The Gordon Commission
on the Future of Assessment in Education

Technological Implications for Assessment Ecosystems: Opportunities for Digital Technology to Advance Assessment

John T. Behrens
Pearson

Kristen E. DiCerbo
Pearson

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educator is to create, together with the students, the conditions under which knowledge at the level of the doxa is superseded by true knowledge at the level of the logos.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 & 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, cell phones' spatial positioning and accelerometer

information 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, has 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 policy makers (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 tasks. 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 multi-user 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, Contexualization, 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 to 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 of 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 now 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, & 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 assessment does 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 on-line, what we do with our money (bank records), and where we look on-line. 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 every day 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 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.