Virtuous cycles of learning: a digital revolution

Paper for the International School for Mind, Brain, and Education.

Ettore Majorana Science Center, Erice, Sicily.

August, 2011

Introduction: redesigning testing during the digital revolution

Advances in technology increasingly influence educational practice and research. For example, during the first ten years of the 21st Century large scale educational testing has become almost entirely computerized, and classroom testing isn't far behind. This paper explores the future of educational testing in the context of emerging post-industrial testing infrastructures and advances in the learning sciences. As background, we provide a brief history of the relation between technology and standardized testing, exploring how technological advances have shaped the way we think about and employ educational tests. Our look at the history of testing leads us to conclude that, to date, testing practices have been influenced more by technological innovations in test design and administration than by advances in the learning sciences. The history of testing also sheds light on our current testing practices and what the near future might hold.

Internet technology and other "edu-tech" trends will transform standardized testing in the coming decades. We argue that the learning sciences should assume more responsibility for shaping the development and application of new testing technologies. Appealing to the legacy of constructivism and the emerging field of mind, brain, and education, we suggest new *learning-centric* directions for the development of testing technologies. We argue that most, if not all, tests can and should be designed to aid directly in the process of learning and teaching. This is the ideal of testing in the service of learning, made possible through advances in computer technology, the learning sciences, and psychometrics. The DiscoTest Initiative serves as an example of these positive new directions for testing.

At the core of the DiscoTest initiative are a growing number of subject-specific, research-based, embedded, formative, and diagnostic assessments, all of which are standardized to Fischer's Skill Scale (Fischer, 1980; Fischer & Bidell, 2006) through the use of the Lectical Assessment System (Dawson, 2010; Dawson-Tunik, 2004). These resource-rich assessments support multi-level learning in educational systems by advancing the understanding of students, teachers, decision makers, and researchers (Dawson & Stein, 2008; Stein, Dawson, & Fischer, 2010). We look at the research and development behind DiscoTests and the technological advances that make them possible. Finally, we explore the educative and measurement functions of DiscoTests in schools, where they provide real-time developmentally appropriate feedback to students, individual and classroom level insights into student learning for teachers, systems-level profiles of learning for administrators, and unprecedented micro and macro developmental data for researchers. As we explain, DiscoTests aim to promote *virtuous cycles of learning* for everyone involved, making good on the possibilities of the digital revolution by leveraging the new science of learning.

A brief history of testing and technology

Advances in computer technology have enabled governments and industries to begin fundamentally *redesigning* basic infrastructures for transportation, communications, and the delivery of energy. Educational systems as a whole, and testing infrastructures in particular, are also being redesigned in the wake of the digital revolution. If current trends continue, the goals of providing government-subsidized broadband and “one laptop per child” are likely to be realized in the near future. Tomorrow’s “edu-tech” will enable new forms of teaching and learning, as well as new forms of testing, administration, and research. But will these new forms of testing merely serve as more streamlined vehicles for serving up conventional assessments, or will educational testing itself be transformed? We argue that transformation is possible if the development of new testing infrastructures is guided by a conception of testing that embodies the best of what we know about teaching and learning.

An important theme in the history of standardized testing is the connection between advances in technology and advances in the ambitions, scope, and efficiency of testing practices. For this brief discussion we focus only on the history of testing in the United States. Importantly, US testing practices have set the pace for many countries, although not all (Haney, Madaus, & Lynos, 1993). The goal of this section is to set a context for discussing the effects that contemporary advances in information technology will have on standardized testing in the decades to come.

Technological advances have changed the way tests are administered and scored, from paper and pen tests scored and commented by humans, to pencil and bubble-sheet tests scored by machines—and now to fully computerized test administration processes that involve automated scoring algorithms and complex statistical analyses. Each major technologically induced change in educational testing has been accompanied by changes in its form and function. Prior to the introduction of standardized multiple-choice tests, both classroom and qualifying assessments took many forms. They pulled on a wide range of skills and knowledge, including those for writing, reasoning, oral expression, and knowledge application, and often involved one-on-one interaction with a teacher. However, these tests were scored subjectively and were not standardized. There was no way of knowing if students in one classroom were being evaluated on the same terms as students in another classroom. These problems were solved with the advent of pencil and bubble-sheet tests that made it possible to administer standardized objective tests to thousands of students. However, there were costs. The new tests were composed exclusively of multiple-choice items with right and wrong answers, limiting the range of skills they could address, and the role of testing as a means of sorting students began to overshadow the role of testing as part of the learning cycle.

The first of the pencil and bubble-sheet assessments were primarily used to “sort” students at the level of the school, having little effect upon classroom instruction or testing beyond the gradual introduction of multiple-choice items into classroom assessments. Hundreds of test publishers sold tests to thousands of schools, and while local practices differed, test-based tracking (and exclusion) became the norm. As this new form of testing gained in popularity, advances in test development and automated scoring would make it possible to create a national testing infrastructure, realized in the United States with the founding of the Educational Testing Service. Eventually, federal policies would expand the national testing infrastructure to include all K-12 schools, a move that has dramatically altered testing practices, deeply affecting classroom practices, straining existing testing resources, and creating a demand for improved testing infrastructures and higher quality tests with greater educational value.

We trace connections between testing and technology across three broad historical eras: industrial (1880 - 1930); late-industrial (1930 - 1980); post-industrial (1980 – present). These eras are differentiated for heuristic purposes and reflect comparable divisions found in the historical sources cited below.

Industrial Era (1880-1930)

The roots of modern educational testing can be found in the work of early psychologists (O'Donnell, 1985). Precisely engineered electrical equipment and newly invented psychometric techniques allowed for the kinds of precise measurements that brought early prestige to psychophysics, and later, behaviorism. As psychology grew beyond the laboratory, the importance and prestige of measurement remained, and the practice of mental testing became central to the identity of the emerging discipline (Brown, 1992). IQ testing, in particular, captured the public imagination and found its way into the plans of policy makers and educational reformers (Gould, 1981).

Testing was labor intensive in the days when the dominant technologies for education were the textbook, the chalkboard, and pen and paper. Test administrators were required to handle test-takers directly, often one-on-one, and numerous personnel were needed to score tests and compute scores by hand. But the invention of the multiple-choice item changed everything (Samelson, 1990). During mobilization efforts for the First World War, American psychologists were recruited to administer examinations—the US Army Alpha Tests—to over two million soldiers. This large scale testing effort was possible mainly because of the invention of the multiple-choice test item format. Multiple-choice tests could be easily administered, then scored rapidly and objectively with stencils that overlaid answer sheets. The Army efforts demonstrated that an IQ-testing infrastructure (of "industrial scale") could be built and maintained (Sokal, 1990).

At the turn of the century, industrialization, urbanization, and immigration forced major changes in American public education, especially in large urban districts, which experienced skyrocketing student numbers (Cremin, 1988). Thanks in part to the example of the war time testing efforts, IQ-testing and related forms of "standardized" assessment figured prominently in the bureaucratic and pedagogical reforms that created the schooling practices characteristic of the industrial era. By 1930, IQ-style multiple-choice standardized testing was ubiquitous in American public schools, and was typically used to track students into different groups for instruction and management (Chapman, 1988). This growing testing infrastructure also allowed for the beginnings of the "scientific management" of school operations (Callahan, 1964). Educational researchers began, for the first time, to systematically use test results to inform their thinking about school improvement (Lagemann, 2000). The greatest strengths of these first large-scale testing efforts were their contributions to organizational efficiency and educational research, where they (ostensibly) offered much needed objectivity and professional clout in the increasingly complex and politicized world of education.

However, early testing infrastructures were expensive, error prone, and widely criticized, with contentious debates about IQ testing in the public sphere, scandals, and the legacy of Eugenics (Gould, 1981). Test-based tracking practices and exclusion took on increasing significance as economic changes made educational attainment increasingly necessary for occupational success. The weaknesses of early large-scale testing efforts include their ruinous effects on the educational trajectories of some students, rampant bias and frequent errors in item construction and scoring, and the increasing use of multiple-choice, rather than open-ended, items in classroom tests. The young testing industry responded to public criticism by placing an emphasis on objectivity and standardization, which it pursued by constraining item content and making further innovations in test-scoring technology, statistical modeling, and test design. But they largely ignored concerns expressed by teachers and educational psychologists about the effects of this kind of testing on teaching and learning in the classroom. Automated test scoring technology and growing federal influence on the shape of public education would soon create a major role for testing in America's increasingly large and complex educational system ((Cremin, 1988; Lemann, 1999; Spring, 1989).

Late-Industrial Era (1931-1980)

In 1931 a young high school science teacher from Michigan solved a technological problem that IBM had been working on unsuccessfully for years: automated test scoring. The implications and subsequent technical developments—such as the Scantron machine—would facilitate the construction of the first national standardized testing infrastructure (Lemann, 1999).

The centerpiece of this newly automated testing infrastructure was the Educational Testing Service (ETS), which emerged, in part, as a result of the federal government's interest in exercising quality control and determining how best to fund research and development. The SAT—a multiple choice test with direct lineage to the Army Alphas—changed the nature of college admissions processes and shaped the educational careers of the millions who flooded into the higher education system in the post-war decades (Spring, 1989).

Several waves of sweeping federal educational legislation during the Cold War consolidated the shape of late-industrial era education. These included federal funding to create the NSF and promote STEM education in K-12, civil rights, desegregation, and the War on Poverty programs, such as Head Start. The federal government began recruiting the testing industry to aid in policy and program evaluation studies (Campbell, 1975). As the "scientific management" of educational reform drew widening support, federally mandated educational testing expanded, eventually to include the NAEP, the Iowa tests, and a host of other national K-12 tests. The strengths of this first wave of national testing efforts included its emphasis on social justice and the promotion of educational equity; ETS was founded with the explicit goal of expanding college access and replacing the existing aristocracy with a meritocracy (Lemann, 1999).

However, discourse continued about the limitations of standardized tests, including critiques of the multiple-choice item itself, new concerns about bias in testing practices, and a growing discontent about the overall shape of the emerging test-based meritocracy (Hoffman, 1962; Nairn, 1980). A key weakness in the post-war testing infrastructure was its continued reliance on a narrow range of item designs, with the vast majority of tests (including the SAT) wedded to the multiple-choice format and almost entirely uninformed by advances in learning theory and educational psychology. The federal influx of support for testing led also to broad concerns that the educational system was being turned into a kind of "sorting machine" for human capital, rewarding a limited set of educational trajectories through the use of tests that focus on mostly STEM related skills and competencies (Spring, 1989). Relatedly, the negative effects of increased testing on classroom practice continued to mount, as the high-stakes multiple-choice exam came to symbolize American education itself (Sacks, 1999; Samelson, 1990)

Nevertheless, thanks to federal and private interests, and the ostensive objectivity tests provided for educators vying for professional and scientific respectability, the standardized testing industry continued to expand and diversify (Haney, et al., 1993). Computer technology, in particular, fuelled ambitions for building testing infrastructures of increasing size and sophistication. During the seventies and into the eighties the role of large-scale standardized testing infrastructures in public education became increasingly significant and more complex, with the first appearance of high-stakes state-wide graduation exams, the AP, GRE, LSAT, and an ever-growing test-prep and tutoring sector (Sacks, 1999).

Post-industrial (1980- present)

Continued advances in the computerized administration and scoring of tests, as well as advances in data management techniques, fueled policy makers' ambitions to build a national K-12 testing infrastructure for use in program evaluation and systems-level accountability (Phillips, 2003). The final push toward a comprehensive federal testing infrastructure began with President Bush's *America 2000* (Bush, 1991), in which he proposed a national testing apparatus that would be tied to a national curriculum and used to assure the equitable distribution of educational opportunity as well as American competitiveness in the global marketplace. President Clinton later endorsed the plan but backed off the idea of a national test. In 2001, *No Child Left Behind* (Hess & Petrilli, 2006) provided a federal mandate to build a K-12 testing infrastructure for accountability purposes, but not in the form of a single national test. Instead, a decentralized competition-oriented set of testing practices was rolled out, with great variation from state to state, ushering in an expanding for-profit computer-intensive testing industry (Toch, 2006).

The verdict on NCLB as a testing initiative is officially still out, but the emerging results are mostly negative (Hursh, 2008), even among those with sympathies toward NCLB's broad goals (Koretz, 2008; Ravitch, 2010). A 2011 report from RAND argues that NCLB testing has significantly and negatively constrained classroom practice in many places, and calls for new approaches to educational measurement, especially those that contribute to learning and leverage new technologies (Schwartz, Hamilton, Stecher, & Steele, 2011). Calls for seriously rethinking the testing infrastructure have also come from the National Research Council (2001), which likewise argued for a testing infrastructure built around the learning sciences and new technologies. Prominent learning scientists and educators have taken strong stands against current testing practices (Darling-Hammond, 2010).

Central to many of these criticisms is the fact that tests built in the wake of NCLB have continued to heavily rely on a WWI era technology—the multiple-choice item. In part, this is because the testing industry is stretched thin, without adequate scientific personnel or research and development capabilities (Toch, 2006). The adoption of new technologies in test design, administration, and scoring is driven primarily by concerns about efficiencies for dealing with economies of scale, as now millions of students must be tested every year. Likewise, as history has shown, advances in the learning sciences do not drive test design or frame the adoption of new technologies for testing. Putting testing in touch with the new sciences of learning is not on the agenda of most large testing firms (Farley, 2009). The narrow focus of item construction and the narrow range of academic content has begun to drastically affect classroom practice, moving many schools toward a focus on test-prep, and sometimes toward outright cheating (Hursh, 2008; Nichols & Berliner, 2007) Nevertheless, federal policy is likely to mandate more testing, and to continue to support the high-stakes test-driven reforms, school closures, and charter school rotations that dominate many school systems (Ravitch, 2010).

The Obama administration has sent mixed messages about the future of testing. They have been openly critical of NCLB, specifically noting the negative effects of testing (Obama, 2008; White house office of the press secretary, 2012). While at the same time, Obama's

Department of Education has funded two large test-development consortia tasked with building the beginnings of a new national testing infrastructure, which is to be wedded to the new Common Core Standards (US department of education, 2011). The call is for new assessments that measure 21st century skills and go beyond multiple-choice items and shallow assessments of skill. So, while the shape of tomorrow's national testing infrastructure is indeterminate, we can be sure it will be larger, more complex, and integrated with new technologies.

Dovetailing with some of these central issues, which directly concern the upcoming reauthorization of NCLB by the Obama administration, some educators argue that, irrespective of the efficacy of prior testing policies, technology has rendered the prior era's testing infrastructure obsolete (Collins & Halverson, 2009). A report from the National Science Foundation task force on cyber learning (NSF task force on cyberlearning, 2008), focusing on the future of educational technology, describes burgeoning markets for educational technologies that rely on hi-speed internet and powerful computing. Post-industrial era testing infrastructures will be built upon whatever new technologies come to play a dominant role in education. The NSF suggests that testing will likely become wedded to technologies that allow for embedded testing that is repeated, formative, and in the service of real-time learning (*Ibid*).

There is a great deal of speculation concerning tomorrow's educational technology—the so-called emerging "edu-tech." In particular, there are some important characteristics and trends in emerging educational technologies that are relevant to testing (Collins & Halverson, 2009; NSF task force on cyberlearning, 2008):

- **Technology saturation:** Computers and other networked devices will reach an increasingly large portion of the population, especially students. Smart phones and other portable devices are already ubiquitous. The One Laptop Per Child initiative is a testament to this trend.
- **Just in time learning:** Learners will use technologies that organize databases of resources that allow for instant access to whatever needs to be learned.
- **Customization:** Learners will use technologies that are responsive to individual differences, as educational opportunities are guided by user preference and performance.
- **Scaffolding:** Learners will use technologies that structure the delivery of tasks and learning opportunities based on close-to-real-time assessments of performance.
- **Reflection:** Learners will use technologies that document the history of user performances and then present comparisons between users' histories and those of others. Technology will enable detailed portfolio management systems, templates, scoring interfaces, and databasing.
- **Distance learning and online education:** More education will take place at a distance through on-line learning environments, with improved efficacy due to advances in video conferencing and content delivery systems.

- **Databases and electronic learning records:** With embedded assessments and automated progress and behavior monitoring technologies, educational record keeping will become as detailed and complex as medial record keeping.

This list of trends and characteristics is not exhaustive, but it is nevertheless suggestive of what tomorrow's educational institutions will have to work with. The history of testing should teach us that advances in technology can radically alter testing practices and force us to reconsider the function of testing in educational systems. If this list captures the characteristics of the technologies that will become essential to our educational configurations, then we should be inquiring into the preferable directions for testing that they make possible. In particular, we follow the aforementioned NRC and NSF recommendations, and assert that testing infrastructures should promote learning and be based on a science of learning, as opposed to being used primarily for the purposes of efficiency, tracking, and accountability. Our task below is to explore one approach to creating a testing infrastructure based in the new science of learning, which leverages emerging educational technologies. However, this depends first on clarifying what we mean by the learning sciences, which involves, in part, some answers to questions about the nature of learning itself.

Getting the new science of learning into tomorrow's edu-tech

We have raised a broad question facing many educators: how to remain committed to the best of what is known about teaching and learning while also incorporating diverse new educational technologies. There is widespread excitement about the possibilities for high-tech educational futures (Christenson, Johnson, & Horn, 2008; Zucker, 2008). But in order to responsibly design and use technologies that aid student learning we must have some explicit theory of learning to work from (NSF task force on cyberlearning, 2008). Not all informational environments are educational environments, and not all new technologies should be widely adopted simply because they exist. We argue that the development of new testing infrastructures should be guided by a conception of testing that complements the best of what we know about its role in teaching and learning. New technologies for testing should serve—not drive—learning and teaching.

We view learning as a kind of *virtuous cycle,* a repeated process of goal-directed action, which is cyclically improved through the processing of feedback from the environment. Learning is a creative, constructive, trial-and-error process that is deeply biological. Broadly speaking this view can be placed under the banner of *constructivism,* one of the most coherent and comprehensive families of learning theories (Baldwin, 1906; Piaget, 1960). Recent theorists in this tradition have broadened the research horizons of constructivism into neuroscience, computer science, and the field of Mind, Brain, and Education (Fischer, 2009; Mareschal et al., 2007). The virtuous cycle view of learning is broadly applicable, characterizing neural-network models (Spitzer, 1999), language acquisition (Tomasello, 2005) and even political theory and governance (Buck & Villines, 2007). We believe that the natural, well functioning, virtuous cycle of learning is a simple but powerful model that is well worth embedding in new educational technologies.

Goal-action-feedback: The foundational "feedback loop" of life and learning

Most learning scientists—although they offer differing accounts—agree that as a general model of learning the "goal-action-feedback loop" presents learning at its most fundamental and general. This loop contains four essential components: goal, action, feedback, and repeat. When integrated dynamically, they allow us to focus on the goal, choose an appropriate action, interpret the outcome, and decide how to correct or adjust the currently available pool of actions and ideas to better reach the goal on the next attempt.

Cognitive and behavioral learning scientists (Fischer & Bidell, 2006), evolutionary psychologists (Campbell, 1987), and neuroscientists (Mareschal, et al., 2007) have repeatedly rediscovered this most basic mechanism of learning. Piaget and Baldwin first characterized this feedback loop as an ongoing attempt to achieve balance between assimilation and accommodation. More recently, neuroscientists have observed that neurons in the neocortex act in cohorts where repeated stimulation of the network in a variety of contexts fine-tunes the emergence of specific patterns (*Ibid*). We argue that, whether viewing the human learning from the perspective of neurons or behavior, success depends on engaging in a virtuous cycle—the positive feedback loop of learning.

The same process is also thought to characterize organizational learning. Since the 1960s systems theorists (Forrester, 1964), cyberneticists (Beer, 1981), and organizational psychologists (Senge, 1990), have convincingly proposed that institutions should embed feedback-loops that enable goal monitoring and self-correction. A learning organization is one that can adjust its policies based on data about how well they work. This notion is even embedded in the guiding ideals of democratic societies, such as the idea that social policies should be implemented in an experimental mood and remain open to revision in a virtuous cycle that ensures we are continually learning about our schemes of social organization and their effectiveness.

And of course, science itself is a kind of virtuous cycle. Contemporary philosophers of science echo Bacon, Peirce, and Popper in arguing that the scientific method should be understood as a learning process that involves testing a hypothesis against the world experimentally, reconsidering that hypothesis in light of what the experiment reveals, and then testing a new revised hypothesis, and so on (Brandom, 1994; Elgin, 1996). Shifts to new scientific paradigms can be conceived of as part of this ongoing cycle of conceptual revisions and experimentation.

Educators have long sought to foster virtuous cycles of learning at multiple levels in the educational system: for students in classrooms, for administrators as a part of organizational learning in schools, and for researchers as a way of building cumulative knowledge about teaching and learning (Dewey, 1929; Lagemann, 2000). New educational technologies will make many things possible including new ways to enable on-going learning processes that are integrated across different levels of the educational system (Collins & Halverson, 2009). We now turn back to testing and explore new possibilities for coupling constructivist theories of learning with the affordances of new technologies. We argue that it is now possible to build a testing infrastructure that enables multi-level learning in educational systems.

Leveraging tommrow's edu-tech to promote tetsing in the service of learning

The goal of using the learning sciences to inform the design and adoption of new educational technologies should be a top priority (National Research Council committee on testing and assessment, 2001; NSF task force on cyberlearning, 2008). However, the history of testing and technology offered above shows that advances in learning theory have never driven the adoption and design of new testing technologies. Rather, concerns about efficiency and system-level accountability have historically trumped concerns about teaching and learning. Given these historical trends, it is likely that new educational technologies might primarily be used to deliver the same old multiple-choice tests, only faster, to more students, with greater frequency, and more sophisticated data-analytic techniques. The short list of emerging edu-tech trends presented above, including new possibilities for multitudinous, embedded, real-time assessment, could result in a testing infrastructure that is increasingly insensitive to the needs of teachers and students, divorced from research about learning, and used mainly for the real-time systems-level surveillance of teacher, student, and school performance.

But these same technological trends open up possibilities for radically new testing approaches that are built around the best of what we know about the process of learning. Building such tests will involve explicitly adopting a theory of learning, such as the constructivist inspired, multi-level virtuous cycles model introduced above, and using this theory from the outset to structure research and development efforts. This cuts against the grain of the history of testing in so far as an explicit theory of learning would drive test design and technology adoption. We believe that it is now possible to harness new technologies to create a new kind of testing infrastructure that is learning-centric.

Tests combining constructivist theories of learning with new trends in educational technologies would at least conform to the following design principles. They would be:

- **Evidence-based:** priorities should be shifted so that test development is guided by the learning sciences, informed by research about learning.
- **Build knowledge**: given trends toward increasingly large *digital databases of electronic learning records*, tests should aim to contribute to the learning sciences through data collection and housing.
- **Broadly available**: given trends toward *technology saturation* and *online education*, tests should leverage Internet and computer technologies to serve the least advantaged.
- **Support teaching and learning**: tests should enable *customization*, *scaffolding,* and *just in time learning* by organizing the delivery of *online education* resources to educators, students, and parents.
- **Low-stakes**: *just in time learning* requires multitudinous embedded assessments; many low-stakes formative tests, diverse topics, no high-stakes testing anxiety, no reason for cheating.

- **Relevant**: tests should leverage the diversity and ever expanding affordances of *online educational environments* to allow students to operate on knowledge that matters to them and practice essential life skills, while also working toward mastery of the academic competencies targeted by standards.
- **Embeddable**: tests that enable *just in time learning* and *reflection* should be part of classroom lessons and the learning.
- **Formative**: tests that enable *customization* and *scaffolding* should be learning experiences in themselves, directly contributing to student understanding.
- **Diagnostic**: tests that enable *customization*, *scaffolding*, and *reflection* should be based on research about student learning, providing insights into what the student can do now and what would most benefit their learning next.
- **Standardized**: increasingly large *digital databases of electronic learning records* will need to be built around common measures and indexes to enable both scientific and organizational learning; tests should thus be standardized to a universal learning scale.

These design parameters require more elaboration than space provides for here, but they should give a rough sense of the new possibilities for testing that emerge at the interface of the learning sciences and new educational technologies. The history of testing has taught that while new technologies bring transformations in testing practices, these transformations have occurred in isolation from the learning sciences. The invention of the multiple-choice question, and then the Scantron machine, did more to shape testing than any advances in our understanding of learning. Students and teachers paid the price, as the demands of organizational efficiency and systems-level accountability consistently trumped concerns about teaching and learning in the design and adoption of new technologies for testing. As the stage is now set for a new technology-wrought revolution of testing practices, it is critical that educators and policy makers build consensus around a set of design parameters like the ones above in order to shape the new testing infrastructure in ways that will be more beneficial for teachers and students. We now look at an attempt to actualize positive, learning-centric, technology enabled directions for testing—the DiscoTest Initiative. This initiative is an attempt to build a new kind of testing infrastructure based on the design parameters outlined above.

The DiscoTest Initiative: using new technologies to enable virtuous cycles of learning

We have already mentioned reports from the NRC and the NSF that call for a testing infrastructure based more on the science of learning and new digital technologies. Others have also contributed to this call. For example, Toch (2006) describes the underfunded and understaffed state of the post-NCLB testing industry, which for the most part lacks a sufficient number of psychometrians and learning scientists to conduct the necessary research and development to develop high-quality tests. Additionally, there are continuing high-profile calls for an infusion of new ideas into K-12 testing, including calls for reconsideration of the broad role of testing in reform and politics (Ravitch, 2010). There is also increasing pressure to expand

and diversify testing, promote innovative research and development, and to re-think testing practices in light of new technological possibilities (Collins & Halverson, 2009; Schwartz, et al., 2011). In this section we present an overview of one current research and development effort that brings these themes together.

In Latin, *disco* means ‘to learn,” and now serves in English as the root for such words as *disco*urse and *disco*very. Coincidentally, in modern times, *disco* also evokes the image of young people joyfully interacting with music, an image that sits well with the notion of learning as fun. DiscoTests are research-based, subject-area specific, embeddable, formative, computer-administered assessments. They consist of “best practices” formative assessments composed of open-ended essay questions that ask students to respond to *ill-structured* real-world problems—problems without clear cut answers—and provide standardized scores as well as rich real-time educative feedback for students and teachers. DiscoTests will eventually number in the hundreds, and are intended to be administered frequently in low-stakes contexts to support teaching and learning. Because they are standardized to a universal learning metric, they are also capable of providing administrators with system-level data for program evaluation purposes. Additionally, they are uniquely powerful tools for researchers due to their research-intensive construction, the frequency with which they can be taken, their diversity, and their standardization.

Below we outline the research and development efforts that go into the construction of DiscoTests, and then review their educative and measurement functions. It should become clear that the ability of DiscoTests to enable multi-level learning in educational systems stems, is due in large part, to a carefully crafted interface of the learning sciences with advances in computer technology.

DiscoTest research and development

Building DiscoTests involves a set of related undertakings: 1) establishing collaborations with educators; 2) conducting basic research into how students learn the specific concepts and skills targeted by test items; 3) building computer-based low-inference scoring rubrics that are based on this research; 4) compiling expertly vetted learning resources. The finished product is a mature DiscoTest, complete with coding rubrics; diagnostic reports for students, teachers, and parents; targeted learning resources for individual students; a range of teaching resources, including a variety of real time reports and lesson plans that can be tailored to the needs of specific classrooms; and real-time group-level data for decision makers. The broad goal of the initiative is to build dozens of tests, spanning numerous academic subject areas, in order to provide a new kind of testing infrastructure[i].

Here we quickly sketch the broad test development process and then take a closer look at some of its most important elements, especially those made possible through emerging educational technologies. More detailed accounts of our research and development efforts have been presented elsewhere (Dawson & Stein, 2008; Stein, et al., 2010).

The process begins with the selection of subject-area specific concepts and skills in collaboration with educators. Researchers then set out to study how students learn these concepts and skills. Using methods from cognitive developmental psychology, researchers build a set of empirically-grounded *learning sequences*, which present in detail the development of concepts central to the topics targeted by the assessment (Dawson & Stein, 2008; Dawson-Tunik, 2004). These learning sequences inform the creation of *low-inference rubrics*, which are tested, calibrated to assure standardization, and finally used by teachers and students to score mature DiscoTests. The learning sequences also inform the creation of topic-specific learning resources and lesson plans, which are vetted for quality and organized developmentally along the same scale as the learning sequences. Some of these curated learning resources make up a part of the individualized feedback each student receives after taking a DiscoTest. They also contribute to the teacher resource area in which individual and classroom-level profiles of student performance are accompanied by relevant learning resources and lesson plans.

Aside from their function in teaching and learning, DiscoTests also serve as standardized assessments that provide scores for all performances along a single learning scale[ii]. Over time, as students take DiscoTests in different subjects and grades, the accumulated history of their scores and reports forms a complex, detailed electronic learning record, providing rich data for researchers and detailed learning records for schools, districts, and educators. But before exploring the educative and measurement functions of DiscoTests, we discuss the research and development process in more detail.

Learning sequences and low inference rubrics

The lineage of constructivism invoked above, including Baldwin, Piaget, and Fischer, has shown that improvement in cognitive performance reflects more than the incremental accumulation of information. Both knowledge and cognitive skills develop through a series of hierarchically organized levels (or *stages*). Each successive level is more complex and abstract than the level that preceded it. In other words, successive levels are more abstract, complex, and integrated. Development within a level manifests as an increasingly elaborate repertoire of knowledge and skills at that level. Moving from one level to the next occurs when the current level reaches a tipping point, the system reorganizes, and a new way of thinking emerges. This developmental model describes a *natural* virtuous cycle of learning.

During the latter years of the 20th century, several researchers developed metrics based upon this developmental model, ultimately identifying 13 qualitatively distinct levels through which knowledge and cognitive skills develop. The most precise of these metrics is the Lectical® Assessment System (LAS) (Dawson, 2010). The most recent version of the LAS can reliably measures progress through the last 7 of these levels in ¼-level increments called *phases*[iii], and has been shown to have a validity and reliability profile that make it suitable for use in almost any assessment context, from the classroom to high-stakes testing (*ibid*). The LAS is at the core of the DiscoTest research and development process, and is the common scale along which student performances, learning sequences, rubrics, and learning resources are aligned, permitting the delivery of developmentally appropriate educative feedback.

The LAS makes it possible to use a combination of longitudinal and cross-sectional data to construct accurate scoring rubrics and learning sequences that detail the development of specific concepts and skills (Dawson-Tunik, 2004). The research process for building rubrics and learning sequences begins with the design of an interview instrument composed of a set of open-ended items. Then, researchers conduct probed, clinical interviews. The interviews are independently (1) scored with the LAS to determine their developmental phase and (2) submitted to a comprehensive analysis of their content. When both analyses have been completed, analysts begin an iterative process of reconstructing the relation between level of performance and conceptual content. There are several steps in this process:

1. Identify themes and subthemes into which codes can most readily be divided, then
2. produce an organized empirical inventory of conceptions and exemplars, organized by themes and subthemes, and ordered according to the developmental phases in which particular conceptions were found, and
3. vet the integrity of this inventory, exploring other possible thematic combinations of concepts, ensuring that exemplars represent the concepts with which they are associated, and beginning the process of identifying defensible learning sequences.

At this point, analysts begin rubric and sequence construction, which requires identifying clear evidence of conceptual learning sequences, and describing how specific concepts develop over time. We call this process *rational reconstruction*.

Table 1: The physics of energy—"What is happening to the energy of the ball as it hits the floor?"

Exemplar	**Conceptions**	**Sequence description**
It was falling down really fast, then it hit the floor and bounced back up, but not as fast. When I have lots of energy, I run really fast.	energy is motion; energy is a feeling that makes a person or animal *want* to move	Experience with moving objects, combined with the use of the term *energy* to describe inner states connected with a desire for motion, provide a concrete basis for connecting *energy* with *movement*.
It had a lot of energy when it was falling down, but it lost some of its energy when it came back up.	energy is something that is *associated with* motion	The connection of *energy* with *movement* at the previous level sets the stage for the differentiation of these concepts. Students begin to understand that energy and movement are not exactly the same thing, even though energy is always associated with movement.
It had the most energy when it hit the ground, because it was speeding up while it was falling, then some of its energy went into the floor and made a noise.	energy is a kind of quasi-substance that moves between objects; energy causes movement, heat, and sound	Once the relation between energy and motion is understood as an *association*, it is possible to further differentiate these concepts. Energy can be connected to heat or sound (as a cause), and early conceptions of *kinetic* energy (the energy of motion) and *potential* energy (the potential for energy to happen) emerge. These conceptions are essential for a beginning understanding of energy transfer, in which energy is viewed as a kind of quasi-substance that moves between objects.
It had the most kinetic energy when it hit the ground, because that was when it was moving fastest, but when it hit the ground, some of the	energy can take different forms and exist in different states	Once energy has been differentiated into states and forms, it is possible to construct an understanding of transformations. These can be represented in quantitative terms,

kinetic energy was transformed into sound energy and elastic potential energy.		allowing students to prove to their own satisfaction that energy can neither be created nor destroyed.

Relying upon the empirical sequences, developmental theory, and knowledge about related learning sequences, analysts attempt to explain why one form of a concept might follow another in an empirical sequence. Table 1 illustrates the flow of this work. Beginning with exemplars that are associated with particular sub themes and phases, analysts gradually tease out chains of thematically related conceptions that span the targeted developmental range. First, they describe the concepts represented by groups of similar exemplars (as shown in the second column of Table 1). Once these descriptions have been completed, analysts examine how they are related to one another across levels (as shown in the third column of Table 1). When this process is successful, each new level of understanding can be seen to build upon the conceptions of the preceding level. The full process of rational reconstruction is described more thoroughly elsewhere (Dawson-Tunik, 2004; Dawson & Stein, 2008).

During the rational reconstruction process, analysts identify a number of learning sequences like the one in the energy example. These sequences, along with the organized empirical inventory of conceptions and exemplars, inform the construction of low-inference rubrics. The learning sequence described in Table 1 contributed, along with other sequences, to the development of several rubrics focused on kinetic energy, potential energy, and energy transfer/transformation.

Although rubrics are informed by learning sequences, and scores calculated from rubric scores can be aligned along the same scale as the learning sequences, low inference rubric selections do not look much like learning sequences. Instead, rubric selections are more like the specific things students say in response to a particular problem. Table 2 shows one of 5 rubrics (for kinetic energy, potential energy, forces (gravity), work, energy transformation, and energy conservation) used to score student responses to an item that asks students to describe the energy of a *falling ball*. Each way of describing the energy of a falling ball is associated, in an empirical inventory of conceptions and exemplars, with the particular developmental phase during which it became common in student performances.

Table 2: Rubric codes for kinetic energy—falling ball scenario

Codes	**Score**				
	1	**2**	**3**	**4**	**5**
Mentions energy but fails to explain how it relates to the problem	X				
Claims that the ball has energy while it is falling		X			
Claims that the ball's energy increases as it falls			X		
Claims that the ball's energy increases as it falls because it speeds up				X	
Claims that the ball's KINETIC energy increases as it falls because it speeds					X

up (or accelerates)					

The ultimate goal of rubric construction is to make teacher and student friendly coding menus that can be used in the classroom to score DiscoTests in real time. The research techniques used to create the learning sequences provide a window into the development of key concepts, and insights into how these concepts are expressed at different levels of complexity. This is the basis of the rubric creation process, which aims to capture as clearly as possible the key developmental differences in student performances. These rubrics are powerful and informative because they are built from an empirically grounded, developmentally organized inventory of subject-specific concepts. And because the rubrics are based on research into how students learn targeted concepts, they can be used to generate both a developmental score for and targeted educative feedback.

Learning resources

The developmental differences revealed by the learning sequences also inform the compilation of related learning resources. Curriculum specialists work to create, compile, and curate subject-specific learning resources. The materials are then organized in terms of the complexity of their task demands and their conceptual focus, allowing them to be linked to rubric-derived scores and delivered to teachers and students. Ideally, these learning resources will support both teacher and student learning, and foster increasingly effective teaching and learning.

Testing the new DiscoTests

Once an initial set of rubrics is devised, and learning resources are compiled, the new DiscoTest can undergo a first round of testing. Two to three rounds of testing are required to refine coding menus, check the accuracy of learning sequences, evaluate item functioning, and optimize reliability.

Educative and measurement functions of DiscoTests

The extensiveness of the research and development process, and the complexity of the final result make DiscoTests unlike conventional standardized tests in many ways. Once built, a DiscoTest can be used indefinitely; students can take the same one several times without exhausting its potential to help them gain an increasingly sophisticated understanding of targeted concepts. This is because the items are deliberately constructed to be answerable at several different levels of sophistication. Moreover, because the primary role of DiscoTests is educative (and items don’t have single correct answers) concerns about cheating are minimal. Furthermore, as already mentioned, DiscoTests are both educative and standardized. All performances are placed on the same domain independent, general scale. This makes it possible to compare learning trajectories across any range of subjects or contexts. This increases the power of DiscoTests as data collection instruments. Eventually they will yield large longitudinal databases that will allow researchers and administrators to construct increasingly refined accounts of how students learn important skills and concepts.

We have discussed how new technological possibilities are changing the shape of education, and testing in particular. As the technological base of our educational practices have transformed, so has the broad function of testing. By combining advances in technology with the new science of learning, DiscoTests are able to facilitate virtuous cycles of learning at multiple levels in the educational system. As noted above, this is a powerful way to fit the best of what we know about teaching and learning into tomorrow's educational technologies. In the next section we provide an overview of the central educative and measurement function of DiscoTests, looking at what they do for students, teachers, administrators, and researchers.

Tests that help students learn

DiscoTests support student learning by gathering evidence about the level of a student's performance, then providing the student (and his or her teacher) with rich feedback that points to the specific concepts and skills that student is most likely to benefit from learning next. DiscoTests engage students in a virtuous cycle wherein their responses prompt immediate and useful feedback about where they are in their learning and how they might improve. This feedback sets the stage for further learning and, in time, another round of assessment, which in turn generates new feedback, and so on.

As students take repeated DiscoTests across numerous subject areas, a digital learning record is produced—the DiscoTest Report Card (see Figure 1). This report card presents a complex picture of the student's conceptual development, shows a variety of growth trends, and connects to a report for every DiscoTest taken by the student.

------------------------Insert Report Card about here----------------------------

Tests that help teachers teach

DiscoTests don't just help learners; they also help educators. DiscoTest reports (1) tell educators how well students understand and think with new ideas; (2) provide diagnostic information that aids instruction; and (3) show how close students are to achieving learning goals. DiscoTests measure more than just what student know or don't know; they inform educators about the way students apply the material they are learning. Moreover, DiscoTests provide educators with a window into the learning process itself, allowing them to continually improve their grasp of how learning unfolds in their domain. With repeated use of DiscoTests as part of a reflective practice, educators, along with their students, engage in a virtuous cycle of learning.

------------------------Insert Teacher feedback about here----------------------------

Tests that help leaders understand and develop schools

Standardized tests have been used for nearly a century by educational leaders seeking information about how their institutions are performing. Contemporary educational reform efforts are built on the idea that standardized tests can be used to measure the effectiveness of new polices and practices. As noted in our account of the history of testing, recent trends in particular have emphasized the use of tests for accountability purposes, monitoring the effectiveness and improvements of schools and districts. Standardized test have become, for better and worse, an instrumental part of managing our school systems.

These system-monitoring standardized tests may be able to provide an overview of who is doing well and who is not, but this knowledge is divorced from any insight into the learning processes in question. Most traditional tests say nothing about the kinds of confusions that are common or about the range of conceptions and ideas held by different individuals or groups and how these are related to future prospects for learning.

As already noted, DiscoTests differ from most standardized tests in important ways. Because they measure the major transformations that occur during learning they provide a score that is meaningful as an index of important learning events. This means that it is possible to track the *learning* of individuals and cohorts over time, providing a more informative method for monitoring individual and cohort progress. Because they are standardized they can be used with specific ends in view, such as evaluating the effectiveness of curricula. Organizational leaders can use DiscoTests to institute virtuous cycles that will allow them to continually learn about how their policies and procedures are affecting the learning of the individuals in their organization.

------------------------Insert admin feedback about here----------------------------

Tests that help researchers build usable knowledge

Finally, all DiscoTests play a role as data collection instruments. Eventually they yield large longitudinal databases that allow researchers to construct increasingly refined accounts of the pathways through which students learn important skills and concepts. This will yield insights into how learning processes unfold in a wide range of important areas. Researchers will also be able to ask questions about the differential effects of pedagogical approaches or leaning environments and to model growth using rich data gleaned from frequent test times and complex indexes of student learning.

The DiscoTest initiative aims, in part, to facilitate the emergence of a cumulative knowledge base of research about teaching and learning. Because all DiscoTests are aligned along a common scale—the LAS—researchers can begin to ask important questions concerning, for example, the differential distribution of student capabilities across subject areas (from history to physics), the factors that shape the educational trajectories of individuals students (thanks to detailed student learning profiles), and the value added of innovations in pedagogy or curricula (both within and between subject areas). Researchers themselves are engaged in a virtuous cycle of learning that becomes a kind of action research, as the research instruments themselves directly benefit student learning and teacher practice (Dawson & Stein, 2008).

Discussion: toward learning-centric testing technologies

The DiscoTest initiative has served as an example of what new computer technologies are making possible for large-scale standardized testing infrastructures. Importantly, this approach to building standardized testing infrastructures conforms to the design parameters listed above, which codify key lessons from the learning sciences that we believe ought shape the development of new testing technologies. It is worth noting more specifically how DiscoTests fit these design parameters, clarifying it as an example of learning-centric testing technology.

DiscoTests are administered on-line and with minimal software and hardware requirements, making them *broadly available*. The ideals of universal access and universal design that infuse a great deal of educational technology could be leveraged to bring high-quality tests to everyone. DiscoTests are built to be *embedded* in the curriculum and *relevant* to the lives of students, making them *low-stakes* and capable of disappearing into the flow of classroom practice. They aid pedagogy through their *formative* and *diagnostic* functions and are built firstly to *support teaching and learning*. The broad goal of building unobtrusive teacher-enabling educational technology has important consequences for testing. Informed by the learning sciences, new forms of testing, like DiscoTest, should move us towards tests that do not impose upon the teacher and student, but rather catalyze better teaching and learning.

Because DiscoTests are *standardized* and will yield large databases of detailed student learning records, they are useful for informing system-level decision making and *building knowledge* about learning and educational reform in general. Emerging educational technologies are already gathering unprecedented amounts of data about student performances. It is important that new testing infrastructures be built that can handle the “data-deluge” produced by real-time embedded assessments. Data needs to be made useful to policy makers and researchers, who will able to ask new questions about student growth, the causes differential performance, and the effects of low-stakes testing on motivation and development.

This example shows that the interface of learning science and new technology could produce tests capable of facilitating multi-level learning in educational systems. Of course, there is nothing that insures that DiscoTest or other comparable learning-centric assessment approaches will take standardized testing in the 21st Century in positive new directions. Not all testing made possible through computer technologies will be in the service of learning. If current trends continue tomorrow’s testing infrastructures could be simply more complex forms of the high-stakes, accountability-oriented tests that dominate education today. It is important that educators seize the opportunity to shape the use of emerging educational technologies in way that are beneficial to students. Large-scale standardized testing infrastructures in particular are likely to remain a key element in schooling and provide a unique point of leverage for changing schooling practice broadly (Stein, et al., 2010)

Conclusion: technogies that change edutcaion change tetsing

This discussion has been guided by questions about how to fit the best of what we know about teaching and learning into tomorrow’s educational technologies. In particular, the future of large-scale standardized testing infrastructures was considered. The history of standardized testing reviewed suggests that as the technological base supporting education shifts so do many of its core practices, including testing. Since the end of WWII standardized testing infrastructures have grown increasingly large and are now at the center of federal educational policy. Yet these transformations and expansions of testing have been divorced from advances in the sciences of learning. As new communications technologies proliferate and the demands on educational systems change, testing practices will change as well. The question is to what extent these changes will be shaped in the interest of teachers and students.

Tomorrow's educational technologies will allow for frequent, embedded, formative assessments that dynamically deliver real-time feedback, blurring the lines between testing with learning. Advances in database technologies and networked computing will enable detailed electronic learning records for each student, as well as complex system-level growth modeling capabilities for administrators and researchers. And while these trends have yet to dislodge dominant testing practices, which still rely heavily on summative, high-stakes testing divorced from teaching and learning and deeply politicized (Hursh, 2008; Ravitch, 2010), the future is likely to be greatly influenced by the "disruptive" power of new educational technologies (Christenson, et al., 2008; Collins & Halverson, 2009). Testing in particular will change in the wake of the digital revolution. It is the responsibility as educators and learning scientists to shape these changes in ways that are beneficial to everyone, by working to construct a learning-centric testing infrastructure.

The DiscoTest Initiative was presented as an example of positive new technology-enabled directions for testing. It answers some questions about how to embed important lessons from the sciences of learning in new educational technologies. By leveraging technology to build tests that promote learning for everyone who uses them, DiscoTests make good on many of the enthusiastic claims made by those foreseeing a digital revolution in education.

Baldwin, J. M. (1906). *Mental development in the child and the race: Methods and processes*. New York: The Macmillan Company.

Beer, S. (1981). *The brain of the firm: the managerial cybernetics of organization*. Newy York: J. Wiley.

Brandom, R. (1994). *Making it Explicit*. Cambridge MA: Harvard University Press.

Brown, J. (1992). *The definition of a profession: The Authority of a metaphor in the history of intelligence testing*. Princeton Princeton University Press.

Buck, J., & Villines, S. (2007). *We the People: Consenting to a Deeper Democracy; A Guide to Sociocratic Principles and Methods*. Washington DC: Sociocracy.info.

Bush, G. H. W. (1991). America 2000: address to the nation on national educational strategy, from http://bushlibrary.tamu.edu/research/public_papers.php?id=2895&year=1991&month=4

Callahan, R., E. (1964). *Education and the cult of efficiency*. Chicago: University of Chicago Press.

Campbell, D. T. (1975). Assessing the impact of planned social change. In G. M. Lyons (Ed.), *Social research and public policy: the Dartmouth/OECD Conference* Hanover, NH: Public Affairs Center, Dartmouth College.

Campbell, D. T. (1987). Evolutionary Epistemology. In Radinitzky & Bartly (Eds.), *Evolutionary epistemology, theory of rationality, and the sociology of knowledge*. La Salle, IL: Open Court.

Chapman, P. (1988). *Schools as sorters: Lewis M. Terman, applied psychology, and the intelligence testing Movement, 1890-1930*. New York: New York University Press.

Christenson, C., Johnson, C., & Horn, M. (2008). *Disrupting class: how disruptive innovation will change the way the world learns*. New York: McGraw-Hill

Collins, A., & Halverson, R. (2009). *Rethinking education in the age of technology: the digital revolution and schooling in America*. New York: Teachers College Press.

Cremin, L. (1988). *American education: the metropolitan experience, 1876-1980*. New York: Harper and Row.

Darling-Hammond, L. (2010). *The flat world and eductaion*. New York: Teachers college press.

Dawson, T. L. (2010). The Lectical™ Assessment System Retrieved March 30, 2010, from http://lectica.info

Dawson, T. L., & Stein, Z. (2008). Cycles of research and application in education: Learning pathways for energy concepts. *Mind, Brain, & Education, 2*(2), 90-103.

Dawson-Tunik, T. L. (2004). "A good education is…" The development of evaluative thought across the life-span. *Genetic, Social, and General Psychology Monographs, 130*(1), 4-112.

Dewey, J. (1929). *The sources of a science of education*. New York: Liveright.

Elgin, C. (1996). *Considered Judgment*. Princeton, New Jersey: Princeton University Press.

Farley, T. (2009). *Making the grades: my misadventures in the standardized testing industry*. Sausalito, CA: PoliPoint Press.

Fischer, K. W. (1980). A theory of cognitive development: The control and construction of hierarchies of skills. *Psychological Review, 87*, 477-531.

Fischer, K. W. (2009). Mind, brain, and education: building a scientific groundwork for learning and teaching. *Mind, Brain, and Education, 3*(1), 3-16.

Fischer, K. W., & Bidell, T. R. (2006). Dynamic development of action, thought, and emotion. In W. Damon & R. M. Lerner (Eds.), *Handbook of child psychology: Theoretical models of human development* (6 ed., Vol. 1, pp. 313-399). New York: Wiley.

Forrester, J. (1964). *Industrial dynamics*. Cambridge, MA: MIT Press.

Gould, S. J. (1981). *The mismeasure of man*. New York: Norton.

Haney, W. M., Madaus, G. F., & Lynos, R. (1993). *The fractured market place for standardized testing*. Norwell, MA: Kluwer Academic.

Hess, F., & Petrilli, M. (2006). *No Child Left Behind*. New York: Peter Lang.

Hoffman, B. (1962). *The tyranny of testing*. New Yerk: Crowell-Collier Press.

Hursh, D. (2008). *High-stakes testing and the decline fo teaching and learning*. New York: Rowman & Littlefeild.

Koretz, D. M. (2008). *Measuring up: what educational testing really tells us*. Cambridge: Harvard University Press.

Lagemann, E. (2000). *An elusive science: the troubling history of educational research*. Chicago: University of Chicago Press.

Lemann, N. (1999). *The big test: the secret history of the American meritocracy*. New York: Farrar, Straus and Grioux.

Mareschal, D., Johnson, M., Sirois, S., Spratling, M., Thomas, M., & Westermann, G. (2007). *Neuroconstructivism: Volumes I & II* Oxford: Oxford Univiertsiy Press.

Nairn, A. (1980). The rein of ETS: the corporation that makes up minds: The Ralph Nader Report on the Educational Testing Service.

National Research Council committee on testing and assessment. (2001). *Knowing what students know: the science and design of educational assessment*. Wahington, D.C.: National Academy Press.

Nichols, S. L., & Berliner, D. C. (2007). *Collateral damage: how high-stakes testing corrupts America's schools*. Cambridge: Harvard Education Press.

NSF task force on cyberlearning. (2008). Fostering learning in the networked world: the cyberlearning opportunity and challenge. Washington, DC: National Science Foundation.

O'Donnell, J. M. (1985). *The origins of behaviorism: American psychology 1870-1920*. New York: New York University Press.

Obama, B. (2008). Speech to the 146th annual meeting of the 87th representative assembly of the national education association. .

Phillips, K. R. (2003). *Testing controversy: a rhetoric of education reform* Cresskill, NJ: Hampton Press.

Piaget, J. (1960). *The psychology of intelligence*. Patterson, New Jersey: Littlefield, Adams.

Ravitch, D. (2010). *The life and death of the great American school system: how testing and choice are undermining education*. New York: Basic Books.

Sacks, P. (1999). *Standardized minds: the high price of America's testing culture and what we can do to change it*. Cambridge, MA: Perseus Press.

Samelson, F. (1990). Was early mental testing: a) racist inspired, b) objective science, c)a technology for democracy, d) the origen of the multiple-choice exam, e) none of the above. In M. Sokal (Ed.), *Psychological testing in american society: 1890-1930* (pp. 113-128). New Brunswick: Rutgers University Press.

Schwartz, H., Hamilton, L. S., Stecher, B. M., & Steele, J., L. (2011). Expanded measures of school performance. Aliginton, VA.: RAND Education.

Senge, P. M. (Ed.). (1990). *The fifth discipline: The art and practice of the learning organization*. New York: Doubleday.

Sokal, M. (Ed.). (1990). *Psychological testing in american society: 1890-1930*. New Brunswick: Rutgers University Press.

Spitzer, M. (1999). *The mind within the net: Models of learning, thinking, and acting*. Cambridge, MA: Massachussetts Institute of Technology.

Spring, J. H. (1989). *The sorting machine revisited: national eductaional policy since 1945*. New York: Longman.

Stein, Z., Dawson, T., & Fischer, K. W. (2010). Redesigning Testing: Operationalizing The New Science of Learning. In M. S. Khine & I. M. Saleh (Eds.), *New Science of Learning: Cognition, Computers and Collaboration in Education* (pp. 207-224). New York: Springer.

Toch, T. (2006). Margins of error: the education tetsing industry in the No Child Left Behind Era. Washington D.C.: Education Sector.

Tomasello, M. (2005). *Constructing a language: a usage-based theory of language acquisition*. Cambridge MA: Harvard University Press.

US department of education. (2011). Race to the top assessment program Retrieved July, 2011, from http://www2.ed.gov/programs/racetothetop-assessment/index.html

White house office of the press secretary. (2012). Adminsitration's statement on educational policy. Retrieved July, 2012, from http://www.whitehouse.gov/issues/education

Zucker, A. (2008). *Transforming schools with technology: how smart use of digital tools helps achieve six key educational goals*. Cambridge, MA: Harvard Education Press.

[i] These tests will be available to individual teachers and students at no cost (to ensure that the least advantaged are served), and delivered to schools and school districts for a low per-assessment fee. Fees will support the maintenance and development of DiscoTests and the dissemination of knowledge about learning.

[ii] This universal learning scale is Fischer's Skill Scale (Fischer, 1980; Fischer & Bidell, 2006) as operationalized by the Lectical Assessment System (Dawson, 2010), as discussed in the body of the text. Importantly, supplemental to scores for developmental level, DiscoTests are also scored for argumentation quality, using a set of standard rating scales. Students receive ongoing feedback about their developing argumentation skills as part of the student report, and teachers can track progress at the level of the classroom. Students also receive feedback about how often they recognize the salience of concepts targeted by test items by mentioning them in their answers.

[iii] Five to seven of these phases are generally identified in an individual K-12 classroom.