, R. J., Bauer, M., Williamson, D. M., & Levy R. (2004). Introduction to evidence centered design and lessons learned from its application in a global e-learning program. *The International Journal of Testing, 4*, 295–301.

Behrens, J. T., Frezzo, D. C., Mislevy, R. J., Kroopnick, M., & Wise, D. (2008). Structural, Functional, and Semiotic Symmetries in Simulation-Based Games and Assessments. In E. Baker, J. Dickieson, W. Wulfeck, & H. F. O'Neil (Eds.), *Assessment of problem solving using simulations* (pp. 59-80). New York: Earlbaum.

Binkley, M., Erstad, O., Herman, J., Raizen, S., Ripley, M., & Rumble, M. (2010). *Defining 21st-Century Skills*. Retrieved from http://atc21s.org/wp-content/uploads/2011/11/1-Defining-21st-Century-Skills.pdf

Brooks, S., Gelman, A., Jones, G., Meng, X. L., (2011) (Eds.). *Handbook of Markov Chain Monte Carlo Methods*. Boca Raton, FL: Chapman.

Chin, J., Dukes, R., &Gamson, W. (2009). Assessment in simulation and gaming: A review of the last 40 years. *Simulation & Gaming, 40*, 553-568.

Csikszentmihalyi, M. (1990). *Flow: The Psychology of Optimal Experience*. New York: Harper and Row.

Dede, C., Nelson, B., Ketelhut, D. J., Clarke, J., & Bowman, C. (2004). Design-based research strageties for studying situated learning in a multi-user virtual environment. *ICLS Proceedings of the 6th international conference on learning sciences.*

Dikli, S. (2006). An Overview of Automated Scoring of Essays. *Journal of Technology, Learning, and Assessment, 5*(1). Retrieved from http://ejournals.bc.edu/ojs/index.php/jtla/article/view/1640.

Engeström, Y. 1987. *Learning by expanding: An activity theoretical approach to developmental research*. Helsinki: OrientaKonsultit.

Freire, P. (1993). *Pedagogy of the Oppressed*. New York: Bloomsbury.

Frezzo, D. C. (2009). Using activity theory to understand the role of a simulation-based interactive learning environment in a computer networking course (Doctoral dissertation). Retrieved from ProQuest http://gradworks.umi.com/33/74/3374268.html

Frezzo, D.C., Behrens, J.T., &Mislevy, R.J. (2010). Design patterns for learning and assessment: facilitating the introduction of a complex simulation-based learning environment into a community of instructors. *The Journal of Science Education and Technology*. Springer Open Access http://www.springerlink.com/content/566p6g4307405346/

Frezzo, D.C., Behrens, J. T., Mislevy, R. J., West, P., & DiCerbo, K. E. (2009). Psychometric and evidentiary approaches to simulation assessment in Packet Tracer Software. Paper presented at the International Conference on Networking and Services, Valencia, Spain.

Gee, J. P. (2003). *What Video Games Have to Teach Us About Learning and Literacy*. New York: Palgrave/Macmillan.

Hunt, E., & Pellegrino, J.W. (2002). Issues, examples, and challenges in formative assessment. *New Directions for Teaching and Learning, 89*, 73-86.

Jensen, F. V. (1996). *An Introduction to Bayesian Networks.* New York, NY, USA: Springer-Verlag.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., Haussler, D. (2002).The human genome browser at UCSC. *Genome Research, 12,* 996-1006.

Ketelhut, D. J., Dede, C., Clarke, J., Nelson, B., & Bowman, C. (in press). Studying situated learning in a multi-user virtual environment. In E. Baker, J. Dickieson, W. Wulfeck & H. O'Neil (Eds.), *Assessment of problem solving using simulations*. Mahwah, NJ: Lawrence Erlbaum Associates.

Khatib, F., Cooper, S., Tyka, M. D., Xu, K., Makedon, I., Popovic, Z., Baker, D., & Players. (2011). Algorithm discovery by protein folding game players. *Proceedings of the National Academy of Sciences, 108*, 18949-18953.

Kounin, J. (1970). *Discipline and group management in classrooms*. New York: Holt, Rinehart, & Winston.

Krathwohl, D. (2009). *Methods of educational & social science research: An integrated approach* (3rd ed.). Long Grove, IL: Waveland Press.

Kuhn, T., S. (1962). *The structure of scientific revolutions* (1st ed.). Chicago: University of Chicago Press.

Lenhart, A., Kahne, J., Middaugh, E., Macgill, A., Evans, C. &Vitak, J. (2008). Teens, video games, and civics. Washington DC: Pew. Retrieved from: http://www.pewinternet.org/Reports/2008/Teens-Video-Games-and-Civics.aspx

Lepper, M. R., & Malone, T. W. (1987). Intrinsic motivation and instructional effectiveness in computer-based education. In R. E. Snow & M. J. Farr (Eds.), *Aptitude, learning, and instruction. Volume 3: Cognitive and affective process analysis*. Hillsdale, NJ: Erlbaum.

Levy, R., & Mislevy, R. J (2004). Specifying and refining a measurement model for a computer-based interactive assessment. *International Journal of Testing, 4*, 333-369.

Margolis, M. J., & Clauser, B. E. (2006). A regression-based procedure for automated scoring of a complex medical performance assessment. In. D.W. Williamson, I. I. Bejar, & R. J. Mislevy (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 123-168) Mahwah, NJ: Lawrence Erlbaum.

Mislevy, R.J. (1994). Evidence and inference in educational assessment. *Psychometrika*, ***59***, 439-483.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). Design and analysis in task-based language assessment. *Language Testing*, 19(4), 477-496. doi:10.1191/0265532202lt241oa

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3-62.

Mislevy, R. J., Behrens, J. T., DiCerbo, K. E., & Levy, R. (in press). Data mining versus psychometrics in educational assessment: An Evidence Centered Design approach. Journal of Educational Data Mining.

Noddings, N. (2001).Care and coercion in school reform. *Journal of Educational Change, 2*, 35-43.

Olsen, S. (2000). Web browser offers incognito surfing. *CNET News.* Retrieved from http://news.cnet.com/2100-1017-247263.html

Peacock, Al, Ke, X., & Wilkerson, M. (2004). Typing patterns: a key to user identification. *Security and Privacy, IEEE, 2*, 40-47.

Pearl, J. 1998. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Prensky, M. (2001). Digital natives. Digital immigrants. *On the Horizon, 9*(5), 1-6.

Rowland, S. (2001, December 21). Use passion to counter culture of compliance. *Times Higher Education*. Retrieved from http://www.timeshighereducation.co.uk/story.asp?storyCode=166322&sectioncode=26

Rupp, A. A., Levy, R., DiCerbo, K. E., Benson, M., Sweet, S., Crawford, A. V., Fay, D., Kunze, K. L., Caliço, T. & Behrens, J. T. (in press). The Interplay of Theory and Data: Evidence Identification and Aggregation for Product and Process Data within a Digital Learning Environment. *Journal of Educational Data Mining*.

Rupp, A. A., Gushta, M., Mislevy, R. J., & Shaffer, D. W. (2010). Evidence-centered design of epistemic games: Measurement principles for complex learning environments. *Journal of Technology, Learning, and Assessment*, 8(4). Available online at http://escholarship.bc.edu/jtla/vol8/4.

Salen, K. (2012). Seminar. Presented at Educational Testing Services, Princeton, NJ.

Shute, V. J., Hansen, E. G., & Almond, R. G. (2008). You can't fatten a hog by weighing it – or can you? Evaluating an assessment for learning system called ACED. *International Journal of Artificial Intelligence in Education, 18*, 289-316.

Shute, V. J., & Kim, Y. J. (2011). Does playing the World of Goo facilitate learning?. In D. Y. Dai (Ed.), *Design research on learning and thinking in educational settings: Enhancing intellectual growth and functioning* (pp. 359-387). New York, NY: Routledge Books.

Shute, V. J., Ventura, M., Bauer, M., & Zapata-Rivera, D. (2009). Melding the power of serious games and embedded assessment to monitor and foster learning: Flow and grow. In U. Ritterfeld, M. J. Cody, & P. Vorderer (Eds.), *The social science of serious games: theories and applications* (pp. 295-321). Philadelphia, PA: Routledge/LEA.

Wainess, R., Koenig, A., & Kerr, D. (2011). *Aligning instruction and assessment with game and simulation design*. CRESST Research Report. Los Angeles: The National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA. Retrieved from: http://www.cse.ucla.edu/products/reports/R780.pdf

Williamson, D.W. (2012). The conceptual and scientific basis for automated scoring of performance items. In R. W. Lissitz & H. Jiao (Eds). *Computers and their impact on state assessments: Recent history and predictions for the future* (pp. 157-194). Charlotte, NC: Information Age Publishing.

Williamson, D. W., Almond, R. G., Mislevy, R. J., & Levy, R. (2006). An application of Bayesian networks in automated scoring of computerized simulation tasks. In. D.W. Williamson, I. I. Bejar, & R. J. Mislevy (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 123-168) Mahwah, NJ: Lawrence Erlbaum.

7. PREPARING FOR THE FUTURE: WHAT EDUCATIONAL ASSESSMENT MUST DO

Randy Bennett

There is little question that education is changing, seemingly quickly and in some cases dramatically. The mechanisms through which individuals learn are shifting from paper-based ones to electronic media. Witness the rise of educational games available on the personal computer, tablet, and mobile phone, as well as the attention being given to those games by the academic community (e.g., Gee & Hayes, 2011; Shaffer & Gee, 2006). Simultaneously, the nature of what individuals must learn is evolving, in good part due to an exponential accumulation of knowledge and of technology to access, share, and exploit that knowledge. In the U.S., the reconceptualization of school competency in the form of the Common Core State Standards (National Governors Association Center for Best Practices & Council of Chief State School Officers [NGO & CCSSO], 2010) signals one attempt to respond to that change. Finally, how education is organized, offered, and administered is undergoing transformation, most apparently–but not only–in higher education. The possibility of assembling one's post-secondary education from free Internet course offerings, with achievement documented through certification "badges," appears to be rapidly coming to reality (Young, 2012).

With potentially seismic changes in the mechanisms, nature, and organization of education must also come changes in educational assessment (Bennett, 2002). Otherwise, education and assessment will work against one another in ever-increasing ways. This paper offers a set of 13 claims about what educational assessment must do if it is to remain relevant, and even more so, if it is to actively and effectively contribute to individual and institutional achievement. The claims are that educational assessment must:

1. Provide meaningful information
2. Satisfy multiple purposes
3. Use modern conceptions of competency as a design basis
4. Align test and task designs, scoring and interpretation with those modern conceptions
5. Adopt modern methods for designing and interpreting complex assessments

6. Account for context
7. Design for fairness and accessibility
8. Design for positive impact
9. Design for engagement
10. Incorporate information from multiple sources
11. Respect privacy
12. Gather and share validity evidence
13. Use technology to achieve substantive goals

Each of these claims is discussed in turn.

Provide Meaningful Information

It should be obvious that in order to make sensible decisions about the effectiveness of education systems and the preparedness of populations, policymakers need meaningful information. Similarly, teachers and students need meaningful information if they are to effectively plan and adjust instruction. The implication of this claim is that, to be relevant, future educational assessment systems will need to provide trustworthy and actionable summative information for policymakers (including local administrators) as well as formative information for teachers and students.

For both types of assessment, the provision of "meaningful information" implies results that faithfully reflect the state of educational achievement for an individual or a group. That reflection will be at a finer grain size in the formative case and at a larger one for the summative case. "Faithfully" implies the existence of an evidentiary argument that supports the fidelity of that reflection (Mislevy, Almond, and Lukas, 2003). Ideally, that reflection should carry with it implications for action – whether microadjustments to learning or macroeducation-policy changes – which also should be supported by an evidentiary argument.

There is no indication that the need will subside for such information or for assessment mechanisms to provide that information. If anything, the need will increase because of the international competition enabled by a global economy and by the belief that a productive and educated workforce is central to maintaining (or increasing) one's standard of living in such an economy (Bennett & Gitomer, 2009). The rapid growth of international assessments is one indicator of this need for summative educational information. In 2000, 43 countries/economies participated in PISA, the *Programme for International*

Student Assessment, sponsored by the Organizations for Economic Co-operation and Development (OECD, n.d. a). By 2012, 64 entities were being assessed (OECD, n.d. b). Interestingly, the proportional increase in participation was overwhelmingly due to an influx of *non*-OECD countries/economies, which tend to be less economically developed than the Organization's membership. In 2000, 14 of the 43 entrants were non-OECD members, whose participation was presumably motivated by the chain of reasoning stated above (i.e., an educated workforce leads to an improved standard of living). In 2012, 30 of the 64 participants were non-OECD members.

A similar case can be made with respect to the need for effective formative assessment. Interest in formative assessment has grown dramatically since publication of the 1998 position and review papers by Black and Wiliam (1998a, 1998b, 1998c). This interest is fueled by the belief that formative assessment actually does what proponents claim – i.e., causes large improvements in learning. Although the research findings and conceptual grounding underlying such claims have been over-stated at best (Bennett, 2011; Coffey, Hammer, Levin, & Grant, 2011; Kingston and Nash, 2011), to remain relevant the educational assessment community must do its best to produce tools and practices that do, in fact, enhance achievement. Educators expect it and students deserve it.

The question, then, is not whether summative and formative assessments will continue to be necessary but rather the form(s) they will take and the competencies they will measure, claims to which we will soon turn.

Satisfy Multiple Purposes

The previous claim indicated that educational assessment must provide meaningful information for summative and formative purposes. As stated, that claim is somewhat oversimplified because, in reality, the demand for meaningful information centers upon *multiple* summative and *multiple* formative purposes. Education officials demand information to assist in evaluating students for promotion and graduation; schools (and school staff) for rewards and sanctions; and intervention programs for continuation and expansion. Educators also demand more fine-grained information for deciding what to teach, when, and to whom, and for helping teachers refine their instructional practice, and improve educational programs.

It should be obvious that this array of purposes cannot possibly be satisfied with a single test because an assessment built for one purpose won't necessarily be suited to other purposes. Building an assessment to serve many purposes is also unlikely to work because an assessment designed for multiple ends may prove optimal for none of its targeted goals. A formative assessment used to generate summative information incidentally is likely to do a poor job at both purposes (for reasons to be discussed later). Multiple purposes might best be served by different, related assessments designed to work in synergistic ways — i.e., through modular systems of assessment. The modular systems approach is the one taken by the Smarter Balanced (Smarter Balanced Assessment Consortium, 2010) and Partnership for Assessment of Readiness for College and Careers (2010) assessment consortia, as well as by such research initiatives as CBAL (Bennett, 2010; Bennett & Gitomer, 2009).

Use Modern Conceptions of Competency as a Design Basis

Across competency domains, the knowledge, processes, strategies, and habits of mind that characterize communities of practice differ fundamentally. At the same time, there are competencies that appear to be more general (Gordon, 2007). Our knowledge about the nature of these general, as well as domain-based, proficiencies is constantly evolving. In addition, the proficiencies our society considers to be important are evolving. The implication of this claim is that assessment design must be firmly grounded in up-to-date conceptions of what it means to be a proficient performer within valued domains, as well as in those competencies that have more general applicability (including socio-emotional ones). Either a domain-based focus or a general focus alone will not suffice (Perkins and Salomon, 1989).

Unfortunately, the conceptions of competency that underlie many current tests, especially those used in primary and secondary school assessment programs, have their grounding in a behaviorist learning theory circa 1950 rather than in the modern learning sciences (Shepard, 1991). In general, those assessment programs do not directly measure knowledge construction, knowledge organization, knowledge schema, procedural fluency, the coordination and integration of competencies required for complex performance, and the problem-solving process, to name a few key constructs. Nor do those tests account for the qualitative orderings in competency development, or learning progressions, that are emerging from theory and research (Corcoran, Mosher, & Rogat, 2009; Daro, Mosher, & Corcoran, 2011; Educational Testing Service, 2012). Such progressions could potentially increase the relevance of test results for teachers and students.

One implication of this claim is that although content standards, such as the Common Core State Standards (NGA and CCSSO, 2010) help, those standards do not necessarily reflect findings from the learning sciences in ways that can effectively guide test design. A bridge from content standards to test design can be provided by competency models that identify the components required for successful performance within and across domains, and how those components might be organized; learning progressions describing hypothesized paths to competency development; and principles for good teaching and learning practice (Bennett, 2010). Describing the literature base underlying the models, progressions, and principles is a key to making the case for those entities as a credible design basis.

Align Test and Task Designs, Scoring, and Interpretation with Those Modern Conceptions

It is one thing to espouse grounding design in modern conceptions of competency. It is another thing to do it. Doing it means, at the least, developing competency models that propose what elements make for proficiency in a domain (and across domains), how those elements work together to facilitate skilled performance, and how they might be ordered as learning progressions for purposes of instruction. Second, it means extracting from research a set of principles for good teaching and learning practice to guide assessment design. Finally, it means developing an assessment design, the tasks composing it, and mechanisms for the scoring and interpretation of examinee performance that are logically linked to the competency model, learning progressions, and/or principles for good teaching and learning practice. That linkage should be documented in a detailed design document that becomes part of the interpretive argument for the test (Kane, 2006).

An important implication of aligning with modern conceptions of competency, at least in the world of primary and secondary schools, is that educational assessment will need to go well beyond traditional item formats (Bennett & Ward, 1993; Pellegrino, Chudowsky, & Glaser, 2001). Modern conceptions recognize the importance of posing reasonably realistic problems that call upon examinees to connect knowledge, processes, and strategies to conditions of use. Those conceptions also posit the importance of problems requiring students to exercise control over multiple competencies simultaneously, and then deploying and integrating those competencies in planful ways to achieve a desired result. Such conceptions will make mandatory the use of more complex tasks, including simulations and other extended constructed-response formats. That use, however, needs to be clearly motivated by the need to measure competencies that cannot be assessed through less labor-intensive means (or by some other important benefit).

Although modern conceptions of competency will make the use of complex tasks unavoidable, that use should not necessarily dominate. More elemental, discrete tasks are needed to decompose complex performance for formative purposes; i.e., to help teachers and students identify which subcompetencies might be responsible for failure on a complex task. For summative purposes, discrete items also can play a role by helping to reduce the impact of such unwanted task effects as lack of generalizability (Linn & Burton, 1994).

Finally, future-scoring mechanisms, regardless of whether human or automated, will need to align with relevant domain processes. Ideally, more sophisticated scoring methods should bring with them the ability to recover the very knowledge structures, problem-solving processes, strategies, and habits of mind that tasks are designed to evoke. One might try to justify scoring responses through methods that don't attempt to account directly for the target competencies (e.g., machine learning, regression of human scores on nonaligned response features), but that justification would be a weak one.

Adopt Modern Methods for Designing and Interpreting Complex Assessments

To align design, scoring, and interpretation to modern conceptions of competency, we will need to adopt modern methods. Methods such as evidence-centered design (ECD) (Mislevy, Almond, & Lukas, 2003) and assessment engineering (Luecht, 2009) offer well-founded inferential structures and mechanisms to aid in the creation of assessments and in making sense of the results. Frameworks like ECD offer: a) a way of reasoning about assessment design, b) a way of reasoning about examinee performance, c) a data framework of reusable assessment components, and d) a flexible model for test delivery.

Reasoning about assessment design begins with specifying the claims to be made about individuals or institutions on the basis of assessment results. Those claims should derive directly from competency models and learning progressions. Specified next is the evidence needed to support those claims. Finally, the tasks required to elicit that evidence are described.

In assessment design, the reasoning chain is as follows: examinees whose competencies are consistent with a given claim will provide particular evidence in responding to the described tasks. Reasoning about examinee performance proceeds in the reverse direction. That is, when a given examinee offers evidence consistent with a claim in response to an aligned task, we can infer with some estimable level of uncertainty that the examinee meets the claim. As more task responses from that examinee are gathered to provide evidence about the claim, our belief in examinee standing with respect to the claim is updated and our level of uncertainty generally gets smaller.

Evidence is accumulated through a measurement model that generates a score, a qualitative characterization (e.g., a level in a learning progression, a diagnosis), or both. That measurement model also provides an estimate of the uncertainty associated with that score or characterization. The operational infrastructure in most large testing programs today can accommodate simple measurement models, generally models that array examinees along a single dimension. The operational infrastructure needs to be created for multidimensional models — i.e., models that extract evidence from an item for more than one dimension simultaneously.

Measurement models are only important, of course, if the purpose of assessment is to characterize student performance in some way that requires the notion of uncertainty. Inferences about some latent attribute of the student (e.g., that the student has achieved proficiency in some domain, or has a given standing with respect to some variable of interest), the likelihood that the student will perform acceptably in some other environment, or the likelihood that the student is a member of a particular diagnostic category, all bring with them such uncertainty. In contrast, if the purpose of assessment is simply to judge a student's performance qua performance–as in an Olympic sporting event–without any attribution beyond describing the observed result, then no inference is needed, no uncertainty is implied, and no measurement model is required. That a student achieved a particular score or ranking in an event, and won a medal (or didn't), are facts. (See Messick, 1992, for discussion of these two situations in the context of performance assessment.)

A third benefit of modern design methods is the potential for a data framework of reusable assessment components. For example, task models can be created to specify the elements of a family of questions (e.g., competency model and learning progression claim, stimulus characteristics, stem characteristics, response format). Generalized rubrics then can be created for scoring that family of questions (Bennett, Morley, & Quardt, 2000). Evidence model fragments that accumulate responses across some specified set of tasks can be generated. These task models, generalized rubrics, and evidence model fragments can, in principle, be stored in a data library. Creating a new assessment then proceeds by selecting data components that match the claims of interest.

A last design benefit is a flexible infrastructure delivery model. The four-process model consists of activity selection, presentation, response processing, and summary scoring (evidence accumulation). Creating the delivery infrastructure so that the four processes are separate allows for assembling new assessments, or changing old ones, in modular fashion. For example, the activity selection and presentation processes might be set to use members from the same task model in both a summative test and a diagnostic assessment but the response processing and summary scoring processes might be differently configured for those two use cases. For the summative case, response processing might extract a correct/incorrect judgment for each answer and accumulate across answers so as to estimate standing on a single dimension, whereas for the diagnostic assessment, aspects of the examinee's answer process might be judged and accumulated to produce a qualitative characterization.

Account for Context

A student's performance on an assessment – that is, the responses the student provides and the score the student achieves – is an indisputable fact. *Why* the student performed that way, and in particular, what that performance says about the student's competencies, is an interpretation. For many decision-making purposes, to be actionable, that interpretation needs to be informed by an understanding of the context in which the student lives, learns, was taught, and was assessed.

This need is particularly acute for large-scale tests for which decisions typically center upon comparing individuals or institutions to one another, or to the same competency standard, so as to facilitate a particular decision (e.g., graduation, school accountability, postsecondary admissions). Because of the need to present all students with the same tasks (or types of tasks) administered under similar conditions, those tests, in contrast to classroom assessment, will be far more distant in design, content, and format from the instruction students actually encounter. That distance is predicated upon the intention to measure competencies likely to manifest themselves across a variety of contexts, rather than in any particular one. In this sense, such tests are "out of context."

At present, our attempts to factor context more finely into the interpretation of large-scale test results take a variety of forms. In college and graduate admissions, for example, context is provided indirectly by grade-point-average and transcripts, and more directly by letters of recommendation and personal statements. These factors are combined clinically by admissions officials in decision making. For federal school accountability purposes, under *No Child Left Behind*, limited contextual data must be reported in addition to test-related information, including tabulations concerning "highly qualified teachers" and attendance and dropouts (State of New Jersey, Department of Education, n.d. b).

States may choose to compile additional information outside the requirements of *NCLB*. The complete New Jersey state "School Report Card" includes average class size, length of school day, instructional time, student/computer ratio, Internet connectivity, limited English proficiency rate, disability rate, student attendance rate, dropout rate, graduation rate, student suspensions and expulsions, student/faculty ratio, faculty attendance rate, faculty mobility rate, faculty and administrator credentials, National Board of Professional Teaching Standards certification, teacher salaries, and per pupil expenditures (State of New Jersey, Department of Education, n.d. a). Although New Jersey provides a wealth of information about the school-level context in which students are being educated, it offers no guidance about how to use that information for interpreting test results. Further, the state offers very little insight into the instructional context that characterizes any given classroom or into the home environment in which its students reside. How those factors should shade the interpretation of assessment results, and inform action, is left for teachers and parents to gauge for themselves.

Embedding assessment directly into the learning context – i.e., more closely integrating assessment with curriculum and instruction – should make assessment information more actionable for formative purposes. Such embedded assessments will be integral components of anytime/anywhere, online learning environments into which those assessments can be seamlessly fit. For a variety of reasons, this in-context performance might not be useful for purposes beyond the classroom or learning environment that is generating the data (e.g., for school accountability, college admissions, teacher evaluation). The large number and wide diversity of such learning environments may not make aggregation meaningful. In addition, attaching significant consequences to activity in environments built to facilitate learning may unintentionally undermine both the utility of the formative feedback and achievement itself (Black & Wiliam, 1998a). Last, the constant and potentially surreptitious surveillance of student behavior may pose privacy issues significant enough that some students opt out.

Design for Fairness and Accessibility

Among our country's social values is the idea of fairness in the form of equal opportunity for individuals, as well as for traditionally underserved groups. In standardized testing, fairness for individuals was a motivating concern from the earliest implementations of the practice, going back to the ancient Chinese civil service examinations (Miyazaki, 1976), which were instituted to ensure that jobs were awarded on merit rather than social class or family connections.

In the United States, concern for fairness did not originally extend to groups. In fact, several of the field's progenitors expressed racist views, perhaps most obviously in their interpretations of test results (e.g., Brigham, 1923) and most destructively in their failure to object to the use of their work to support racist and anti-immigration political agendas. Among the earliest statements of concern for group fairness from within the field was that of Carl Brigham (1930, p. 165) who, ironically, was a former eugenicist:

> For purposes of comparing individuals or groups, it is apparent that tests in the vernacular must be used only with individuals having equal opportunities to acquire the vernacular of the test. This requirement precludes the use of such tests in making comparative studies of individuals brought up in homes in which the vernacular of the test is not used, or in which two vernaculars are used. The last condition is frequently violated here in studies of children born in this country whose parents speak another tongue. It is important, as the effects of bilingualism are not entirely known.

He went on:

> This review has summarized some of the more recent test findings which show that comparative studies of various national and racial groups may not be made with existing tests, and which show, in particular, that one of the most pretentious of these comparative racial studies – the writer's own – was without foundation. (p. 165)

Brigham's concern unfortunately did not take root for many years to come (with the notable exception of the *SAT*, which was instituted in the 1930s to *increase* access for economically diverse students to Harvard and other selective institutions [Bennett, 2005]). Among other things, tests were used well into the 1960s as a component of state-sanctioned, institutionalized racism. Reading test performance was used in some states as a registration requirement, thereby denying many African-American citizens the right to vote (U.S. Department of Justice, n.d.).

The measurement community began to turn concerted attention to defining, identifying, and removing unfairness in tests in the late 1960s and early 1970s as part of a larger societal movement to redress racial discrimination (Cole & Zieky, 2001). Similar concerns surfaced in the 1970s around accessibility and fairness for individuals with disabilities, most particularly with respect to postsecondary admissions tests (e.g., Sherman and Robinson, 1982; Willingham et al., 1988). Current concerns for the fairness and accessibility of tests for English language learners bring Brigham's (1930) statement full circle.

As noted, concerns for fairness are a social value, emerging first for fairness at the individual level and, later, for groups. Appreciation of the need for group fairness has been aided by the growing diversity of our society and the activism of those who were disenfranchised.

Concern for fairness will continue regardless of the form that future educational assessments take. Those tests will have to factor fairness into test design, delivery, scoring, analysis, and use. That concern will not be restricted to consequential tests but extend to formative assessment as well. Formative assessments entail a two-part validity argument: a) that the formative instrument or process produce meaningful inferences about what students know and can do, leading to sensible instructional adjustments and b) that these inferences and instructional adjustments consequently cause improved achievement (Bennett, 2011). Fairness would seem to require that this argument hold equally well across important population groups–that is, a formative assessment instrument or process should provide similarly meaningful inferences about student competency, suggest similarly sensible instructional adjustments, and lead to similar levels of instructional improvement. Conceivably, a differentially valid formative assessment, used indiscriminately, could have the unwanted effect of *increasing* achievement gaps among population groups. Preventing such an occurrence might require the design and use of *demographically sensitive* formative assessments, in concept like pharmaceuticals created to target particular population groups (Saul, 2005). In a free-market system, however, development will be most concentrated on the needs of those most able to pay, leaving to government and advocacy organizations the task of ensuring that attempts are made to address instances of differential validity that disfavor underserved groups, when such instances do occur.

Design for Positive Impact

It is generally acknowledged that, for consequential assessments, test design and use can have a profound impact – sometimes intended, sometimes not – on individuals and institutions (Koretz & Hamilton, 2006). Examples of impact may be on the behavior of teachers and students, or on the behavior of organizations (e.g., schools). *No Child Left Behind* was premised on intended positive impact. That is, test use was intended to focus educators in underachieving schools on the need to improve and, in particular, on improvement for underserved student groups.

Test design and use also can have unintended effects. In the case of *No Child Left Behind*, those effects are commonly asserted to include large amounts of instructional time spent “teaching to the test,” in essence, an extreme curricular narrowing caused by the interaction of the Act’s focus on reading and mathematics, a patchwork of mostly low-quality content standards among the states, the constrained methods used to measure achievement of those standards, and the sanctions placed on schools that fail to achieve required levels of proficiency.

The reasoning behind the *Race to the Top Assessment Program*, which the U.S. Department of Education instituted to fund development of Common Core State Assessments, appears to be that, if low-quality standards and narrow assessments can have negative effects, then high-quality standards and assessments ought to be able to have a positive impact (U.S. Department of Education, 2010). The implication of this claim is that impact must be explicitly taken into account at the assessment-design stage. By using principles and results from learning sciences research, summative assessments can be built to model good teaching and learning practice (Bennett, 2010). That modeling can occur via: a) giving students something substantive and reasonably realistic with which to reason, read, write, or do mathematics or science; b) routinely including tools and representations similar to ones proficient performers employ in their domain practice; c) designing assessment tasks to help students (and teachers) connect qualitative understanding with formalism; d) structuring tests so that they demonstrate to teachers how complex performances might be scaffolded; and e) using learning progressions to denote and measure levels of qualitative change in student understanding.

Designing for positive impact might also mean preserving the idea of a consequential test—i.e., an event for which students must prepare. If the test is a faithful representation of the competencies and situations of use at which education is targeted, intensive preparation can have beneficial effects. Among other things, practice leads to automaticity, and to knowledge consolidation and organization. Testing can have positive effects by strengthening the representation of information retrieved during the test and also slowing the rate of forgetting (Rohrer and Pashler, 2010).

Design for Engagement

Assessment results are more likely to be meaningful if students give maximum effort. Electronic game designers seem to have found ways to get students to give that effort. Assessment designers will also need to find new ways to enhance engagement. Designers might start by: a) posing problems that examinees are likely to care about; b) providing motivating feedback; c) using multimedia and other game elements; and d) employing delivery hardware preferred by the target population (e.g., smart phones, tablets), where that hardware is appropriate to the task demands of the domain.

Why not simply embed assessment into a game, thereby creating an engaging assessment? For formative purposes, that strategy might work to the extent that the game was designed to exercise relevant competencies and game play can be used to generate meaningful information for adjusting instruction, either inside or outside of the game. For summative purposes, game performance might offer useful information *if*, among other things, everyone plays the same game, or a common framework can be devised for meaningfully aggregating information across students playing different games intended to measure the same (relevant) competencies. That latter idea is employed in the *Advanced Placement Studio Art* assessment, for which students undertake different projects, all of which are graded according to the same criteria (Myford & Mislevy, 1995).

In short, assessments of the future will need to be designed for engagement but not for the purpose of simply making assessments fun. Rather, they will need to be designed for engagement that facilitates, better than current assessments, measuring the competencies of interest for the assessment purposes at hand.

Incorporate Information from Multiple Sources

All assessment methods–tests, interviews, observations, work samples, games, and simulations–*sample* behavior. Further, each method is subject to its own particular limitations, or method variance. In combination, these facts argue for the use of multiple methods in generating information, certainly for the making of consequential decisions about individuals and institutions. Multiple sources are commonly used for such consequential decisions as postsecondary admissions, where grade-point-average and tests scores are often combined with one another through decision rules, and further clinically integrated with information from interviews, personal statements, and letters of recommendation.

To the extent practicable, this claim also would suggest using multiple sources of evidence for formative decision making. Rather than adjusting instruction for the class or an individual on the basis of a single interaction or observation, the teacher would be wise to regard the inference prompted by that initial observation as a “formative hypothesis” (Bennett, 2010), to be confirmed or refuted through other observations. Those other observations could be past classroom behavior, homework, quizzes, or the administration of additional tasks directly targeted at testing the hypothesis. As technology infuses learning and instruction, the amount and type of other information available only will increase.

Respect Privacy

In a technology-based learning environment, assessment information can be gathered ubiquitously and surreptitiously. Some commentators have suggested that this capability will lead to the “end of testing” (Tucker, 2012). That is, there will be no reason to have stand-alone assessments because all of the information needed for classroom, as well as for accountability purposes, will be gathered in the course of learning and instruction.

Whereas this idea may seem attractive on its surface, students (as well as teachers) have privacy rights that assessment designers will need to respect. For one, Individuals should know when they are being assessed and for what purposes. Their knowledgeable participation in assessment thereby becomes their informed consent. Second, having every learning (and teaching) action recorded and potentially used for consequential purposes is, arguably, an unnecessary invasion of the student's (and teacher's) right to engage freely in intellectual activity. That privacy invasion could potentially stifle experimentation in learning and teaching, including the productive making of mistakes (Kapur, 2010). Third, as a functionary of the state, the public school's right to ubiquitously monitor student and teacher behavior is debatable at best. In the U.S., at least, the state can monitor public behavior–as in the use of traffic and security cameras–particularly when that monitoring is in the interest of public safety. Except in very circumscribed instances, private behavior cannot be monitored without a court order. Whether the state can monitor learning behavior (as separate from testing behavior), and use that behavior to take actions that affect a student's life chances is an open question.

A compromise position that attempts to respect individual privacy and provide information for making consequential, as well as instructional, decisions might be a model similar to that used in many sports. In baseball, the consequential assessment of performance that counts toward player statistics and team standing occurs during the game, and only during the game. Spring training, before-game practice, in-between inning practice, and in between-game practice are primarily reserved for learning. We might consider doing the same for assessment embedded in learning environments–use separately identified periods for consequential assessment versus learning (or practice).

Gather and Share Validity Evidence

However innovative, authentic, or engaging they may prove to be, future assessments will need to provide evidence to support the inferences from, and uses of, assessment results. Legitimacy is granted to a consequential assessment by a user community and the scientific community connected to it. Among other things, that legitimacy depends upon the assessment program providing honest evaluation, including independent analysis, of the meaning of assessment results and the impact of the assessment on individuals and institutions; reasonable transparency in how scores are generated; and mechanisms for continuously feeding validity results back into the improvement of the assessment program.

With respect to score generation, transparency must be apparent at least to members of the scientific community who are experts in the field, for it is these individuals who represent and advise the user community on technical matters. The need for transparency implies that score generation methods (e.g., automated scoring of constructed responses) cannot be so closely held by test vendors as to prevent independent review. In essence, "Trust us" approaches don't work when people's life chances are at stake.

One method for protecting intellectual property and permitting independent review is patent. A second, but less desirable approach from a transparency point of view, would be to grant access under a nondisclosure agreement to the user community's scientific advisors (e.g., members of a testing program's technical advisory committee). Those advisors could then report back to the user community in general terms that preserve the vendor's confidentiality but assure the technical quality of the scoring method.

Use Technology to Achieve Substantive Goals

The final claim is that future assessments will need to use technology to do what can't be done as well (or at all) with traditional tests. Among those uses will be to measure existing competencies more effectively (and efficiently), for example, by scoring complex responses automatically or administering tests adaptively. A second use will be to measure new competencies. New competencies could include aspects of competencies we currently measure; for example, current tests measure the result of problem solving but technology also could be used to measure features of the examinee's problem-solving process (Bennett, Persky, Weiss, & Jenkins, 2010). Third, technology might be deployed to have positive impact on teaching and learning practice. Using technology without the promise of a clear substantive benefit ought to be avoided.

Conclusion

Education, and the world for which it is preparing students, is changing quickly. Educational assessment will need to keep pace if it is to remain relevant. This paper offered a set of claims for how educational assessment might achieve that critical goal.

Many of these claims are ones to which assessment programs have long aspired. However, meeting these claims in the face of an education system that will be digitized, personalized, and possibly gamified, will require significantly adapting, and potentially reinventing, educational assessment. Our challenge as a field will be to retain and extend foundational principles, applying them in creative ways to meet the information and decision-making requirements of a dynamic world and the changing education systems that must prepare individuals to thrive in that world.

References

Bennett, R. E. (2002). Inexorable and inevitable: The continuing story of technology and assessment. *Journal of Technology, Learning, and Assessment, 1*(1). Retrieved from http://escholarship.bc.edu/jtla/vol1/1/

Bennett, R. E. (2005). *What does it mean to be a nonprofit educational measurement organization in the 21st-century?* Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/Nonprofit.pdf

Bennett, R. E. (2010). Cognitively Based Assessment of, for, and as Learning: A preliminary theory of action for summative and formative assessment. *Measurement: Interdisciplinary Research and Perspectives, 8*, 70-91.

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy and Practice 18*, 5-25.

Bennett, R. E., & Gitomer, D. H. (2009). Transforming K-12 assessment: Integrating accountability testing, formative assessment, and professional support. In C. Wyatt-Smith & J. Cumming (Eds.), *Educational assessment in the 21st-century* (pp. 43-61). New York, NY: Springer.

Bennett, R. E., Morley, M., & Quardt, D. (2000). Three response types for broadening the conception of mathematical problem solving in computerized tests. *Applied Psychological Measurement, 24*, 294-309.

Bennett, R.E., Persky, H., Weiss, A., & Jenkins, F. (2010). Measuring problem solving with technology: A demonstration study for NAEP. *Journal of Technology, Learning, and Assessment, 8*(8). Retrieved from http://escholarship.bc.edu/jtla/vol8/8

Bennett, R. E., & Ward, W. C. (Eds). (1993). *Construction vs. choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment*. Hillsdale, NJ: Erlbaum.

Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education, 5*(1), 7-74.

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

Black, P., & Wiliam, D. (1998c). *Inside the black box: Raising standards through classroom assessment*. London, England: Kings College, London School of Education.

Brigham. C. C. (1923). *A study of American intelligence*. Princeton, NJ: Princeton University Press.

Brigham, C. C. (1930). Intelligence tests of immigrant groups. *Psychological Review, 37*, 158-165.

Coffey, J. E., Hammer, D., Levin, D. M., & Grant, T. (2011). The missing disciplinary substance of formative assessment. *Journal of Research in Science Teaching, 48*, 1109–1136.

Cole, N. S., & Zieky, M. J. (2001). The new faces of fairness. *Journal of Educational Measurement, 38*, 369-382.

Corcoran, T., Mosher, F. A., & Rogat, A. (2009). *Learning progressions in science: An evidence-based approach to reform*. New York, NY: Teachers College-Columbia University.

Daro, P., Mosher, F. A., & Corcoran, T. (2011). *Learning trajectories in mathematics: A foundation for standards, curriculum, assessment, and instruction*. Philadelphia, PA: Center for Policy Research in Education.

Educational Testing Service (ETS). (2012). *The CBAL English Language Arts (ELA) Competency Model and Provisional Learning Progressions.* Princeton, NJ: Author. Retrieved from http://elalp.cbalwiki.ets.org/

Gee, J. P., & Hayes, E. R. (2011). *Language and learning in the digital age.* Milton Park, Abingdon, England: Routledge.

Gordon, E. W. (2007). Intellective competence: The universal currency in technologically advanced societies. In E.W. Gordon & B. R. Bridglall (Eds.), *Affirmative development: Cultivating academic ability* (pp. 3-16). Lanham, MD: Rowan & Littlefield.

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education and Praeger.

Kapur, M. (2010). Productive failure in mathematical problem solving. *Instructional Science, 38*(6), 523-550.

Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice, 30*,(4), 28–37.

Koretz, D. M., & Hamilton, L. S. (2006). Testing for accountability in K-12. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 531–578). Westport, CT: American Council on Education and Praeger.

Linn, R. L., & Burton, E. (1994). Performance-based assessment: Implications of task specificity. *Educational Measurement: Issues and Practice, 13*(1), 5–8.

Luecht, R.M. (2009). Adaptive computer-based tasks under an assessment engineering paradigm. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*. Retrieved from www.psych.umn.edu/psylabs/CATCentral/

Messick, S. (1992). *The interplay of evidence and consequences in the validation of performance assessments* (Research Report 92-39). Princeton, NJ: Educational Testing Service.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief introduction to evidence-centered design* (Research Report 03-16). Princeton, NJ: Educational Testing Service.

Miyazaki, I. (1976). China's examination hell: The civil service examinations of Imperial China. New York, NY: Weatherhill.

Myford, C. E., &Mislevy, R. J. (1995). *Monitoring and improving a portfolio assessment system* (MS-94-05). Princeton, NJ: Educational Testing Service.

National Governors Association Center for Best Practices & Council for Chief State School Officers. (2010). *Common core state standards.* Washington, DC: Author.

Organisation for Economic Cooperation and Development. (n.d. a). Programme for International Student Assessment (PISA): PISA 2000 participants. Retrieved from http://www.oecd.org/pisa/participatingcountrieseconomies/pisa2000listofparticipatingcountrieseconomies.htm

Organisation for Economic Cooperation and Development (OECD). (n.d. b). Programme for International Student Assessment (PISA): PISA 2012 participants. Retrieved from http://www.oecd.org/pisa/participatingcountrieseconomies/pisa2012participants.htm

Partnership for Assessment of Readiness for College and Careers. (2010). *The Partnership for Assessment of Readiness for College and Careers (PARCC) application for the Race to the Top Comprehensive Assessment Systems Competition.* Retrieved from http://www.fldoe.org/parcc/pdf/apprtcasc.pdf

Pellegrino, J. W., Chudowski, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academy of Sciences.

Perkins, D. N., & Salomon, G. (1989). Are cognitive skills context bound? *Educational Researcher, 18*, 16-25.

Rohrer, D., & Pashler, H. (2010). Recent research on human learning challenges conventional instructional strategies. *Educational Researcher, 39*, 406-412.

Sherman, S. W., & Robinson, N.M. (Eds.). (1982). *Ability testing of handicapped people: Dilemma for government, science, and the public.* Washington, DC: National Academy Press.

Smarter Balanced Assessment Consortium. (2010). *Race to the Top Assessment Program application for new grants: Comprehensive assessment systems CFDA Number: 84.395B.* Retrieved from: http://www.k12.wa.us/SMARTER/RTTTApplication.aspx

Saul, S. (2005, June 24). F.D.A. Approves a Heart Drug for African-Americans. *New York Times.* Retrieved from http://www.nytimes.com/2005/06/24/health/24drugs.html

Shaffer, D. W., & Gee, J. P. (2006). *How computer games help children learn.* Houdsmills, UK: Palgrave MacMillan.

Shepard, L. A. (1991). Psychometricians beliefs about learning. *Educational Researcher, 20*, 2-16.

State of New Jersey, Department of Education. (n.d. b). *Guide to the New Jersey school report card 2011*. Trenton, NJ: Author. Retrieved from http://education.state.nj.us/rc/rc11/guide.htm

State of New Jersey, Department of Education. (n.d. a). *NCLB school, district, and state reports.* Trenton, NJ: Author. Retrieved from http://education.state.nj.us/rc/index.html

Tucker, B. (2012). Grand test auto: The end of testing. *Washington Monthly.* Retrieved from http://www.washingtonmonthly.com/magazine/mayjune_2012/special_report/grand_test_auto037192.php

US Department of Education. (2010). *Race to the Top Assessment Program: Application for new grants.* Washington, DC: Author. Retrieved from http://www2.ed.gov/programs/racetothetop-assessment/resources.html

US Department of Justice. (n.d.). *Introduction to federal voting rights laws.* Washington, DC: Author. Retrieved from http://epic.org/privacy/voting/register/intro_a.html

Willingham, W. W., Ragosta, M., Bennett, R. E., Braun, H. I. Rock, D. A., & Powers, D. E. (1988). *Testing handicapped people*. Boston, MA: Allyn & Bacon.

Young, J. (2012, January 8). 'Badges' earned online pose challenge to traditional college diplomas. *Chronicle of Higher Education.* Retrieved from http://chronicle.com/article/Badges-Earned-Online-Pose/130241/

8. TO ASSESS, TO TEACH, TO LEARN: A VISION FOR THE FUTURE OF ASSESSMENT IN EDUCATION

Edmund W. Gordon[8]

Toward a Vision for the Future of Assessment in Education

I often think of a discussion a few years ago, in which some of us were musing concerning the changing nature of the program and services of organizations like Educational Testing Service, where the original concern for testing has grown into a concern with education and approaches to its assessment. I recall one of us expressing the view that such organizations could drop the reference to testing from their institutional titles and simply bill themselves as education services. We were thinking that testing, as an isolated function, would be in decline while the other processes involved in education (assessment, teaching and learning) would be ascending. Where testing would be involved, we thought, an assessment perspective would replace the more traditional emphasis on measurement. This implicit preference may have been percipient of the future of measurement in education. I have advocated for a conception of education as a differentiated unity of pedagogy, as a Troika of essential, interdependent and dialectical component processes—assessment, teaching and learning.

From Testing to Assessment to Education

The education service institutions of the future would be equally concerned with assessment (not just testing), teaching and learning. I find it interesting, as I look at the wide variety of activities and programs that constitute the agenda of measurement and assessment, there is represented my pedagogical Troika, still dominated by a concern with testing, but progressively more inclusive of concerns for assessment, teaching and learning — the primary processes of education. I predict for the future of assessment in education a continuing presence but less dominating role for measurement and testing. The field and the institution will become much more deeply involved in teaching as inquiry, inspiration, instruction, mediation, modeling and assessment (Armour-Thomas and I call it dynamic pedagogy). The center stage will be occupied by the facilitation of human learning, much of which will be self-directed, independently mediated, and comprehensively accessed from multiple sources.

[8] The Author acknowledges with deep appreciation the editorial and research assistance of Emily B. Campbell, Paola C. Heincke and Paola A. Valencia in the preparation of this commentary.

I see the assessment component greatly influenced by human judgment grounded in evidentiary reasoning, the disconfirmation of continuously changing inferences, and the relational analysis of data generated from systems of probes distributed throughout assessment, teaching and learning transactions. I envision a shift away from the assessment OF education for the purposes of accountability, accreditation, selection and placement, and toward assessment FOR education through diagnosis, inquiry, instruction, modeling, and mediation for the purpose of informing and improving teaching and learning processes and outcomes. While we have honored the assessment ***OF*** and ***FOR*** distinctions, it is important to recall that the purposes implied by the ***OF/FOR*** distinction may be of less importance than the nature of the processes that are engaged. In some instances it may be necessary to take measurements "of" in order to assess "for" education. In other instances it may be necessary to include information from assessments made for education in judgments made concerning the quality of education and its achievement.

To Assess, To Teach and To Learn

The field will continue to be concerned with informing summary judgments concerning accountability, certification, placement and selection, but I hope these needs will not continue to so dominate in educational policy that it will sometimes distort what we are trying to achieve through teaching and learning. Rather, it is my hope that the field will privilege our concern for informing formative and summary judgments concerning the directed development of human capacities. I think the assessment enterprise in education will become an educative service, concerned with informing and improving teaching and learning, and modeling the adaptive, intellective and learning behaviors that exemplify the intended outcomes of education. Yet, in thinking about education and its assessment for the world of the 21st century, we are confronted with notions of intelligence, knowledge and even human abilities that are constrained by conceptions of fixity, stability, predictability, veridicality, and some would include "truth." My colleagues on the Gordon Commission and I have been debating our emerging concerns for what the assessment enterprise can become and the compatibility of such visions with the more traditional concerns and conceptions of educational measurement.

My highly respected friend and cognitive psychologist Robert Sternberg devoted more than half his career struggling with the componential analysis of human intelligence and established his tri-archaic theory of intelligence. This theory holds that meta-components, performance components, and knowledge acquisition components (1985) operate together to enable intelligent behavior. Most of our effort at understanding human adaptive capacity has privileged some conception of intelligence and most of those conceptions have included the mastery, memory and manipulation of knowledge and technique. And, our efforts at assessment and measurement have focused on documenting the amount and status of what we know and know how to do, as reflected in what and how much we know. Lately I have been thinking that more important than the focus on "what students know and know how to do," – the amount and status of what one knows, we should instead attend to a better understanding of human adaptive capabilities and potentials, and the processes by which they are being developed and modified in the learning person. Such understanding may require a sharper focus on the processes by which knowledge and techniques are acquired and human adaptation is achieved. Rather than a primary focus on the measurement of the status of one's knowledge and skills, the stage could be shared with a concern for the adequacy of the processes by which the capabilities of mind are developing and being utilized. It is interesting that in one of Sternberg's most recent contributions, he uses the kaleidoscope as a metaphor for imaging this process. Writing about college admissions, he sees the phenomena of focus in making judgments concerning human abilities as being in much more fluid and dynamic flux than could be revealed in his earlier efforts at componential analysis.

Education as the Cultivation of Intellective Competence

Our distinguished colleague on the Commission, the late Professor Michael Martinez, would enthusiastically nominate the cultivation of intellect – the nurturance of the mental abilities and capacities for which the study of any but certainly academic knowledge and technique are propaedeutic.

In such a world, the mastery of extant knowledge and technique is likely to be recognized as the pre-learning that is essential to the subsequent development of the personal command of one's access to adaptive and adaptable mental capacities that can be applied to any knowledge and technique, even though it is the learning of some specific knowledge and technique that was the genesis of the mental ability. The mental capacities do not function and cannot be acquired without access to knowledge and technique, but the mastery of the knowledge and technique is not the end toward which teaching and learning should be directed.

Some colleagues on the Gordon Commission have suggested that what is really involved concerns the problem of transfer learning. Just what is the cargo of transfer in learning?

What is actually transferred in human learning and intelligence appears to me to be the capacity for the intentional orchestration of the several components of one's developed mental abilities. What we know and know how to do are important, but the quality of this ability to put the pieces together in ways that make sense to self and others and that serve one's intentions appear to be more important. However, we cannot afford to forget or neglect the fact that without content – knowledge and technique – there is nothing of which to make sense. Bereiter and Scardamalia remind us that the more information, knowledge, and techniques that are accessible in the learner's repertoire, the greater is the potential for making connections and sense making. But what happens in new learning situations is not limited to making connections between what is quotidian and what is new. The real challenge is to be found in generating nonexisting relationships between things that are novel as well as between the novel and that which is transferred from prior learning situations. I think these processes involve the mental abilities or capacities that Snow and his student Martinez reference in their claim that education can and should cultivate intelligence.

Similarly, with access to knowledge, technique and production-models ubiquitously available, alongside technologies that are informed and enabled by digital information, the ends of teaching and learning and assessment will of necessity be manifestations of the relevant mental capabilities, even if these manifestations of mental capacity are reflected only in the possession and use of specific knowledge and technique. But in this conceptualization, the specific knowledge and technique should not be the only targets of assessment. The targets of assessment might better be the interchangeable, transferable and adaptable mental capabilities and, of course, the capacity to intentionally access them in the context of some system of knowledge or techniques. The specific knowledge and technique may change with time. It is the mental capabilities and the capacity to access, manage, understand and use them with intentionality that will become the ends toward which we direct teaching and learning, as well as the targets of what we will seek to assess for education, and in our continued assessments of education as we move through the 21st century. But in this new century, it appears that changing paradigms, shifting epistemologies, and emerging developments in the sciences and technology all point in the direction of making it more likely that we are, indeed, able to deliver this kind of teaching and learning. The members of the Gordon Commission have been considering whether it is possible to develop assessment instruments and procedures that are both sensitive to, and enabling of, nurturing or stimulating productive interactions between engagement with knowledge and technique and the development of mental abilities, as well as intentional access to them.

New Standards and Common Core Standards

We saw some movement in this direction in the New Standards initiative and more recently in some of the concerns driving the Common Core Standards movement. In New Standards we saw an intensive focus on the teaching and learning of knowledge and technique through activities that encouraged and demanded the intentional use of mental capacities. The New Standards Assessment probes were designed to model and elicit the kinds of responses that reflected high levels of content mastery through the demonstration of one's command of critical mental capacities. There is difference of opinion here concerning the relative emphasis on the content to be learned and the bi-products of such learning that may be reflected in enabled mental abilities or capacities. I am convinced that this should not be a debate about a choice between the two emphases, rather the continuing problem concerns the effective management of the interactions between knowledge and technique mastery and the processes by which the capacity for acquiring and accessing complex mental operations is achieved. I think mental capacities are developed and enabled by serious engagement with substantive knowledge and technique. I think Resnick's "accountable talk" (Michael, O'Connor & Resnick, 2008) is an example of such serious engagement and her data suggest that the use of this instructional and assessment technique results in more effective learning, and even in the transfer of such learning. The Gordon Commission has sought to sponsor a serious retrospective study of the New Standards Initiative as one of the nation's most important studies in assessment of and for education.

I had encountered what I saw in the New Standards Initiative once before. My friend W.E.B. Du Bois is my model of a well-developed scholar. He was also an excellent teacher. Du Bois was known for posing the Socratic type question in which he modeled in his questions the kind of answer or learning for which he was seeking. To answer W.E.B.'s probes, one had to have some knowledge, to which one had to apply a rich complement of mental capacities in order to make sense of the chunks of knowledge, to gain some understanding of how the knowledge chunks fit together, and to communicate the mechanisms and the meanings of the phenomena in question and how the respondent was dealing with it to others. Dr. Du Bois employed assessment, teaching and learning iteratively and bidirectionally. I see a model for the future of assessment in what was common practice for my mentor.

A Pedagogical Troika: Assessment, Teaching and Learning

Some of us argue that the expectation of such shifts and changes in our understanding will require us to transform what we do in, through, and with education. Educational assessment will certainly need to be responsive to these changed conceptions of human intellective competence and the implicit conditions by which intellective competence is acquired and manifested. Some of us believe that in the process assessment in education will have to become an integral part of the Troika – assessment, teaching and learning that we call pedagogy. We argue that assessment can and should inform the processes of teaching and learning. To do so, I believe that the conditions and the processes of assessment in education will change in the direction of what I have described as a tripartite multifunctional system.

The Gordon Commission on the Future of Assessment in Education has considered a vision of pedagogy – the central mechanism operative in education – that is interactively, dialectically and transformatively inclusive of assessment, teaching and learning. Our inclination to make concern for teaching and learning conjoint with the traditional concerns of assessment rests on a vision of teaching and learning as reciprocal human processes, which are directed at the understanding, enablement and achievement of high levels of intellective competence in all learners. That is, the interconnectedness of assessment, teaching, and learning affords all learners pathways toward the best attachment of information and intention to use it in relation to the contexts in which they are experienced. In this vision "to teach" is to enliven, enable, and empower learners through deliberately orchestrated learning experiences, guided exploration, mediated inquiry, didactic instruction, imagination and modeled explication. In contrast to earlier notions of teaching involving the transfer of knowledge, skills, and values, this view of pedagogy makes the teaching person a guide, a coach a model, an orchestrator, a stimulator, and a resource person. The reference to teaching and learning is bifocal and bidirectional and references the assimilation and accommodation of that which is old, as well as the active construction and integration of that which is new. While not rejecting the traditional emphasis on associative memory and endogenous retrieval processes, the new vision of assessment, teaching, and learning privileges constructive, trans-active and transformative social processes which are endogenous, exogenous, and situative in teaching as well as learning persons.

The products of these assessment, teaching, and learning endeavors are reflected in the achievement of intellective competence which references the developed abilities and dispositions to understand as well as to know, to perceive critically, to explore widely, to bring rational order to chaos, to bring knowledge and technique to bear on the solution of problems to test ideas against explicit and considered moral values–as well as empirical evidence–and at the same time to recognize and create material and abstract relationships between real and imaginary phenomena. These achievements are less focused on what we want learners to know and know how to do and are better aimed at what we want our students to aspire to become, to be disposed toward, and to actually be, i.e., thinking, agentic, anti-fragile and compassionate human beings.

Intellective Character and Competence

In our vision of assessment, teaching, and learning, achievement standards are central, but the explication of what we want learners to know about specific disciplines and to be able to do in meeting these standards must be considered as instrumental to the achievement of what we want learners to be and become. I have referred to this state as intellective competence. My friend and colleague James Greeno thinks that what I am talking about also includes aspects of character. Should we ever complete our joint effort on this subject, we will no doubt refer to "character and competence." The old "scholastic aptitudes" may not have been so far from the mark in the effort at the time to achieve some distance from the specific academic content covered in the diverse curricula of the nation. Those "scholastic aptitudes," I prefer abilities, can be thought as generalized developed abilities and dispositions that not only reflect the capacity to handle academic knowledge and techniques, but, more importantly, reflect the adaptive abilities and dispositions that result from effective education of high quality. Instead of scholastic aptitudes it may be more appropriate to think of developed intellective abilities, or competencies. These developed abilities are not so much reflected in the specific discipline-based knowledge and skills a student may have acquired, but in the ability and disposition to use the meta-products of having experienced education to engage and solve quotidian, as well as novel, problems intentionally. I suggest that these manifestations of intellective competence[9] become the ends toward which education should be directed and the targeted indicators upon which assessment should be focused.

[9] I may eventually back away from my use of intellective as the qualifier for competence and return to intellectual because of the cool reception my use of the term has received. Some people simply don't understand. Some think it is jargon, and some are simply distracted or let themselves become distracted from the more substantive issues involved in my use of the term. I use intellective for three reasons. 1) I do not want my use of intellectual to be confused with the more colloquial use of intellectual as it refers to those of us who are associated with the academy or esoteric knowledge. 2) I certainly want to distinguish what I am talking about from the notion of fixed or inherited intelligence. 3) I do want to reference the developed ability and disposition to express human agency in deliberately adaptive activity; to explore and engage purposefully; to seek and utilize informational, human and material resources; to solve quotidian as well as novel problems; to bring order to chaos; and to make sense of the world that one encounters. My construction is in some ways similar to social competence, except that my use of the construct is inclusive of the social as one instance of the domain covered by the intellective.

Teaching, learning, and assessment are dialectical and transactive components of the pedagogical process, and increasingly these components are viewed as functioning in symbiotic relationships. Although each of these components has an independent history and a separate traditional constituency, they are, perhaps, best viewed as parts of a whole cloth, which are differentially emphasized at different times and for different purposes, but always in the context of the whole fabric. I have sometimes referred to this process as orchestration. In some situations, it can be productive to use assessment data to diagnose and prescribe. In other situations, the purpose may be to hold someone accountable. Members of the Gordon Commission are increasingly persuaded that the primary purpose of assessment in educations should be to inform, as well as improve teaching and learning processes and outcomes. We concur with the position advanced by the National Research Council (Knowing What Students Know) that it can be dysfunctional to have the same assessment instruments and procedures serve multiple purposes. Economical as such practices may appear, assessment procedures used for multiple purposes can be disturbing and destructive to the ends intended to be served. This is especially likely to be the case when high stakes are attached to the use of data from assessment used for multiple purposes.

Appropriate articulation between assessment, teaching, and learning processes requires that the development of instruments and procedures for assessment be informed by an intimate understanding of the processes of teaching and learning. Many recent advances in psychometric technology reflect subtle aspects of instructions and special features of the psychology of subject matter learning. Modern conceptions of teaching, learning, and assessment as components of pedagogy are changing, and in each of these components we see aspects of the other components embedded: teaching is moving toward guided exploration and inquiry; learning is depending more on experience, construction, explication, and reflection; and assessment has begun to incorporate tasks involving problem solving, application, and comparative interpretation. Some of the most effective teaching that I have seen recently has been embedded in assessment probes and assessment situations, while some of the most productive assessments that I have observed have been embedded in curriculum materials and teaching/learning transactions.

Diversity, Excellence and Equity

Increasingly, concern for excellence, equity, and fairness in assessment requires that responsible approaches to educational assessment include attention to the quality of teaching and learning transactions and the sufficiency of learner access to appropriate opportunities to learn. Given the changes in the demographics in the U.S., and the co-mingling of world populations that is associated with globalization, systems of assessment, teaching and learning that are incapable of concurrently addressing the issues of population diversity, equitable opportunities to learn, and the achievement of academic excellence will simply become marginalized in the 21st-century. Assessment, teaching, and learning will – out of necessity – have to be appropriate to the diversity in the populations that must be served, and informative of the teaching and learning processes in which they will be embedded. This may constitute a monumental problem for the education and the education assessment enterprises in the absence of changes in the ways in which the functioning of these enterprises are conceptualized and the values that are currently privileged. Assessment of status for purposes of sorting will need to be turned to the assessment of process and becoming in the interest of the development of capacity. Standardization in achievements and outcomes may have to coexist with systematization of diverse routes to and documentation/measures of ways of adaptation. Epistemologies of cultures, languages and schemata may reveal greater varieties of competent function than the hegemonic standards traditionally honored. Respect for the concurrent valuing of diversity, excellence and equity will demand a shift from assessment OF education toward assessment FOR education.

Potentials – Exploring Emerging Developments in Science and Technology

The challenges of the 21st-century may be even more complex. The Gordon Commission has begun the exploration of emerging developments in the exploitation by human beings of nature, science, technology, and scientific imagination. Thought has been given to the possibilities that reside in the exploration of several domains. Buried amongst these are:

1. The Potential of the Combination of Electronic and Digital information Technologies[10]
    - Electronic management of digital information
    - Amplified and new human abilities and capacities
    - Communication computation and fabrication made possible by digitization
    - Technological sup-plantation of human functions

[10] See Behrens, J. T. Technological Implications for Assessment Ecosystems in http://www.gordoncommission.org/rsc/pdf/behrens_dicerbo_technological_implications_assessment.pdf

2. The Potentials that Reside in Changing Social Relations
    - Changes in human to machine relations
    - Changes in demographic distribution
    - Variations in the characteristics of humans living in closer proximity
3. The Potentials Related to Bio-Chemical and Mechanical Change
    - Electrochemical mechanisms
    - Evolutionary biology
    - Transformations in human energies, as in the transformation of energy by transducers
4. The Potentials of Existential Human Realities
    - Imagination and virtual realities
    - Extra sensory perception and telepathic communication

Gordon Commission member James Gee argues that discussions of teaching, learning, and assessment need to be placed in the context of dramatic changes in our contemporary and future worlds. These changes are fueled by advances in technology and by the multiple interacting social, environmental, economic, global, and conflicting civilizations in our world. This context of change involves emerging technologies whose effects are already being felt ("low-hanging fruit"), and technologies on the horizon that can shape a better or worse future depending on how we prepare now for that future ("high-hanging fruit"). This context of emerging and longer-term possibilities is, unfortunately, not usually an overt part of discussions on school reform or the future of learning and the practice of assessment in schools and in the broader society. In a personal note to me, he has called our attention to a few of the most salient items that compose the context of change relevant to what students should know and be able to do in a 21st-century world. None of these changes are good or bad in and of themselves. All of them hold out both great promises for good, and dangerous perils for ill, depending on how we engage with them:

1. *The Producer/Participant Movement.* Thanks to digital technologies, many more people than ever before are becoming (and demanding to be) makers, participants, and designers, not just consumers and spectators. Everyday people are producing, often collaboratively, media of all sorts, science and knowledge, news, ads, new technologies and businesses, and Internet interest-driven learning communities devoted to almost any topic one can imagine.

2. *The Fab Movement.* Part of the Producer/Participant Movement, the Fab Movement involves 3D printers and extractors that can make anything from human skin to houses and nearly any other physical object one can think of. The Fab Movement erases the barrier between atoms and bits, since 3D reality-capture technologies can digitize an object that can then be digitally changed and "printed" out as a new physical object. In the near future, people will be able readily to print houses for the poor or bombs for terrorism.

3. *The DIY Biology Movement.* Also part of the Producer/Participant Movement, the DIY Biology Movement uses low-cost technologies now available to almost anyone to investigate and redesign cells, viruses, DNA, and other biological materials. DIY biologists are seeking cures for cancer in their homes, but also redesigning viruses that could have good or ill effects.

4. *The Amateur-Expert Phenomenon.* Also connected to the Producer/Participant Movement, today amateurs can use the Internet and readily available technologies to compete with and sometimes out-compete experts in a great many domains. Credentials mean much less than they used to.

5. *Big Data.* New technologies have allowed for the collection of massive amounts of data of all sorts and its use in real time, across time, and after action for learning, knowledge building, and successful action for individuals, groups, institutions, and society at large. Data-collecting devices are being incorporated into objects and even people's bodies, allowing people to plan and act in their daily lives based on copious data.

6. *The Dangerous Expert Effect.* Big Data and recent research have shown that credentialed experts in a great many domains make very poor predictions (no better than chance), and that their predictions get worse, not better, when they get more data. Such experts often undervalue what they don't know, overvalue what they do know, and look at data through often unwarranted generalizations to which they are professionally attached. Networked groups of people and tools, using diverse perspectives, make better predictions.

7. *Crowd Sourcing and Collective Intelligence.* Thanks to the failures of narrowly focused experts (like economists in terms of the 2008 recession), there has been, in science and business, a push towards systems of collective intelligence that network diverse points of view from experts and amateurs in different fields with knowledge stored in smart tools and technologies.

8. *Jobs.* Changes in technology—for example, in generalized robots that can be programmed to carry out different functions, and in tools for digital fabricating—look like they will soon remove the low labor-cost advantage that led to out-sourcing and the temporary success of countries like China. They will dramatically change the nature of work, the types of skills needed for success, and the types (and number) of jobs available. Many new businesses will leverage consumers and digital tools rather than workers for design and production.

9. *Longer Lives.* New research in biology and new technologies—for example, digitally designing new viruses and new forms of life—hold out the possibility of greatly extending human life, some claim even to a form of "immortality." In an already crowded world, this is good news for individuals, but, perhaps, bad news for the world.

10. *Growing Inequality.* Inequality between the rich and the poor is growing ever greater in the United State and across the world. In the United States, inequality is as bad or worse than it was in the 1890s, the Age of the Robber Barons. Class has, for the first time, passed race in terms of educational gaps. Research has clearly shown that high levels of inequality in a society lead to poor levels of health and high levels of social problems for both the rich and poor in the society.

11. *New Technologies for Solving our Major Problems*. New technologies are emerging and on the horizon that have the potential to actually solve some of our most serious problems, problems such as global warming, public health, environmental degradation, energy consumption, and housing for the poor. We hear less about these because of the academic urge to stress disaster and the negative.

12. *Sustainability, Resilience, and Anti-Fragility.* Though technologies are emerging or are on the horizon that can potentially solve our problems, there is evidence that they may come too late. The effects of global warming and other human-environmental interactions are coming so much faster than predicted that there may not be time to leverage new technologies and practices. This has led some people to argue that it is too late for "sustainability" as a goal (which means that people and systems sustain themselves through change). We need to move to either "resilience" (people and systems adapt and transform amidst change) or "anti-fragility" (people or systems are designed actually to get better with change and chaos).

13. Mainstream discussions of school reform mainly frame issues of learning and assessment in terms of a narrow focus on current technological changes (e.g., adaptive technologies and customization) and not more broadly on the interactions between technology and our fast-changing and high-risk global world. Such discussions risk being rendered irrelevant by change and, worse, forestalling the contributions education, learning, and assessment can make to saving our world and making a better long-term future for all.

Purpose and Fidelity to Intent

Considerable attention is given to the purposes of assessment in the work of the Gordon Commission. Why do we make such large investments of time, human effort and material resources in assessment of and for education? In the major resource paper prepared for the Commission, the focus is directed at “purpose drift” or the tendency of assessment instruments and procedures to be used for purposes other than those for which they were developed. However Ho (2012) draws heavily on the literature concerning purposes of assessment which have been the subject of serious attention in the measurements sciences.

There are multiple purposes to be served by assessment in education. These purposes range from accountability and admissions to diagnosis, curriculum design, and guidance. However, as we have noted, assessment for the purpose of facilitating accountability has come to dominate the use of assessment in education in the United States. It is not simply the use of assessment for this purpose, but the use of assessment in a very narrow form, that is, the use of standardized educational achievement tests to make high-stake decisions concerning institutions, students, and teachers. This imbalance or misuse has become so influential that in too many instances, teaching and learning transactions have become misdirected with the result that too much time is directed at the excessive use of poor test preparation practices. We see numerous reports of tests’ data having been changed or misrepresented. In some of these instances it is not as much cheating by students, as it is malpractice by teachers and administrators in response to the penalizing use of test data. We recognize the important functions served by the use of good assessment information for accountability purposes, but we are concerned that the accountability function has become distorted and has over-balanced the other and possibly more important functions of assessment in education. The correction of this imbalance in national and state education policy is thought to be one of the most critical and neglected problems in education as we focus on the future of assessment in education.

The members of the Gordon Commission have considered the possibility that the capacity of assessment to inform curriculum design, as well as guide teaching and learning processes, renders assessment capable of serving a wide range of purposes, including, of course, accountability. In the work of the Gordon Commission, one can find expressions of concern that accountability not be the primary driver of assessment in education. But we also find strong support for the idea that greater balance should be achieved in the privileging of specific of the multiple purposes of assessment, such as in the context of more broadly distributed efforts at the use of assessment in education to improve and inform teaching and learning processes and outcomes. The most radical expression of this idea involved the integration of assessment in the teaching and learning transactions. Less radical examples include assessment probes that are distributed throughout the course or school year. A colloquial expression of this concern can be seen in the discussions of formative and summative assessments. In public hearings and consultative conversations we heard repeated calls for greater balance in the use of assessment to serve the multiple purposes identified for assessment. We were reminded that it is legitimate that assessment data be used to drive school reform and change teacher practice, but that it is better used constructively rather than punitively. We heard calls for greater access to assessment information by teachers and students, and on a more timely basis. Teaching and learning persons were asking that assessment data be made available and interpreted to them in ways that inform and guide what they can do to improve teaching and learning processes and outcomes.

Candidates for Assessment Capacity and Practice by Mid-21st Century

Few of us seem prepared to give up the capability of assessment to determine learner status and to monitor learner progress. I recall a very rich discussion in which accountability and responsibility were conjoined as purposes of assessment that require more complex information than can be provided by a test score. The argument offered was that the assessment should address status, process and their contexts to be useful as explanatory information. Why then do we assess? We assess in order to better understand the people we teach, the processes by which we teach them, the situations in which they learn or fail to do so, and to enhance their intellective character and competence. What then might well be the characteristics of systems of assessment in education that embrace assessment, teaching and learning as privileged processes? My preferred candidates for assessment capacity and practice by mid-21st-century follow:

- I propose a system of inquiring assessment probes, embedded in teaching and learning transactions. There are at least three ideas included in this proposal:
    - The first concerns a gradual replacement of standalone tests with systems of assessment (multiple and varied assessment opportunities), which are distributed over time, and throughout the teaching and learning transaction.
    - The second involves the integration of assessment probes as instruments of inquiry, instruction, modeling, and mediation.
    - The third would separate responsibility for the use of data drawn from rich descriptions of these transactions for administrative, and the use for student development purposes. Teachers would be enabled to interpret these data diagnostically and prescriptively. We psychometricians would be responsible for distilling from these in vivo learning and teaching transactions the data needed for accountability and certification.
- The integration of assessment with teaching and learning will demand a view of assessment as diagnostic inquiry, exploratory mediation, and intensive accountable exchange ("accountable talk" to use Resnick's term). There is a rich history of the use of questioning as a part of instruction. Good teachers know the art of posing questions that stimulate thought (Socratic dialogue) as well as probing for evidence of status or process. Most good teachers do not depend solely on standardized tests to know where their students are and what they need. Whimby (1980) makes extensive use of exploratory mediation through which teacher and student inquiry are used in the search for explication of meaning and processes utilized. In the integration of assessment with teaching and learning, the unique character of each of these processes may be lost, as each serve functions that can be interchanged with the other.
- The unbundling and explication of the cognitive demands of knowledge and technique mastery. What is the cargo of transfer learning? I have already given extensive discussion to my concern for the complementarities between the worlds of knowledge and technique on one hand, and developing mental capacities on the other. I have also discussed the possibilities for distilling from the items of standardized test clearer indices to the cognitive demands of test items. In this approach, we recognize the importance of knowledge content in teaching and learning, but I argue that the mastery of such content may be less important than is the achievement of intentional command of the mental abilities that one (1), have been developed in the course of the study of this content and two (2), are essential to the processing of information represented in the knowledge and technique.

- Modern information technologies afford students access to almost limitless quantities and varieties of information resources. Competence in accessing and utilizing available resources could replace the more traditional privileging of memory store. Assessment and education by mid-21st century will be capable of documenting and determining the status of one's competence in determining resource need, accessing needed resources, help seeking, and the utilization of these resources.
- Distance learning and the use of epistemic games have already reached epidemic levels among age groups of learners under thirty. Current predictions suggest continued growth in the use of these educative and recreational media. The almost colloquial anticipation is that this genre of electronic digital information exchange carries with it a trove of information that can be used for educational purposes. In the near future such information will be distilled from the records of these transactions, even as the genre gains in sophistication relative to its capacity to generate useful information. The assessment challenge will be the systematization of relevant indicators as well as the data distillation techniques utilized.
- I like to think of the digital and electronic technologies as amplifiers of human abilities; however, these technologies do not simply enhance the existing human abilities, they appear to have the potential for creating new human capacities. Future assessments in education will need to be capable of documenting human abilities in their amplified state as well as these newly emerging human capabilities. Even at this time we can anticipate increasing demands for abilities that relate to adaptation to randomization: pattern recognition and generation of patterns; rationalization of contradictions; the adjudication of relational paradoxes; and the capacity for virtual problem solving.
- In the 20th century, testing and measurement of developed abilities dominated assessment. In the 21st century, assessment for the development of human capacities will be the demand. Assessments in that new age will need to be diagnostic, prescriptive, instructive and capable of documenting what exists, capturing the processes by which abilities are developing, and modeling the achievements that are the ends of assessment, teaching and learning. Assessments will continue to be conducted and interpreted by the professional other, but assessment will also be ubiquitously conducted by oneself and layperson others, in what Torre and Sampson (2012) describe as cultures of assessment, where evidentiary reasoning will become a colloquial basis for action, based on data that are ubiquitously generated in commerce, in life, in play, in study and in work.

Projected Process-Analytic Function for Assessment

I feel strongly that traditional practice in measurement has focused too much on the measurement of respondents' relative status with respect to knowledge and skills mastery, and insufficiently on the nature of the processes involved in their learning and the teaching that enables it. Yet, it is these processes of learning and the directed engagement of students' mental abilities and dispositions in the study of relevant knowledge-skills content, with which teachers should be concerned, and which should be the focus of assessment activity. Content mastery in teaching, as well as in measurement, should not be ignored. But analyzing and documenting the processes by which they are engaged will be the likely source of our enhanced capacity to inform and improve teaching and learning. Emerging developments in our understanding of the nature of knowledge, what it means to know, and in the education relevant sciences and technologies, increasingly, will enable and demand that attention be given to this projected process-analytic function for assessment in education.

Consider for a moment my conception of pedagogy—assessment, teaching, and learning as interrelated processes of inquiry; as exercises in the collection of information relevant for understanding human performance; as involving the explication, mediation and modeling of information; and as the thoughtful engagement with information (knowledge and technique) for the purpose of enhanced understanding to inform action directed at the facilitation of learning and development. There are two implicit moral obligations in this theoretical model. First, there is the moral obligation to seek knowledge and understanding. Intentional human action should be informed. I also consider that we have the moral obligation to act on the basis of one's informed understanding. Education is one domain of human activity in which this moral imperative is essential. Directed learning and development demands guidance from the best and most complete information available. In most of the work of the Gordon Commission we have elaborated an essentially epistemological rationale for new directions in our approach to assessment, but there is also a deontic rationale, which may be even more powerful than the epistemological. If the intent in assessment in education is to inform and improve teaching and learning, the moral obligation is to generate, interpret and make available the relevant evidence that is necessary for intervention as action on this enabled understanding.

"Those who have the privilege to know, have the duty to act." Albert Einstein

References

AERA-APA-NCME *Standards for Educational and Psychological Testing, W.P. Fisher, Jr. Rasch Measurement Transactions, 2011, 24:4, 1310.*

Allalouf, A. and Sireci, S. G. (2012), Guest Editorial: Dissemination of measurement concepts and knowledge to the public. *Educational Measurement: Issues and Practice,* 31: 1. doi: 10.1111/j.1745-3992.2012.00227.x.

Anastasi, A., & Urbina, A. (1997). *Psychological Testing* (7th ed.). Upper Saddle River, NJ: Prentice.

Andersen, Chris (2012). *Makers: The New Industrial Revolution*. New York: Crown Business.

Armour-Thomas, E. and Gordon I, E. W. (2012). Toward an Understanding of Assessment as a Dynamic Component of Pedagogy. *The Gordon Commission on the Future of Assessment in Education*. Retrieved January 13, 2013, from http://www.gordoncommission.org/rsc/pdf/armour_thomas_gordon_understanding_assessment.pdf

Behrens, J.T., and DiCerbo, K. (2012). Technological Implications for Assessment Ecosystems: Opportunities for Digital Technology to Advance Assessment. *The Gordon Commission on the Future of Assessment in Education*. Retrieved January 13, 2013, from http://www.gordoncommission.org/rsc/pdf/behrens_dicerbo_technological_implications_assessment.pdf

Bereiter C. and Scardamalia, M.(2012). What Will It Mean to Be an Educated Person in Mid-21st-Century? *The Gordon Commission on the Future of Assessment in Education*. Retrieved January 13, 2013, from http://www.gordoncommission.org/rsc/pdf/bereiter_scardamalia_educated_person_mid21st_century.pdf

Cauce, A. M. and Gordon I, E.W. (2012). Toward the Measurement of Human Agency and the Disposition to Express It. *The Gordon Commission on the Future of Assessment in Education*. Retrieved January 13, 2013, from http://www.gordoncommission.org/rsc/pdf/cauce_gordon_measurement_human_agency.pdf

Chung, G. (2012). Toward the Relational Management of Educational Measurement Data. *The Gordon Commission on the Future of Assessment in Education*. Retrieved January 13, 2013, from http://www.gordoncommission.org/rsc/pdf/chung_toward_relational_management_educational_measurement.pdf

Church, George, M. & Regis, Ed. (2012). *Regenesis: How Synthetic Biology Will Reinvent Nature and Ourselves*. New York: Basic Books.

Committee on Defining Deeper Learning and 21st-Century Skills. (2012). "Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st-Century." National Science Research Council.

Diamandis, Peter H. & Kotler, Steven (2012). *Abundance: The Future is Better Than You Think*. New York: Free Press.

Dixon-Román E. and Gergen, K. (2012). Epistemology and Measurement: Paradigms and Practices – Part I. A Critical Perspective on the Sciences of Measurement. *The Gordon Commission on the Future of Assessment in Education*. Retrieved January 13, 2013, from http://www.gordoncommission.org/rsc/pdf/dixonroman_gergen_epistemology_measurement_paradigms_practices_1.pdf

Gee, J. P. (2013). *The Anti-Education Era: Creating Smarter Students Through Digital Learning*. New York: Palgrave/Macmillan.

Gershenfeld, Neil (2007). *Fab: The Coming Revolution on Your Desktop—From Personal Computers to Personal Fabrication*. New York: Basic Books.

Gordon, E.W., and B. L. Bridglall, eds. (2007). *Affirmative Development: Cultivating Academic Ability*. Lanham, MD: Rowman & Littlefield.

Gordon, E. W. (2007). Intellective competence. *Voices in Urban Education*, 14, 7-10.

Gorin, J. (2012). Assessment as Evidential Reasoning. *The Gordon Commission on the Future of Assessment in Education*. Retrieved January 13, 2013, from http://www.gordoncommission.org/rsc/pdfs/gorin_assessment_evidential_reasoning.pdf

Hakuta, K. (2012). Assessment of Content and Language in Light of the New Standards: Challenges and Opportunities for English Language Learners. *The Gordon Commission on the Future of Assessment in Education*. Retrieved January 13, 2013, from http://www.gordoncommission.org/rsc/pdf/hakuta_assessment_content_language_standards_challenges_opportunities.pdf

Hill, C. (2012). Assessment in the Service of Teaching and Learning. *The Gordon Commission on the Future of Assessment in Education*. Retrieved January 13, 2013, from http://www.gordoncommission.org/rsc/pdf/hill_assessment_service_teaching_learning.pdf

Hitt, Jack (2013). *Bunch of Amateurs: Inside America's Hidden World of Inventors, Tinkerers, and Job Creators*. New York: Broadway.

Ho, A., (2012). Variety and Drift in the Functions and Purposes of Assessment in K–12 Education. *The Gordon Commission on the Future of Assessment in Education*. Retrieved January 13, 2013, from http://www.gordoncommission.org/rsc/pdf/ho_variety_drift_functions_purposes_assessment_k12.pdf

Jenkins, Henry (2006). *Convergence Culture: Where Old and New Media Collide*. New York: New York University Press.

Joseph E. Stiglitz (2012). *The Price of Inequality: How Today's Divided Society Endangers Our Future*. New York: Norton.

Kaestle, C. (2012). Testing Policy in the United States: A Historical Perspective. *The Gordon Commission on the Future of Assessment in Education*. Retrieved January 13, 2013, from http://www.gordoncommission.org/rsc/pdf/kaestle_testing_policy_us_historical_perspective.pdf

Linn, R. L. (2012). Test-Based Accountability. *The Gordon Commission on the Future of Assessment in Education*. Retrieved January 13, 2013, from http://www.gordoncommission.org/rsc/pdf/linn_test_based_accountability.pdf

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

Mendoza-Denton, R. (2012). A Social Psychological Perspective on the Achievement Gap in Standardized Test Performance Between White and Minority Students: Implications for Assessment. *The Gordon Commission on the Future of Assessment in Education*. Retrieved January 13, 2013, from http://www.gordoncommission.org/rsc/pdf/mendoza_denton_social_psychological_perspective_achievement_gap.pdf

Meroe, A. S. (2012). Democracy, Meritocracy and the Uses of Education. *The Gordon Commission on the Future of Assessment in Education*. Retrieved January 13, 2013, from http://www.gordoncommission.org/rsc/pdf/meroe_democracy_meritocracy_uses_education.pdf

Michaels, S., O'Connor, C., & Resnick, L. B. (2008). Deliberative discourse idealized and realized: Accountable talk in the classroom and in civic life. *Studies in Philosophy and Education, 27*(4), 283-297.

Mislevy, R. (2012). Four Metaphors We Need to Understand Assessment. *The Gordon Commission on the Future of Assessment in Education*. Retrieved January 13, 2013, from http://www.gordoncommission.org/rsc/pdf/mislevy_four_metaphors_understand_assessment.pdf

Nassim, Nicholas Taleb (2012). *Antifragile: Things That Gain From Disorder*. New York: Random House.

Nielsen, Michael (2012). *Reinventing Discovery: The Era of Networked Science*. Princeton, NJ: Princeton University Press.

Pickett, K., & Wilkinson, R. (2011). *The Spirit Level: Why Greater Equality Makes Societies Stronger*. New York: Bloomsbury Press.

Resnick, L. B., Michaels, S., & O'Connor, C. (2010). How (well-structured) talk builds the mind. In D. Preiss & R. Sternberg (Eds.), *Innovations in educational psychology: Perspectives on learning, teaching and human development* (pp. 163-194): Springer.

Shirky, Clay (2010). *Cognitive Surplus: How Technology Makes Consumers Into Collaborators*. New York: Penguin.

Silver, Nate (2012). *The Signal and the Noise.* New York: Penguin.

Sireci, S.G., and Forte, E. (2012), Informing in the Information Age: How to communicate measurement concepts to education policymakers. *Educational Measurement: Issues and Practice, 31(2) 27-30.*

Smolan, Rick & Erwitt, Jennifer (2012). *The Human Face of Big Data.* New York: Against All Odds Production.

Snow, Richard E. (1996). Psychology, Public Policy, and Law, Vol 2(3-4), Sep-Dec, 536-560.

Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of human intelligence.* Cambridge: Cambridge University Press.

Thurlow, M. (2012). Accommodation for Challenge, Diversity and Variance in Human Characteristics. *The Gordon Commission on the Future of Assessment in Education*. Retrieved January 13, 2013, from http://www.gordoncommission.org/rsc/pdf/thurlow_accommodation_challenge_diversity_variance.pdf

van Barneveld, A. Arnold, K. E. and Campbell, J.P. (2012). Analytics in Higher Education: Establishing a Common Language. *Educase: Learning Initiative*. http://net.educause.edu/ir/library/pdf/ELI3026.pdf

Varenne, H. (2012). Education: Constraints and Possibilities in Imagining New Ways to Assess Rights, Duties and Privileges. *The Gordon Commission on the Future of Assessment in Education*. Retrieved January 13, 2013, from http://www.gordoncommission.org/rsc/pdf/varenne_education_constraints_possibilities.pdf

Weinberger, David (2012). *Too Big to Know: Rethinking Knowledge Now That the Facts Aren't the Facts, Experts Are Everywhere, and the Smartest Person in the Room Is the Room*. New York: Basic Books.

Whimbey, A., &Lochhead, J. (1980). *Problem solving and comprehension: A short course in analytical reasoning*. Philadelphia: Franklin Institute Press.

Wohlson, Marcus (2011). Biopunk: *How DIY Scientists Hack the Software of Life.* New York: Penguin.

Zolli, Andrew & Healy, Ann Marie (2012). *Resilience: Why Things Bounce Back.* New York: Free Press.

9. THE FINDINGS AND RECOMMENDATIONS OF THE GORDON COMMISSION

The members of the Gordon Commission have not met formally to deliberate concerning findings and recommendations that can be drawn from the work of the Commission. The Co-chairpersons of the Commission, however, have agreed on the following conclusions on findings and recommendations that are grounded in the consultations, deliberations, and commissioned papers conducted by the Gordon Commission. Edmund W. Gordon and James W. Pellegrino have concluded that the findings and recommendations of the Commission can be summarized as follows:

FINDINGS

Nature of Assessment

1. Assessment is a process of knowledge production directed at the generation of inferences concerning developed competencies, the processes by which such competencies are developed, and the potential for their development.

2. Assessment is best structured as a coordinated system focused on the collection of relevant evidence that can be used to support various inferences about human competencies. Based on human judgment and interpretation, the evidence and inferences can be used to inform and improve the processes and outcomes of teaching and learning.

Assessment Purposes and Uses

3. The Gordon Commission recognizes a difference between a) assessment OF educational outcomes, as is reflected in the use of assessment for accountability and evaluation, and b) assessment FOR teaching and learning, as is reflected in its use for diagnosis and intervention. In both manifestations the evidence obtained should be valid and fair for those assessed and the results should contribute to the betterment of educational systems and practices.

4. Assessment can serve multiple purposes for education. Some purposes require precise measurement of the status of specific characteristics while other purposes require the analysis and documentation of teaching, learning and developmental

processes. In all cases, assessment instruments and procedures should not be used for purposes other than those for which they have been designed and for which appropriate validation evidence has been obtained.

5. Assessment in education will of necessity be used to serve multiple purposes. In these several usages we are challenged to achieve and maintain balance such that a single purpose, such as accountability, does not so dominate practice as to preclude the development and use of assessments for other purposes and/or distort the pursuit of the legitimate goals of education.

Assessment Constructs

6. The targets of assessment in education are shifting from the privileging of indicators of a respondent's mastery of declarative and procedural knowledge, toward the inclusion of indicators of respondent's command of access to and use of his/her mental capacities in the processing of knowledge to interpret information and use it to approach solutions to ordinary and novel problems.

7. The privileged focus on the measurement of the status of specific characteristics and performance capacities, increasingly, must be shared with the documentation of the processes by which performance is engaged, the quality with which it is achieved and the conditional correlates associated with the production of the performance.

8. Assessment theory, instrumentation and practice will be required to give parallel attention to the traditional notion concerning intellect as a property of the individual and intellect as a function of social interactions – individual and distributive conceptions of knowledge – personal and collegial proprietary knowledge.

9. The field of assessment in education will need to develop theories and models of interactions between contexts and/or situations and human performance to complement extant theories and models of isolated and static psychological constructs, even as the field develops more advanced theories of dialectically interacting and dynamic biosocial behavioral constructs.

10. Emerging developments in the sciences and technologies have the capacity to amplify human abilities such that education for and assessment of capacities like recall, selective comparison, relational identification, computation, etc., will become superfluous, freeing up intellectual energy for the development and refinement of other human capacities, some of which may be at present beyond human recognition.

Assessment Practices

11. The causes and manifestations of intellectual behavior are pluralistic, requiring that the assessment of intellectual behavior also be pluralistic, i.e., conducted from multiple perspectives, by multiple means, at distributed times and focused on several different indicators of the characteristics of the subject(s) of the assessment.

12. Traditional values associated with educational measurement, such as, reliability, validity, and fairness, may require reconceptualization to accommodate changing conditions, conceptions, epistemologies, demands and purposes.

13. Rapidly emerging capacities in digital information technologies will make possible several expanded opportunities of interest to education and its assessment. Among these are:

    a. Individual and mass personalization of assessment and learning experiences;

    b. Customization to the requirements of challenged, culturally and linguistically different and otherwise diverse populations; and

    c. The relational analysis and management of educational and personal data to inform and improve teaching and learning.

RECOMMENDATIONS DRAWN FROM THE WORK OF THE GORDON COMMISSION

The members of the Commission recognize that the future of assessment will be influenced by what the R&D and the assessment production communities generate as instruments and procedures for the assessment in education enterprise. However, we are very much aware that equally determinative of the future will be the judgments and preferences of the policymakers who decide what will be required and what practitioners and the public will expect. In recognition of the crucial role played by policymakers, the Executive Council of the Gordon Commission has given special attention to the development of a policy statement that concludes with three recommendations directed at those who make policy concerning education and its assessment. The statement has been prepared by James Pellegrino, Co-chair of the Commission, and Lauren Resnick, member of the Executive Council, with input from Sharon Lynn Kagan, consultant to the Chair, and other members of the Executive Council — Randy Bennett, Eva Baker, Bob Mislevy, Lorrie Shepard, Louis Gomez and Edmund W. Gordon — and the assistance of Richard Colvin as writing consultant.

This Public Policy statement represents the authors' sense of recommendations that are implicit in the work of the Commission. However, it has not been vetted by the members of the Gordon Commission and thus it should not be concluded that any given member of the Commission endorses the specifics included herein.

A STATEMENT CONCERNING PUBLIC POLICY

Introduction

The Gordon Commission on the Future of Assessment in Education was created to consider the nature and content of American education during the 21st-century and how assessment can be used most effectively to advance that vision by serving the educational and informational needs of students, teachers and society. The Commission's goal in issuing this brief public policy statement[11] is to stimulate a productive national conversation about assessment and its relationship to learning. The work of the Commission and this report come at a propitious time. The Common Core State Standards in Mathematics and English Language Arts adopted by 45 states and the District of Columbia, as well as Next Generation Science Standards that are under development, stress problem solving, creativity and critical thinking over the memorization of isolated facts and decontextualized skills. Assessments meant to embody and reinforce those standards are under development and will be given for the first time in 2015. Over the next few years states will be deeply engaged in implementing the standards and preparing for the new assessments. These developments have heightened awareness among educators and state and federal policymakers of the critical relationships among more rigorous standards, curriculum, instruction and appropriate assessment, and have created an opportunity to address issues of long standing. This policy statement capitalizes on that opportunity to bring about a fundamental reconceptualization of the purposes of educational assessments.

Transforming Assessment to Support Teaching, Learning and Human Development

Although assessment, broadly construed, is a central element of education and must be aligned with both teaching and learning goals, it is not the only — or even the major — tool for improving student outcomes. Indeed, for education to be effective, schools must be designed with clear and precise teaching and learning goals in mind and supported in ways that make them likely to reach those goals; teachers must be provided with the

[11] *This Public Policy statement represents the authors' sense of recommendations that are implicit in the work of the Commission. However, it has not been vetted by the members of the Gordon Commission and thus it should not be concluded that any given member of the Commission endorses the specifics included herein.*

appropriate instructional materials and professional development; and other resources including time, technology and teachers' skills must be deployed strategically.

To be helpful in achieving the learning goals laid out in the Common Core, assessments must fully represent the competencies that the increasingly complex and changing world demands. The best assessments can accelerate the acquisition of these competencies if they guide the actions of teachers and enable students to gauge their progress. To do so, the tasks and activities in the assessments must be models worthy of the attention and energy of teachers and students. The Commission calls on policymakers at all levels to actively promote this badly needed transformation in current assessment practice. The first and most important step in the right direction will require a fundamental shift in thinking about the purposes of assessment. Throughout the long history of educational assessment in the United States, it has been seen by policymakers as a means of enforcing accountability for the performance of teachers and schools. For a relatively low outlay, assessments could expose academic weaknesses and make it possible to pressure schools and teachers to improve. But, as long as that remains their primary purpose, assessments will never fully realize their potential to guide and inform teaching and learning. Accountability is not the problem. The problem is that other purposes of assessment, such as providing instructionally relevant feedback to teachers and students, get lost when the sole goal of states is to use them to obtain an estimate of how much students have learned in the course of a year. It is critical that the nation's leaders recognize that there are multiple purposes of assessment and that a better balance must be struck among them. The country must invest in the development of new types of assessments that work together in synergistic ways to effectively accomplish these different purposes — in essence, systems of assessment. Those systems must include tools that provide teachers with actionable information about their students and their practice in real time. We must also assure that, in serving accountability purposes, assessments external to the classroom will be designed and used to support high-quality education. Finally, the nation must create a demand for improved assessment practices by helping parents and educators understand the need for change.

The transformation of assessment will require a long-term commitment. There will be some who will argue that, with the work of the PARCC and Smarter Balanced state consortia to create new assessment systems not yet complete, it would be better to wait before pursuing major policy changes. The Commission disagrees and believes that *because* that work is unfinished, now is the time to move toward more fundamental changes. Certainly, the new assessment systems will need to be implemented and analyzed and then—based on data—revised, to be sure that they are, indeed, supportive of the standards. The fundamental reconceptualization of assessment systems that the Commission is calling for should guide those inquiries. The states leading the consortia must demand that the assessment systems be robust enough to drive the instructional changes required to meet the standards. In addition, states have to expect that the assessment systems will provide evidence of student learning that is useful to teachers. Finally, states have to demand that the systems be flexible enough to be adapted to new methods of delivery and scoring as they emerge. As of now, the funding for the consortia will run out in 2014, just as the new assessment systems are starting to be used, and the costs will likely be shifted to the states. The states will have a financial as well as educational incentive to make sure the assessment systems are working as intended.

Consistent with the above, the leadership of the Gordon Commission has developed a set of recommendations directed toward federal and state policymakers; private for-profit and nonprofit organizations related to assessment; the scholarly community; and philanthropists. As a context for these recommendations, we briefly summarize major themes that emerged from meetings that the Commission held across the country as well as reviews and syntheses of research regarding assessment history, methods, philosophy, digital technology and policy.

Reconsidering Assessment: Why, What and How We Assess

The purposes of assessment fall into two general categories: first, assessment *of* learning generally involves an appraisal of student achievement after a period of instruction. Such assessments can be used to judge attainment for such purposes as accountability, admission to college or other opportunities, and to evaluate programs or approaches. Second, assessment *for* learning involves a more restricted and focused appraisal of student knowledge during a shorter period. It is designed for purposes such as adjusting and improving instruction. Although both types of assessment share certain features, they each must be tailored to their specific purpose; an assessment designed for one purpose, such as accountability, is seldom best suited for other purposes such as instructional adjustment.

Recognizing that accountability will continue to be an important aspect of educational policy, the Gordon Commission believes that accountability must be achieved in a way that supports high-quality teaching and learning. It must be remembered that, at their core, educational assessments are statements about what educators, state policymakers and, indirectly, parents want their students to learn and — in a larger sense — become. What we choose to assess is what will end up being the focus of classroom instruction. Teachers and students will take their cues from high-stakes tests and will try to score well on them regardless of their type. So, it is critical that the tests best represent the kind of learning students will need to thrive in the world that awaits them beyond graduation.

But changing the nature and quality of external accountability tests will not be enough. An equal, if not greater, investment needs to be made in new assessment resources and tools that better integrate assessment with classroom teaching and learning, and better represent current thinking on how students learn and on changes in the world at large. The globalization of the economy, advancements in technology, the development of the Internet, and the explosion of social media and other communication platforms have changed the nature of what it means to be well educated and competent in the 21st century. Digital technologies have empowered individuals in multiple ways, enabling them to express themselves, gather information easily, make informed choices, and organize themselves into networks for a variety of purposes. New assessments — both external and internal to classroom use — must fit squarely into this landscape of the future, both signaling what is important and helping learners know they are making progress toward productive citizenry.

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

Designing and implementing assessments that support this ambitious vision of education represents a major challenge. Historically, educational assessments have been far more narrowly focused. Assessments have been designed primarily to provide summative information about student, teacher, school and system performance. That information has been used to highlight weaknesses, direct the spending of money, choose students for additional help or advanced classes, and evaluate the effectiveness of programs or teaching methods. Present testing practices enjoy broad support among policymakers because many people accept them as defining educational accomplishment. But this emphasis on measuring student performance at a single point in time and with assessments whose primary purpose is to provide information to constituencies external to the classroom has, to a large extent, neglected the other purposes of assessment. Moreover, developing a new mindset about the contexts and purposes for assessment, as well as new approaches to accomplish it, is not only difficult, but requires an investment of resources. Presently, the federal government is absorbing the lion's share of the costs for the systems of assessment being developed by the PARCC and Smarter Balanced consortia. The conditions of that support stipulate that accountability components be the primary focus of their work. As a result, it is highly likely that the tools and resources needed to support teacher uses of assessment in the classroom will be seriously underdeveloped and in need of significant further work. When this round of federal funding ends, and the states are left with the challenges and costs associated with implementation and further development of accountability systems, there may be little money remaining to devote to formative assessment and practices.

Moving Forward: The Opportunity

Because assessments are, essentially, a claim about a student's competencies, new approaches to assessment must be treated as a process of gathering evidence to confirm or disprove particular claims. That evidence, which in a system of assessments can come from multiple sources, can be used to improve both how and what students are learning. The evidence might include activities ranging from simple to complex performance tasks pursued within classrooms as well as assessments external to regular classroom activities.

Digital technologies hold great promise for helping to bring about many of the changes in assessment that the Commission believes are necessary. Technologies available today and innovations on the immediate horizon can be used to access information, create simulations and scenarios, allow students to engage in learning games and other activities, and enable collaboration among students. Such activities make it possible to observe, document and assess students' work as they are engaged in natural activities—perhaps reducing the need to separate formal assessment for accountability from learning in the moment. Technologies certainly will make possible the greater use of formative assessment that, in turn, has been shown to significantly impact student achievement. Digital activities also may provide information about noncognitive abilities—such as persistence, creativity and teamwork—that current testing approaches cannot. Juxtaposed with the promise is the need for considerable work to be done on issues of scoring and interpretation of evidence before such embedded assessment can be useful for these varied purposes.

Many issues, including some alluded to above, have been discussed and debated among educators and assessment experts for many years. As part of those discussions it is now widely recognized that large-scale standardized testing has exerted a greater and greater influence over American schooling. At the same time, it has been shown repeatedly that teachers have the largest impact on education of any in-school factor. And it is what teachers do and what they teach and how they assess in classrooms that give teachers that influence. Given that fact, it would seem appropriate to identify specific, effective instructional resources such as curricula and classroom assessments and then prepare teachers to use those resources effectively. However, the notion that education must be locally controlled is deeply engrained in our nation's culture and educational politics and that fact has meant that instructional resources must be chosen by those closest to the classrooms, which sometimes means individual teachers. So, states have individually relied on external tests to exemplify and enforce their content standards so as to ensure some degree of consistency of quality and results across classrooms, schools and districts in their jurisdiction. External tests, then, have too often become the de facto curriculum with a range of intended and unintended outcomes, such as impoverishing the development and use of effective classroom assessments. The Common Core standards, and the rethinking of assessment that they are fostering, provide an opportunity to challenge this deeply held belief in local control.

Recommendations

In the Realm of State Collaboration and Policy

The constitution of every state in the nation requires it to provide a free public education to its children. That means that states have the most authority over the assessments used to monitor the quality of the education children are receiving. Although the past several decades have seen some power and authority over schooling and assessment shift to the federal government, this trend is now in the other direction. The states, acting through the National Governors Association and the Council of Chief State School Officers, demonstrated that they recognized the need for better standards and assessments when they led the creation and adoption of the Common Core State Standards. Although the two assessment consortia are federally funded, they are led by the states. The states participating in the consortia have agreed to establish common progress categories. This record of collaboration is something to build upon. Most state education departments are understaffed and poorly funded. That means that taking on the additional responsibility of monitoring how well the assessments are working will be difficult for them to accomplish on their own. They will have an incentive to continue to work together on this important job.

It is recommended that states create a permanent Council on Educational Assessments modeled on the Education Commission of the States to take on this function. Funding for the Council should come from the federal government, states, and a small tax on every assessment sold.

The Council's first responsibility would be to commission an evaluation of the strengths and weaknesses of the Smarter Balanced and PARCC assessment systems and their effect on teaching and learning. The purpose of this evaluation would be to ensure that the new assessments are, indeed, driving instruction that is consistent with the educational vision embodied in the standards. As has been done before with evaluations of important assessment programs such as the National Assessment of Educational Progress (NAEP), such an evaluation might be conducted by an independent panel assembled under the auspices of the National Academy of Sciences or the National Academy of Education.

In addition, the Council should:

- Conduct research on how assessments are changing, help inform states so that they make good purchasing decisions, and address issues as they arise. The Council also would oversee the process of setting cross-state performance level targets.
- Mount a public education campaign targeting parents, educators, school board members and the media explaining the importance of good assessment to quality education.
- Create a Study Group on the Challenges of Equitable Assessment to explore issues related to diversity, equity and excellence.

- Commission research on policies designed to secure the privacy of assessment data while also creating protocols for making large quantities of such data available to qualified researchers.

In the Realm of Federal Policy

Significant pieces of federal educational legislation are awaiting reauthorization, including the Elementary and Secondary Education Act (ESEA) of 2002, the Individuals with Disabilities Education Act, and the Higher Education Act. The reauthorization of these major pieces of legislation provides an opportunity to promote new ideas about assessment. The Obama administration has successfully used incentives built into the American Recovery and Reinvestment Act of 2009, the *Race to the Top* competitions and the Investing in Innovation fund to bring about a variety of policy changes and innovations. For example, the *Race to the Top* district competition requires applicants to use “collaborative, data-based strategies and 21st-century tools” to move beyond one-size-fits-all approaches and personalize learning. This has significant implications for assessments and the type of feedback they provide for teachers and learners. The U.S. Department of Education has used its waiver powers to allow states to experiment with measuring students’ year-to-year growth rather than their status at a fixed point in time. This waiver power also was used to free states from some of the onerous accountability aspects of the *No Child Left Behind* act.

It is recommended that the President and Congress consider various models to encourage experimentation with different approaches to assessment and accountability. In reauthorizing ESEA, the Obama administration should press for funds to incentivize states and assessment companies to experiment with radically different forms of assessments, including challenging performance tasks that better represent the learning activities that will help students develop the competencies they will need to succeed in the 21st century.

In the Realm of National Research and Development

The assessments that we will need in the future do not yet exist. The progress made by the PARCC and Smarter Balanced consortia in assessment development, while significant, will be far from what is ultimately needed for either accountability or classroom instructional improvement purposes. This is not a criticism of the Consortia per se but a realistic appraisal of the design constraints and timelines imposed upon their work from the outset. While America certainly can profit from the consortia’s work, the U.S. Department of Education, the Department of Defense, the National Science Foundation, and the National Institute of Child Health and Human Development, in collaboration with the philanthropic community, should commit to a 10-year research and development effort to strengthen the capacity of the U.S. assessment enterprise to broaden the range of behaviors, characteristics and manifestations of achievement and related development that are the targets of assessment in education. This effort should be a partnership between not-for-profit organizations (existing or newly created), the for-profit sector, professional teacher organizations and universities. There are multiple models for this type of public-private research and development effort in biomedicine, defense and other fields.

As discussed earlier, one goal of this effort should be the creation of assessment tasks that exemplify the type of learning that we want to occur in classrooms. Today, teaching to the test is seen as a negative consequence of accountability testing. With the proper assessment tools, it will be easier to encourage teaching to the underlying competencies as standard practice. In order to be practical, new ways of delivering and scoring such assessments will have to be developed. Technologies for presenting rich and varied materials and for capturing and automating the scoring of written responses and other student behaviors currently exist and show promise. But they will need to continue to improve and be adapted for a variety of subjects in order for these new assessments to be widely used for a range of assessment purposes.

This expanded view of assessment will require the training and employment of broadly educated specialists in learning, cognition, measurement and assessment. It is recommended that the government and private philanthropies increase the number of pre- and postdoctoral scholars dedicated to the development of this expertise.

General Recommendations Concerning the Future of Assessment in Education

1. As is traditional in the Medical profession and is rapidly embraced as a guide for all professional activity, the recommendation is made that in assessment policy, practice and use of assessment data, this field should “First Do No Harm.” Responsibility for honoring this value falls at multiple levels – policymakers, administrators, staff and perhaps most heavily on the manufacturers of assessment devices and those of us who sell them. (See Ho’s paper on purpose drift).

2. We could declare as consensus among the members of the Commission that

assessment can serve multiple purposes. There is less agreement concerning the possibility that a single test should be so used, however, the consensus holds concerning the need for balance in the attention given to the use of assessment for different purposes. It is recommended that with the possible exception of "informing and improving teaching and learning," no single purpose should be permitted to distort the valued goals of education. Similarly it is recommended that fidelity to the purpose for which the instrument or procedure is designed be honored. This recommendation references, among other concerns, the difference between our traditional concern with assessment of education and the Commission's emphasis on assessment for education.

3. Assessment in education is essentially grounded in inferential reasoning. It is a process by which evidences collected for the purpose of the disconfirmation of inferences one seeks to make concerning the phenomena being assessed. It is therefore recommended that assessment processes be held to standards similar to those long honored in the tradition of the empirical sciences. However, given the Commission's concern for changing paradigms and shifting epistemologies, it is further recommended that the universal utility of positivist scientific methodologies as a standard for evolving assessment practices be subjected to continuing inquiry.

4. We believe that most members of the Commission embrace concern for differential validities, i.e., the idea that validity may be a relative construct, and that it's relativity must be taken into account in policymaking and practice with respect to assessment in education. It is therefore recommended that the field embrace the notion of differential validities and the imperative that tests of validity be appropriate to the populations and situations in which the construct is being utilized.

5. It is recommended that research and development efforts be intensified around questions related to the implications for assessment in education that flow from questions related to the cargo of learning transfer. Special attention may need to be given to the complementarities between mastery of declarative and procedural knowledge and the intentional command of instrumental mental processes.

6. It is recommended that the targets of assessment in education be broadened to include a wider range of human abilities, ways of adaptation, amplified abilities and human capacities, including those that are the products of exposure to digital electronic technologies.

7. Given the considerable evidence in support of agency, disposition, cultural identities, and existential states as influences on the nature and quality of human performance, it Is recommended that research and development concerning the relationships between human performance and these variables be given considerably greater priority in inquiries concerning assessment in education.

8. Debate continues concerning the idea that intelligence is a characteristic of individuals; intelligence is a collectively produced construct best associated with social groups; and the idea that intelligence originates and is expressed in both contexts. The increased practice of collaboration in the production of knowledge and its application suggests the importance of our recommendation that research and development effort be directed at differentiating assessments to capture intellective competence as a property of individuals and as a function of collaboration between persons.

9. Considerable concern has been expressed in the Commission about the artificiality of "stand-alone" or "drop-in-from-the-sky" tests. Perhaps more problematic than the isolated character of these examinations is concerned with the tendency to treat the data from these tests as independent and sole sources of information concerning the performance and status of students. Some commissioners argued for the greater use of systems of examinations distributed over time embedded in the ongoing teaching and learning of experiences. It is recommended that assessment in education move progressively toward the development and use of diversified assessment systems for the generation and collection of educational assessment data.

10. It is then the final recommendation, implicit in the work of the Gordon Commission, that the academic and philanthropic sectors of the society – cooperatively supported by tax levy funds, consider the creation of a Virtual Institute on the Future of Assessment in Education (VIFAE) to continue the inquiry initiated by the Gordon Commission; to encourage broad and cross-disciplinary collaboration in this work; and to support the attraction to and development of young and new scholars to conceptual, research and development explorations of the relationships between assessment, teaching and learning.

The aim of the VIFAE is to build a study group of scholars to continue inquiry into the changing nature of education and the changing demands on assessment in education that will result from changes in education and society. The MacArthur Foundation has seeded such a virtual Institute with the funding of the Edmund W. Gordon MacArthur

Foundation/ETS Fellowship for 21st-Century Learning and Assessment. The VIFAE would seek to increase the understanding of the issues related to these changes and their implications for assessment in education. At the heart of the VIFAE should be a cadre of senior and junior scholars, in a multi-disciplinary network; working from home sites at institutions across the country, in active communication, and working on different aspects of the common concern with the future of assessment in education. We imagine that the VIFAE will be to the science of assessment what AT&T Bell Laboratories was to communications and information technology.

Several developments in education, in society, and in assessment provide justification for the proposed virtual institute. The epistemologies that inform teaching and learning including their assessment are shifting and the paradigms that inform education practices are changing. The purposes of assessment are multiple and evolving, but appear to be stuck in a conceptual frame that may be limiting, if not distorting, advanced notions of education and progress in the sciences. As a result, assessment and measurement science in the future will require changes in such directions as:

1. The blending or integration of assessment, teaching, and learning;
2. The separation of accountability and certification from diagnostic, prescriptive, and adjudicative functions of assessment;
3. Multicomponential systems of assessment distributed over time and situation;
4. The introduction of contextualism, perspectivism, and situativism in assessment probes and conditional correlates of performance for scoring and consideration in the interpretation of assessment data;
5. The development and identification of what will count as indicators of intellective competence as technological amplifiers of human abilities that enable adaptive capacities more powerful than human memory, relational identification and adjudication, information acquisition, selective comparison, and virtual simulated experience;
6. The incorporation of advanced digital technologies for information management will be ubiquitously represented in general as well as STEM educational at three or more levels of concern:
    a. New representations and presentations of knowledge may have implications for changes in the practices of assessment, teaching, and learning;

b. These technologies present new opportunities for assessment, teaching, and learning and the manner in which these processes are engaged; and

c. Advanced technologies for handling digital information will require improved and different human capabilities and personnel training to gather and interpret these new kinds of data and their presentations, exponential growth in the kinds and quantities of data available, and the ability to comprehend and utilize advances in the meanings of information (principles) made possible by these extensions of human sensory and conceptual abilities;

7. The assessment of new skills such as collaborative problem solving and various "noncognitive" or "soft skills"; and

8. Assessment that will be dynamic and will include a more detailed and complex data collection, such as the contextual and processual analysis of teaching and learning behaviors and situations.

The changes numerated above will call for numerous foundational psychometric advances, so that fair, valid and actionable data can be extracted from complex tasks and simulations; unobtrusive assessments that take place over an extended "real time"; serious games that are also assessments; and other novel forms of assessment. These changes also will require new ways of understanding educational data structures, including:

- Longitudinal data;
- High-dimensional data matrices;
- Detail about teaching and learning processes (scaffolding) as well as products outcomes;
- Mixtures of continuous (e.g., timing) and discrete (correct/incorrect) data; and
- Dynamic models for data collected over time.

10. ABOUT THE GORDON COMMISSION ON THE FUTURE OF ASSESSMENT IN EDUCATION

Commission Background

Conceptions of what it means to educate and to be an educated person are changing. Notions of and demands on practice in the teaching and learning enterprise are broadening and expanding. And the concern with accountability forces this dynamic and eclectic enterprise to constrict and, in the worst of instances, to compromise in the interest of meeting certain accountability criteria. These realities, coupled with changes in epistemology, cognitive and learning sciences, as well as in the pedagogical technologies that inform teaching and learning, are narrowing—possibly even stifling—creativity and flexibility in teaching and learning transactions. These are among the perceived problems that led to the creation of the Gordon Commission on the Future of Assessment in Education.

Although these immediate issues were foundational in the establishment of the Gordon Commission, a second more compelling contextual problem helps to drive its mission. Changing conceptions of and practices in educational assessment are making many of the capabilities of traditional conceptions and practices in educational assessment obsolete. The work of the Commission rests on the assumption that assessment in education can inform and improve teaching and learning processes and outcomes.

Mission of the Commission

The Gordon Commission was created with the mission to study the best of educational assessment policy, practice and technology; consider the best estimates of what education will become and what will be needed from educational measurement during the 21st century; and to generate recommendations on educational assessment design and application that meet and/or exceed the demands and needs of education — present and predicted.

Given the mission of the Gordon Commission, a number of goals were outlined that focused the work of the Commission. The goals of the Gordon Commission are to:

- Inform the field and the public about the need and possibilities for change in education, as well as change in the functions, practices and roles of assessment in education;

- Increase public awareness and knowledge about assessment as an integral component of education and the possibilities for change in assessment practice;
- Encourage the field of educational assessment to strengthen its capacity to factor into measurement practice attention to the influence of human attributes, social contexts and personal identities on human performance;
- Balance emphasis on prediction, selection and accountability with equal concern for informing and improving teaching and learning processes and outcomes; and
- Inform long-term planning and product development in the field of psychometrics.

Chairperson Gordon's Perspectives on Assessment

As part of the work of the Commission, Chairperson Gordon was to lay out some of his existing ideas that he believed would influence his conceptual leadership and the work of the Gordon Commission. He shared these perspectives with the Commissioners, as well as the public at large (Gordon Commission, 2012)[12] and invited an examination and critique of them. The ideas, cited below, provided a starting point for many of the consultative conversations as well as for the continuing seminars of the Gordon Commission Fellows.

- Traditional approaches to testing give too much emphasis to a limited view of the status of a narrow range of cognitive functions, as well as to the neglect of the affective and situative domains of human performance and the processes by which these functions and domains are engaged.
- Current assessment instruments and procedures tend to neglect the diverse contexts and perspective born of different cultural experiences and cultural identities and the influence of these contexts, perspectives, and identities on human performance. However, the most important features of intellective competence may require that the expression of competence be demonstrated independent of such contexts, perspectives, and identities.
- Traditionally, testing has privileged—in its purposes—accountability, prediction, and selection to the neglect of diagnosis, prescription, and the informing and improving of teaching and learning processes and outcomes. I believe that the most important functions and purposes of measurement in education concern informing, as well as improving, teaching and learning processes and outcomes.

[12] The Gordon Commission on the Future of Assessment in Education.(2012). *Assessment, Teaching and Learning Bulletin 2(1).* *Retrieved from* *http://www.gordoncommission.org/publications_reports.html*

- Traditional approaches to assessment have emphasized relative position and competition to the neglect of criterion-based judgments of competence. The meritocratic ideology that dominates in testing may be dysfunctional to developmental democratization, particularly when developmental opportunities are distributed on the basis of prior developmental achievements and when level of prior development may be, in part, a function of the maldistribution of the opportunity to develop, learn, or excel.
- Traditional approaches to assessment privilege knowing, knowing how to, and mastery of veridical knowledge, while intellective competence, emerging epistemologies, and the cohabitation of populations with diverse cultural forms may—increasingly—require multiple ways of knowing, understanding as well as knowing, and the ability to adjudicate competing relationships in our knowledge and in the production of knowledge.
- The pursuit of content mastery should be in the service of the development of mental processes. Michael Martinez's notions in his book, *Education as the Cultivation of Intelligence*, resonate with me. Michael's mentor, the late Richard Snow, left an incomplete idea in which he was developing the argument for the study of content (subject matter) as instrumental to the development of intellect. I am attracted to the notion of the study of any content as a means of nurturing intellect, as well as for the purposes of knowing.
- The term "intellective competence," connotes the effective orchestration of affective, cognitive, and situative processes in the interest of intentional human agency. Affective is placed first to emphasize that human activity appears to begin with affect, and that while cognition ultimately informs affect, it is affect that gives rise to cognitive functions. The primacy of one or the other is not the current issue. Traditional approaches to educational testing have given insufficient attention to the influence of affect on human performance, and this has been done to the disadvantage of the psychometric enterprise. Although affective and situative processes are unstable and messy, attribution, disposition, intentionality, and motivation have an important influence on human performance. They cannot continue to be left out of the calculus of assessment in education. The problem is how they should be included, not whether they should be included.

Commission Members

The Gordon Commission consists of 30 members. The scholars, policymakers and practitioners who comprise the Commission have identified critical issues concerning educational assessment, investigated those issues, and developed position and review papers that informed the Commission's recommendations for policy and practice in educational assessment.

Chairman

Edmund W. Gordon
John M. Musser Professor of Psychology, Emeritus
Yale University
Richard March Hoe Professor of Education and Psychology, Emeritus
Teachers College, Columbia University

Co-Chair

Jim Pellegrino
Liberal Arts & Sciences Distinguished Professor
Distinguished Professor of Education
Co-Director, Learning Sciences Research Institute
University of Illinois at Chicago

Executive Council

Eva Baker
Distinguished Professor, Graduate School of Education and Information Studies, and Director, National Center for Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles

Randy E. Bennett
Norman O. Frederiksen Chair in Assessment Innovation, Educational Testing Service (ETS)

Louis M. Gomez
MacArthur Foundation Chair, Digital Media and Learning, Graduate School of Education & Information Studies, University of California, Los Angeles

Robert J. Mislevy
Frederic M. Lord Chair in Measurement and Statistics, ETS

Lauren Resnick
Senior Scientist, and Project Director, Learning Research and Development Center, and Distinguished University Professor of Psychology and Cognitive Science, University of Pittsburgh

Lorrie A. Shepard
Dean, School of Education, and Professor of Education, University of Colorado at Boulder

Commissioners

J. Lawrence Aber
University Professor and
Albert and Blanche Willner Family Professor of Psychology and Public Policy, Department of Applied, Psychology, Steinhardt School of Education, New York University

Bruce M. Alberts
Professor, Department of Biochemistry and Biophysics, University of California, San Francisco, and Chief Editor, Science Magazine

John Bailey
Director, Dutko Worldwide

John T. Behrens
Vice President, Pearson Center for Digital
Transformation

Ana Mari Cauce
Provost, and Earl R. Carlson Professor of Psychology, University of Washington

Linda Darling-Hammond
Charles Ducommun Professor of Education, and Co-Director, School Redesign Network (SRN), School of Education, Stanford University

Ezekiel Dixon-Román
Assistant Professor, School of Social Work and Social Policy, University of Pennsylvania

James Paul Gee
Mary Lou Fulton Presidential Professor of Literacy Studies, Arizona State University

Kenji Hakuta,
Lee L. Jacks Professor of Education, School of Education, Stanford University

Frederick M. Hess
Resident Scholar and Director of Education Policy Studies, American Enterprise Institute for Public Policy Research

Andrew Ho
Assistant Professor of Education, Graduate School of Education, Harvard University
Freeman A. Hrabowski III
President, University of Maryland, Baltimore County

Michael E. Martinez (1956 – 2012)
Professor, Department of Education, University of California, Irvine

Rodolfo Mendoza-Denton
Associate Professor, Psychology Department, University of California, Berkeley

Shael Polakow-Suransky
Chief Academic Officer and Senior Deputy Chancellor, New York City Department of Education

Diane Ravitch
Research Faculty, Steinhardt School of Culture, Education, and Human Development, NYU

Charlene Rivera
Research Professor, and Executive Director, Center for Equity and Excellence in Education, George Washington University

Lee Shulman
President Emeritus, The Carnegie Foundation for the Advancement of Teaching, and Charles E. Ducommun Professor of Education–Emeritus, School of Education, Stanford University

Elena Silva
Senior Associate, Public Policy Engagement,
Carnegie Foundation for the Advancement of Teaching

Claude Steele
Dean, Graduate School of Education, Stanford University

Ross Wiener
Executive Director, Education and Society Program, The Aspen Institute

Robert Wise
Former U.S. Governor, West Virginia, and President, Alliance for Excellent Education

Constance M. Yowell
Director of Education, The John D. and Catherine T. MacArthur Foundation

Consultants

Through its commitment to influencing the future of assessment in education, the Commission seeks to stimulate a national conversation on possible relationships between assessment, teaching and learning. Toward that end, the Commission consulted with a wide variety of experts, ranging from consumers of tests and test results, to research and development scholars who produce tests and knowledge relevant to assessment, as well as policymakers who determine the broad importance and application of tests.

Consultants to the Chairman

History: **Carl Kaestle**
University Professor and Professor of Education, History, and Public Policy
Education Department
Brown University

Philosophy: **Lucius Outlaw**
Professor of Philosophy
College of Arts and Sciences
Vanderbilt University

Policy: **Sharon Lynn Kagan**
Virginia and Leonard Marx Professor of Early Childhood and Family Policy
Co-Director, National Center for Children & Families
Teachers College, Columbia University
Professor Adjunct, Yale Child Study Center
Yale University

Psychology: **Kenneth Gergen**
Senior Research Professor of Psychology
Swarthmore College

Consultants to the Commission

Jamal Abedi
Professor of Education
School of Education
UC Davis

Russell Almond
Associate Professor
Department of Educational Psychology and Learning Systems
College of Education
Florida State University

Eleanor Armour-Thomas
Professor, Educational Psychology
Queens College, City University of New York

Lloyd Bond
Professor
Department of Educational Research Methodology
University of North Carolina, Greensboro

A. Wade Boykin
Director
Capstone Institute at Howard University

John Bransford
Shauna C. Larson Endowed Professor in Learning Sciences
College of Education
University of Washington

Henry Braun
Boisi Professor of Education and Public Policy
Department of Educational Research, Measurement, and Evaluation
Lynch School of Education
Boston College

Tony Bryk
President
The Carnegie Foundation for the Advancement of Teaching

Li Cai
Co-Director
National Center for Research on Evaluation, Standards, and Student Testing (CRESST)
UCLA Graduate School of Education and Information Studies

Robert Calfee
Professor of Education, Emeritus
University of California, Riverside

Madhabi Chatterjji
Associate Professor of Measurement- Evaluation & Education and
Director, Assessment and Evaluation Research Initiative (AERI)
Teachers College, Columbia University

Greg Chung
Senior Researcher
National Center for Research on Evaluation, Standards, and Student Testing (CRESST)
UCLA Graduate School of Education and Information Studies

Dennis Culhane
Professor
School of Social Policy
University of Pennsylvania

Carol Dweck
Lewis and Virginia Eaton Professor
Department of Psychology
Stanford University

Howard Everson
Professor and Senior Research Fellow
Center for Advanced Study in Education
Graduate School & University Center
City University of New York

John Fantuzzo
Albert M. Greenfield Professor of Human Relations
Graduate School of Education
University of Pennsylvania

Roy Freedle
Research Psychologist

Patricia Gándara
Professor, UCLA Graduate School of Education and Information Sciences;
Co-Director, The Civil Rights Project at UCLA

Angela Glover-Blackwell
President and CEO
PolicyLink

James Greeno
Visiting Professor
School of Education
University of Pittsburgh

Kris Gutiérrez
Professor, Social Research Methodology
Director, Education Minor
UCLA

Edward Haertel
Jacks Family Professor of Education and Associate Dean for Faculty Affairs
Stanford University

David T. Hansen
Professor and Director
Program in Philosophy and Education
Teachers College, Columbia University

Norris Haynes
Professor in the Counseling and School Psychology Department
Southern Connecticut State University

Jeffery Henig
Professor of Political Science and Education and Politics & Education Program Coordinator
Teachers College, Columbia University

Cliff Hill
Arthur Gates Professor of Linguistics and Education
Teachers College, Columbia University

Stafford Hood
Sheila M. Miller Professor and Associate Dean for Research
Director, Center for Culturally Responsive Evaluation and Assessment (CREA)
College of Education
University of Illinois at Urbana-Champaign

Gerunda B. Hughes
Director
Office of Institutional Assessment & Evaluation
Howard University

Daniel Koretz
Henry Lee Shattuck Professor of Education
Harvard Graduate School of Education

Zeus Leonardo
Associate Professor
Language and Literacy, Society and Culture
Graduate School of Education
University of California at Berkeley

Alan Lesgold
Professor and Dean
School of Education
University of Pittsburgh

Charlie Lewis
Professor of Psychology
Fordham University

Robert Lin
Distinguished University Professor
University of Colorado at Boulder

Robert McClintock
John & Sue Ann Weinberg Professor of Philosophy and Education
Teachers College, Columbia University

Raymond McDermott
Professor, School of Education
Stanford University

Fayneese Miller
Dean, College of Education and Social Services
Professor, Human Development Educational Leadership and Social Policy
University of Vermont

Luis C. Moll
Professor of Language, Reading and Culture
College of Education
University of Arizona

Michelle Moody-Adams
Joseph Straus Professor of Political Philosophy and Legal Theory
Department of Philosophy, Columbia University

Aaron M. Pallas
Professor of Sociology and Education
Teachers College, Columbia University

Thomas W. Payzant
Professor of Practice
Harvard Graduate School of Education

David Pearson
Professor of Language and Literacy, Society and Culture
Graduate School of Education
University of California, Berkeley

Douglass Ready
Professor of Education
Teachers College, Columbia University

Judith Singer
James Bryant Conant Professor of Education
Harvard Graduate School of Education

Mary Kay Stein
Professor and Senior Scientist
School of Education
University of Pittsburgh

Donald Stewart
Visiting Professor of Public Policy
University of Chicago

Hervé Varenne
Professor of Education
Teachers College, Columbia University

Ernest Washington
Professor of Educational Psychology
University of Massachusetts Amherst

Dylan Wiliam
Emeritus Professor of Educational Assessment
Institute of Education
University of London

John B. Willett
Charles William Eliot Professor of Education
Graduate School of Education
Harvard University

Mark Wilson
Professor
Policy, Organization, Measurement, and Evaluation; Cognition and Development
Graduate School of Education
University of California, Berkeley

Dennie Palmer Wolf
Clinical Assistant Professor of Education
Director of Opportunity and Accountability
Annenberg Institute for School Reform
Brown University

Staff

Executive Officer:	**Paola Heincke**
Embedded Journalist:	**David Wall Rice** Associate Professor of Psychology Morehouse College
Multimedia Advisor:	**Mikki Harris** Multimedia Consultant and Professor of Journalism University of Mississippi
Senior Research Scientist:	**Ernest Morrel** Professor of Education and Director, Institute for Urban and Minority Education (IUME) Teachers College, Columbia University
	Rochelle Michel Senior Product Manager - Lead Educational Testing Service
Research Assistants:	**Emily Campbell** **E. Wyatt Gordon** **Emile Session** **Paola Andrea Valencia-Cadena**
Editorial Assistant:	**Maralin Roffino** Assistant to the Director of Communications SUNY Rockland Community College

Work of the Commission

Meetings of the Commission

There were two face-to-face meetings of the Gordon Commission. The initial meeting was held May 24-25, 2011, at the Chauncey Conference Center in Princeton, NJ, and the second meeting was held February 12-13, 2012, at the Caribe Hilton in San Juan, Puerto Rico.

Consultative Conversations

The Gordon Commission spent much of its first year gathering and synthesizing information and perspectives concerning the state of the art and sciences of educational measurement and assessment. The chairman and members of the Commission have held individual consultations with experts around the country who provide input into the work and the direction in which the Commission is going. The Commission hosted more than a dozen consultative conversations with groups that advised the Commission on the identification of issues that need to be addressed and the substance of the issues to be considered.

The Gordon Commission Fellows

The Gordon Commission Fellows is a dynamic group of six emerging pre- and post-doctoral scholars in the fields of the learning sciences, anthropology, psychometrics, the sociology of education, and education technology. These Fellows were assembled to analyze and identify emergent themes, critical innovations, similarities and distinctions, and ultimately synthesize the knowledge produced across the body of the commissioned papers in brief papers of their own. The idea behind the creation of this group was that the work of the commission's experienced scholars and policymakers should be complemented by a younger generation who, in their ongoing dialogue and in their syntheses of the more than two dozen papers, would add new life and new ideas to the project. During their work together over the spring and summer, each Fellow selected overlapping cross-sections of the papers to critically analyze and present for a series of Fellows-led group discussions, all under the tutelage of Commission Chairman Dr. Edmund W. Gordon and Dr. Ernest Morrell, the current director of the Institute of Urban Minority Education (IUME) at Teachers College, Columbia University.

Following is a brief biographical sketch of each of these dynamic young scholars and the links to their synthesis papers:

Keena Arbuthnot received a Ph.D. in Educational Psychology from the University of Illinois at Urbana-Champaign, specializing in Psychometrics/Educational Measurement, Applied Statistics and Program Evaluation. She holds a M.Ed. degree in Educational Psychology and a B.S. degree in Mathematics. In 2005, Dr. Arbuthnot became a Lecturer on Education and a Post-doctoral Fellow at the Harvard Graduate School of Education. She is currently an Assistant Professor at Louisiana State University in the Department of Educational Theory, Policy and Practice. Dr. Arbuthnot conducts research that addresses issues such as the achievement gap, differential item functioning, psychological factors related to standardized testing performance, stereotype threat, and mathematical achievement and African-American students. She also is a former high school mathematics teacher.
http://www.gordoncommission.org/rsc/pdfs/21138_arbuthnot_synthesis_papers_05.pdf

Sherice N. Clarke is pursuing a Ph.D. in Education at the University of Edinburgh, anticipating the award of her doctorate in the spring of 2012. Her thesis is titled *The Inclusive Museum: Understanding Adult ESOL in Museums*. Clarke currently holds a M.Ed. with a concentration in Teaching English to Speakers of Other Languages (TESOL) from the University of Edinburgh, as well as a bachelor's degree in Art History from Hunter College. Her research interests include engagement, agency, identity, classroom discourse and narrative. She currently holds a post-doctoral appointment at the University of Pittsburgh's Learning Research & Development Center. Additionally, she has been an instructor in the University of Pittsburgh's Linguistics Department, a teacher trainer at Edinburgh's Institute of Applied Language, and an English teacher and EFL Department advisor at the SathitBangua School in SamutPrakam, Thailand.
http://www.gordoncommission.org/rsc/pdfs/21138_clarke_synthesis_papers_05.pdf

Juliette Lyons-Thomas is a third-year doctoral student in the Measurement, Evaluation, and Research Methodology (MERM) program at the University of British Columbia. Her current research focuses on think-aloud protocols as a validation method in educational assessment. Her interests also include accountability in education, validity and cross-cultural assessment. Juliette received her M.A. from New York University in Educational Psychology, specializing in Psychological Measurement and Evaluation, and her B.Sc. from McGill University in Psychology.
http://www.gordoncommission.org/rsc/pdfs/21138_thomas_synthesis_papers_05.pdf

Jordan Morris is a second-year doctoral student in the Social Welfare program at the University of California, Los Angeles. She received her B.A. in Psychology from the University of Maryland, College Park, and her Ed.M. in School Psychology and Education Policy from Teachers College, Columbia University. Her research interests include child and adolescent development, critical media literacy, and race and schooling. http://www.gordoncommission.org/rsc/pdfs/21138_morris_synthesis_papers_05.pdf

Catherine Voulgarides is a fourth-year Ph.D. student in the Sociology of Education program at New York University, where she currently works as a research assistant at the Metropolitan Center for Urban Education under the leadership of Dr. Pedro Noguera. At the Center, she has worked on and assisted with the Technical Assistance Center on Disproportionality, in Special Education. Before joining the Center, she worked for the AmeriCorps Vista project in Phoenix, Arizona, coordinating and developing ESL programs for recent immigrant parents in the Phoenix school system. She holds a B.A. in Economics and is a graduate of McGill University in Montreal, Canada. She also holds a MST in Special Education from Pace University in New York City, and taught middle school special education for several years in Washington Heights. Her research interests are centered on the intersection between federal disability legislation and racial and ethnic disproportionality. http://www.gordoncommission.org/rsc/pdfs/21138_voulgarides_synthesis_papers_05.pdf

Amanda Walker Johnson received both a Ph.D. and an M.A. in Anthropology (Sociocultural) from the University of Texas at Austin's African Diaspora Program. In 2004, she served as both a Research Associate for the Research and Evaluation Division at the Intercultural Development Research Association in San Antonio, Texas, and as an Assistant Instructor in the Department of Anthropology at the University of Texas at Austin. In 2005, Dr. Johnson was hired as an adjunct faculty member in the College of Humanities, Arts, and Social Sciences at the University of the Incarnate Word in San Antonio, Texas. In 2006, she was hired as an assistant professor in the Department of Anthropology at the University of Massachusetts Amherst. Dr. Johnson's areas of expertise include African American anthropology; critical race theory and political economy of race in the United States; critical educational theory; feminist theories of race, body and nation; anthropology of science; and cultural and identity politics in the African Diaspora. http://www.gordoncommission.org/rsc/pdfs/21138_johnson_synthesis_papers_05.pdf

Science, Technology and Scientific Imagination

Under the auspices of the Gordon Commission on the Future of Assessment in Education, the Arizona State University (ASU) Center for Games and Impact, the ASU Center for Science and the Imagination, and the Carnegie Mellon Project on Working Examples (funded by the MacArthur Foundation and the Gates Foundation), sponsored two concurrent symposia on October 25-27, 2012, at ASU: 1) The Perils and Possibilities of Emerging Technologies for Learning and Assessment, and 2) Science and Imagination–The Future for the Teaching, Learning and Assessment We Want and How to Get There. These symposia are based on longer-term projects related to these areas:

The Perils and Possibilities of Emerging Technologies for Learning and Assessment: This symposium discussed emerging technologies in the context of how we can put them to the best uses for the most people, in the service of vision for schools and society in the modern world. Among other things, Working Examples (a platform designed at Carnegie Mellon and supported by the MacArthur Foundation and the Gates Foundation) was used to resource the discussion of the future just about to grow, for better or worse, and how its growth can be shaped into desirable directions. Working Examples are sketches of a concrete practice in learning, education, or assessment that its author supports and wants to advocate for. This helped to focus the discussion not on critique of what is wrong, but on concrete proposals of what could be right and good.

Science and Imagination: The Future We Want and How to Get There: This symposium was about letting the imagination free to image, in as concrete terms as possible, the desired long-term future and how this can be made to happen over the long haul. The symposium was about "high-hanging fruit" and what the arts, humanities, social sciences, and hard sciences have together to offer for Imagineering a better, more equitable future for all learners across the globe. The symposium sought to address the question, how can learning, education, and assessment address the long term needs for human growth and survival in a world filled with risky complex systems and seriously unaddressed national problems? Among other things, "design fictions" were used to resource the discussion of makeable futures. Design fictions are concrete images of something that can capture a view of what a desired and possible future might look like.

The aforementioned projects invited participants to imagine the longer-term desired future for learning and education and how this could be actualized, not in terms of abstractions, but in terms of what might actually exist and happen in such a future. The projects asked the participants to liberate themselves from the taken-for-granted assumptions that a sole focus on the near future often brings, and to leverage contributions from all areas and domains from fiction to science. There was overlap in the people participating in the two projects and significant crosstalk between them.

Excellence, Diversity and Equity

In the agreement by which the Gordon Commission was funded, the Commission was asked to give special attention to the problems posed for assessment by the concern for the concurrent privileging of the pursuit of excellence and equity in academic opportunity and achievement. Through the Excellence and Equity Project, the Commission has honored that agreement.

One of the commissioned papers in the Knowledge Synthesis Project was directed at this question. Professor Wade Boykin has written a paper, *Human Diversity, Assessment in Education, and the Achievement of Excellence and Equity*. Professor Boykin was advised by a consultative conversation held at Howard University in December 2011. In addition, a small study group has been designed to give extended discussion to this set of problems. The study group should explore issues such as:

- Tensions in the assignment of responsibility for differentials in assessment outcomes between opportunity to learn, adequacy of the assessment probe, and effort by the performing person;
- Conditional correlates of human performance related to tensions between cultural identities and hegemonic cultural practices and attributes assigned to context;
- What issues require attention when assessing the relationship between diversity and excellence/equity in education?; and
- What challenges are posed by shifts in the populations being assessed and advances in the technologies are emerging, while the criteria by which excellence is judged are also changing?

The concern for excellence and equity is further addressed in the work of the Commission in a group of papers directed at the synthesis of knowledge and thought concerning disabling and handicapping conditions, cultural variation, differences in first language and class/ethnic diversity.

Communication and Social Marketing

A bifocal program of communication was developed. As part of the internal communication plan, the Commission created a blog that was used for the Commission members as a working site for document sharing, discussions and live chat. As the commissioned papers were in various stages of completion, all of the commissioned papers were accessible via the blog to the Commission members. External communications were directed toward three target audiences: practitioners and policymakers; students and parents; and the psychometric research and development community. The external communication plan included:

- The creation of a website where selected materials were available to the general public;
- Development of a bimonthly bulletin—Assessment, Teaching, and Learning—which was the Chairman's vehicle for keeping the public informed concerning the work of the Gordon Commission, and for stimulating a national conversation concerning the relationships between assessment, teaching and learning;
- Hosting of public hearings and forums with key stakeholders to ensure access to the Commission by persons in the field; and
- Webinars for information dissemination and the provision of continuing professional development to policymakers and practitioners;
- The use of regular and social media for the dissemination of strategic messages to target audiences.

Bibliographic Resources

From the beginning of the work of the Gordon Commission, staff members and Fellows have worked to compile a comprehensive collection and directory of the bibliographic resources used in the course of this work. Our Resources File is not a definitive collection; however, it does represent what we think of as the most important literature that has relevance for the work of the Gordon Commission. The collected works are organized under the working categories used by staff and can be searched using common search terms and the special search terms indicated in the File. It can be found under "Resources" at www.gordoncommission.org.

The Knowledge Synthesis Project

This decision led to the conduct of the central activity of the Gordon Commission that has been referred to as the Knowledge Synthesis Project. This initiative consisted of the commissioning of 25 reviews of extant knowledge and thought papers concerning the issues that were identified as most important. These papers can be found at http://www.gordoncommission.org/publications_reports.html. The papers that resulted from this work will be published in the series *Perspectives on the Future of Assessment in Education* in four categories:

Assessment in Education: Changing Paradigms and Shifting Epistemologies

1. *Epistemology in Measurement: Paradigms and Practices – Part I. A Critical Perspective on the Sciences of Measurement* (Ezekiel J. Dixon-Román and Kenneth J. Gergen): This essay provides a critical and historical overview of the science of measurement. The authors situate this overview within the context of the new developments in measurement that promise to change assessment.
   http://www.gordoncommission.org/rsc/pdf/dixonroman_gergen_epistemology_measurement_paradigms_practices_1.pdf

2. *Epistemology in Measurement: Paradigms and Practices – Part II. Social Epistemology and the Pragmatics of Assessment* (Kenneth J. Gergen and Ezekiel J. Dixon-Román): This essay builds upon the foundation of Part I of the series by exploring sociocultural models of assessment and recommendations for its future. The authors make a case for assessments founded in social contexts, as opposed to "one size fits all" models.
   http://www.gordoncommission.org/rsc/pdf/dixonroman_gergen_epistemology_measurement_paradigms_practices_2.pdf

3. *PostModern Test Theory* (Robert J. Mislevy): This essay explains the idea of "neopragmatic postmodernist test theory" and offers some thoughts about what changing notions concerning the nature of and meanings assigned to knowledge imply for educational assessment, present and future.
   http://www.gordoncommission.org/rsc/pdf/mislevy_postmodern_test_theory.pdf

4. *What Will It Mean to Be an Educated Person in Mid-21st Century?* (Carl Bereiter and Marlene Scardamalia): This paper comments on the ways in which the intellective demands on educated persons will change in the near future. Attention is called to

issues related to knowledgeability and capacity for adaptability. The authors focus on this shift in the face of tremendous technological advances that continue to affect the field of education.
http://www.gordoncommission.org/rsc/pdf/bereiter_scardamalia_educated_person_mid21st_century.pdf

5. T*oward an Understanding of Assessment as a Dynamic Component of Pedagogy* (Eleanor Armour-Thomas and Edmund W. Gordon): The authors of this paper explore a form of teaching that integrates assessment, curriculum, and instruction in the service of learning. They argue that assessment could be strengthened by its integration into pedagogy.
http://www.gordoncommission.org/rsc/pdf/armour_thomas_gordon_understanding_assessment.pdf

6. *Preparing for the Future: What Educational Assessment Must Do* (Randy Elliot Bennett): This essay explores the forms that summative and formative assessments will take and the competencies that they will measure in the future. The author proposes the core competencies that assessments must provide in order to continue to be relevant.
http://www.gordoncommission.org/rsc/pdf/bennett_preparing_future_assessment.pdf

7. *Changing Paradigms for Education: From Filling Buckets to Lighting Fires to Cultivation of Intellective Competence* (E. Wyatt Gordon, Edmund W. Gordon, John Lawrence Aber, and David Berliner): This essay provides an overview of the factors that promise to shift the ways in which we think about education and assessment in the future. The authors theorize about the ways that education will need to change in order to survive through these paradigm shifts.
http://www.gordoncommission.org/rsc/pdf/gordon_gordon_berliner_aber_changing_paradigms_education.pdf

Changing Targets of Assessment in Education

8. *The Possible Relationships Between Human Behavior, Human Performance, and Their Contexts* (Edmund W. Gordon and Emily B. Campbell): This essay explores how context affects human behavior, performance and the assessments of behavior and performance. The authors argue for a form of assessment that is capable of accounting for the contexts in which individuals exist.
http://www.gordoncommission.org/rsc/pdf/gordon_campbell_implications_assessment_education.pdf

9. *Education: Constraints and Possibilities in Imagining New Ways to Assess Rights, Duties and Privileges* (Hervé Varenne): This essay explores the relationships between the granting of political privilege, United States public schools, and the contemporary uses of assessment. The essay then imagines how a set of new institutions that challenge these dynamics may look and feel.
   http://www.gordoncommission.org/rsc/pdf/varenne_education_constraints_possibilities.pdf

10. *Toward a Culture of Educational Assessment in Daily Life* (Carlos A. Torre and Michael R. Sampson): This essay makes the case for future cultural practices through which lay persons, as well as educators, enjoy reasonably sophisticated understandings of educational assessment data and processes. The authors outline their best estimates of where education is going, where it needs to go, and, therefore, what may be needed from educational self-assessment during the 21st century.
    http://www.gordoncommission.org/rsc/pdf/torre_sampson_toward_culture_educational_assessment.pdf

11. *Toward the Measurement of Human Agency and the Disposition to Express It* (Ana Mari Cauce and Edmund W. Gordon): This paper attempts to bring together multiple bodies of knowledge in developing a multidimensional view of human agency, with a focus on those factors that allow for and facilitate the development and display of human agency. The authors do so with an eye toward the development of ways to assess, facilitate, and foster human agency through strategies that are most relevant to academic achievement and the advancement of intellective capacities.
    http://www.gordoncommission.org/rsc/pdf/cauce_gordon_measurement_human_agency.pdf

12. *Test-Based Accountability* (Robert L. Linn): This essay explores how test-based accountability can increase student achievement and equity in performance among racial-ethnic subpopulations, students who are poor and their more affluent peers. The author expands on the history of test-based accountability and the prospects for its future.
    http://www.gordoncommission.org/rsc/pdf/linn_test_based_accountability.pdf

13. *Variety and Drift in the Functions and Purposes of Assessment in K-12 Education* (Andrew Ho): This essay reviews recent frameworks that differentiate among purposes of educational assessments, particularly purposes of large-scale, standardized assessments and reflects on the forces that shape the purposes of any particular assessment over time. The author uses this discussion to identify migratory patterns for modern assessment programs as they expand across purposes. http://www.gordoncommission.org/rsc/pdf/ho_variety_drift_functions_purposes_assessment_k12.pdf

14. *Testing Policy in the United States: A Historical Perspective* (Carl Kaestle): This essay provides an overview of the history of testing policy in the United States. The author focuses on policy issues in order to allow the reader to reflect upon how current-testing practices came to be. http://www.gordoncommission.org/rsc/pdf/kaestle_testing_policy_us_historical_perspective.pdf

Psychometric Change in Assessment Practice

15. *Four Metaphors We Need to Understand Assessment* (Robert Mislevy): This essay offers four metaphors that serve as facilitators for discussion surrounding assessment. The goal of this paper is to create a common “language” that can be used to talk about assessment and its implementation. http://www.gordoncommission.org/rsc/pdf/mislevy_four_metaphors_understand_assessment.pdf

16. *Assessment as Evidential Reasoning* (Joanna S. Gorin): This essay argues for the expansion of the notion of assessment to include diverse sources of evidence. In particular, the author is concerned with assessments that can measure real world variables and factors. http://www.gordoncommission.org/rsc/pdf/gorin_assessment_evidential_reasoning.pdf

17. *Assessment in the Service of Teaching and Learning* (Clifford Hill): This essay argues for assessments that evaluate individual learners with a view to improving individual instruction. The author’s framework goes beyond computer adaptive testing into the development of a dual model for both testing and project components. http://www.gordoncommission.org/rsc/pdf/hill_assessment_service_teaching_learning.pdf

18. *Testing in a Global Future* (Eva Baker): This essay approaches the future of testing in a globalized context by addressing factors central to predicting the future of assessment. The roles of international comparisons, demography, knowledge expansion, job changes, and of technological growth are central to the author's analysis.
    http://www.gordoncommission.org/rsc/pdf/baker_testing_global_future.pdf

19. *Technological Implications for Assessment Ecosystems: Opportunities for Digital Technology to Advance Assessment* (John T. Behrens and Kristen E. DiCerbo): This essay explores how technology promises to shape the future of assessment. The authors discuss the nature of data and its role in human self-awareness.
    http://www.gordoncommission.org/rsc/pdf/behrens_dicerbo_technological_implications_assessment.pdf

20. *Toward the Relational Management of Educational Measurement Data* (Greg K. W. K. Chung): This essay conceptualizes how technology can be used to help assessment individualize instruction. The author explores how to leverage individualized data to measure what students understand and can do, to derive meaningful measures of cognitive and affective processes, and to develop capabilities for precise diagnosis and targeting of instruction.
    http://www.gordoncommission.org/rsc/pdf/chung_toward_relational_management_educational_measurement.pdf

Assessment in Education and the Challenges of Diversity, Equity and Excellence

21. *Human Diversity, Assessment in Education and the Achievement of Excellence and Equity* (A. Wade Boykin): This essay explores the paradox between the origins of assessment as a vehicle for equity and the contemporary tendency of assessment to be used for exclusionary purposes. The author provides a framework for more proactively addressing issues of race, culture, excellence, equity, and assessment in education.
    http://www.gordoncommission.org/rsc/pdf/boykin_human_diversity_assessment_education_achievement_excellence.pdf

22. *Assessment of Content and Language in Light of the New Standards: Challenges and Opportunities for English Language Learners* (Kenji Hakuta): This essay plays out an imagined scenario in 2017 (five years hence) for the assessment of English language learners, based on assumptions about what the author knows of the Common Core State Standards and how this most recent wave of reform will impact state and local systems in the assessment of content and English language proficiency.
    http://www.gordoncommission.org/rsc/pdf/hakuta_assessment_content_language_standards_challenges_opportunities.pdf

23. *Democracy, Meritocracy and the Uses of Education* (Aundra Saa Meroe and Edmund W. Gordon): This essay grapples with the tension between meritocracy and democracy. The authors focus on the practical outcomes associated with each and problematize current conceptions of both ideas.
    http://www.gordoncommission.org/rsc/pdf/meroe_democracy_meritocracy_uses_education.pdf

24. *Accommodation for Challenge, Diversity and Variance in Human Characteristics* (Martha L. Thurlow): This essay explores the continuing evolution in instructional and assessment accommodations. The author provides background on the theoretical perspectives underlying assessment accommodations, including the history of accommodations, accommodation policies, and validity considerations. http://www.gordoncommission.org/rsc/pdf/thurlow_accommodation_challenge_diversity_variance.pdf

25. *A Social Psychological Perspective on the Achievement Gap in Standardized Test Performance Between White and Minority Students: Implications for Assessment* (Rodolfo Mendoza-Denton): This essay explores the issue of academic motivation and performance from a social psychological perspective. The author focuses specifically on the academic achievement gap between White and minority students. http://www.gordoncommission.org/rsc/pdf/mendoza_denton_social_psychological_perspective_achievement_gap.pdf

The Gordon Commission
on the Future of Assessment in Education

Contact us at:
contact@gordoncommission.org
Gordon Commission • P.O. Box 6005 • Princeton, NJ 08541
www.gordoncommission.org

The Gordon Commission was established by ETS to investigate and advise on the nature and use of educational testing in the 21st century. 22008