SYNTHESIS REPORT ON THE GAMES, LEARNING, AND ASSESSMENT (GLA) WORKSHOP

Valerie Shute, Matthew Ventura, & Yoon Jeon Kim (Florida State University)

INTRODUCTION

GAMES, LEARNING, AND ASSESSMENT

> *That's what learning is, after all; not whether we lose the game, but how we lose and how we've changed because of it and what we take away from it that we never had before, to apply to other games.* Richard Bach (1984)

Scholars from various disciplines have recently shown increasing interest in figuring out how well-designed digital games[1] can be used to support learning (e.g., Gee, 2003; Prensky, 2006; Shaffer, Squire, Halverson, & Gee, 2005; Shute, Rieber, & Van Eck, 2011). A common motivation for studying games as a vehicle to support learning is frustration with the current education system and a desire for alternative ways of teaching – ways that increase student engagement and yield a rich, authentic picture of the learner(s). The main problem boils down to this: most schools in the U.S. are not adequately preparing kids for success in the 21st Century (e.g., Partnership for 21st Century Skills, 2006). Learning in school is still heavily geared toward the acquisition of content within a teacher-centered model, with instruction too often abstract and decontextualized and thus not suitable for this age of complexity and interconnectedness (Shute, 2007). One downside of this adherence to outdated pedagogy is that other developed countries of the world are surpassing us on measures of important competencies (e.g., mathematics problem solving) as assessed by international tests such as the PISA and TIMSS (Gonzales, Williams, Jocelyn, Roey, Kastberg, & Brenwald, 2008; Howard, Paul, Marisa, & Brooke, 2010).

Why should we consider using games as vehicles to assess and support learning? Digital games are a significant part of young people's everyday lives. The Pew Internet and American Life Project surveyed 1,102 youth between the ages of 12 and 17. They reported that 97% of youth – both boys (99%) and girls (94%) – play some type of digital game (Lenhart et al., 2008). Also, a

[1] We use the term "digital game" to refer to a wide range of computer-based, video, mobile, and other electronic games.

recent study by the Kaiser Family Foundation reported that youth (between the ages of 8 and 18) spend an average of nearly six and half hours a day with media including one hour and thirteen minutes playing video games on gaming consoles and mobile devices (Rideout, Foehr, & Roberts, 2010). This finding is fairly consistent among different racial groups. Additionally, Ito and her colleagues (2010) found that playing video games with friends and family is a large and normal part of the daily lives of youth. They further observed that playing video games is not solely for entertainment purposes. In fact, many youth enthusiastically participate in online discussion forums to share their knowledge and skills about a game with other players, or seek help in relation to challenges when needed. Kids use a variety of video- and picture-editing tools to share their playing strategies on the Internet, and sometimes even learn how to modify the game (i.e., *modding*), which requires advanced computer technology skills.

The main claim that researchers make about digital games is that games can facilitate learning because they provide an engaging learning environment thus learners *want* to play and learn (e.g., Gee, 2003; Salen & Zimmerman, 2005; Shaffer et al., 2006). Digital games are engaging due to their unique design elements. These elements include interactivity, immediate and ongoing feedback, adaptive levels of challenge, unpredictability, unique controls, and visual and auditory stimulation (Fullerton, 2008; Shute, Rieber, & Van Eck, 2011). The theories of learning most aligned with these environments include constructivism and situated learning theories (Barab et al., 2007; Brown, Collins, & Duguid, 1989; Duffy & Jonassen, 1992; Lave & Wenger, 1991; Piaget, 1950; Vygotsky, 1978). The idea is that the learner is active in the learning process, and learning is the result of interaction with an interesting problem context where learners construct meaning. Thus well-designed games have the potential to be used in a variety of learning and assessment contexts (e.g., Gee, 2003; Salen & Zimmerman, 2005; Shute, Ventura, Bauer, & Zapata-Rivera, 2009).

Another affordance of digital games that has drawn scholars' attention is that games can collect enormous amounts of data from players, and then use that corpus of data to dynamically change gameplay in an engaging way (Gee, 2008). This ability to dynamically change the gaming environment based on player information is dependent on accurate and ongoing assessments of the player. Now, for these kinds of games to have a chance at being included in the K-16 curriculum, establishing the validity of such embedded assessments will be critical. Thus the next generation of engaging, educational games will require a team effort, involving game designers, learning, assessment, and content experts, and possibly others early in the design process. To address the issue of just how such a heterogeneous team could work together to come up with game-based assessment designs, the Gates and MacArthur Foundations funded the two-day Games, Learning, and Assessment (GLA) Workshop hosted by the University of Southern California (USC) School of Cinematic Arts.

The GLA workshop (January 27-28, 2011) aimed to facilitate discussion among assessment and learning experts, school principals, district leaders, and game-industry persons (designers, developers, and user researchers) to discover how best to create a research process that uses

principles of game design and contemporary learning theories to suit the needs of an assessment system. The first goal of the workshop was to see how feasible it is to combine assessment and game design expertise towards the design of new game-based assessments. A second goal of the workshop focused on putting assessment at the center of the discussion around games and learning rather than as an after-thought. The third goal of the workshop involved examining issues related to using games as assessment vehicles under the Common Core Standards.

Before discussing specific details of the GLA Workshop, we will give a brief overview of some of the major problems with our nation's assessment system at both the classroom- and high-stakes levels.

PROBLEMS WITH THE NATION'S ASSESSMENT PROCESS

Our country's current approach to assessing students has a lot of room for improvement, at the classroom as well as the high-stakes levels. This is especially true in terms of the lack of support that current/traditional assessments provide for students learning new knowledge, skills, and dispositions that are important to succeed in today's complex world. The current means of assessing students infrequently (e.g., at the end of a unit or school year, for grading and promotion purposes) can cause various unintended consequences, such as increasing the dropout rate given out-of-context and often irrelevant test-preparation teaching contexts that the current assessment system often promotes. The goal of an ideal assessment policy/process should be to provide valid, reliable, and actionable information about students' learning and growth that allows stakeholders (e.g., teachers, students, administrators, and parents) to utilize the information in meaningful ways. Following are some specific problems with our current assessment system.

TRADITIONAL CLASSROOM ASSESSMENTS RARELY INFLUENCE INSTRUCTION OR LEARNING

Many of today's classroom assessments do not support deep learning or the acquisition of complex competencies. Current classroom assessments are typically designed to judge a student (or group of students) at a single point in time, without providing diagnostic support to students or diagnostic information to teachers. Instead, assessments *should* be used to: (a) support the learning process for students and teachers; (b) interpret information about understanding and/or performance regarding educational goals (local to the curriculum, and broader to the state or Common Core Standards); (c) provide formative, compared to summative information (e.g., give useful feedback during the learning process rather than a single judgment at the end); and (d) be responsive to what's known about how people learn – generally and developmentally.

TRADITIONAL ASSESSMENTS ARE NOT KEEPING PACE WITH THE NEEDS OF A COMPLEX WORLD

A century ago, a person who acquired basic reading, writing, and math skills was considered to be sufficiently literate. The goal back then was to prepare young people for production jobs because 90% of the students were not expected to seek or hold professional careers (Shute, 2007). But when faced with highly technical and complex problems in today's world, the ability to solve ill-structured problems and think creatively, critically, collaboratively, and systemically is essential (e.g., Shute & Torres, in press; Walberg, & Stariha, 1992). These skills are what many educators call 21st Century (or complex) competencies. Except in rare instances, our current education system neither teaches nor assesses these new complex competencies despite a growing body of research showing that competencies like critical thinking, persistence, creativity, self-efficacy, openness, and teamwork (to name a few) can substantially impact student academic achievement (Noftle & Robins, 2007; O'Connor & Paunonen, 2007; Poropat, 2009; Sternberg, 2006; Trapmann, Hell, Hirn, & Schuler, 2007).

TRADITIONAL ASSESSMENTS AND VALIDITY ISSUES

Assessments are typically evaluated under two broad categories: reliability and validity. Reliability is the most basic requirement for an assessment and is concerned with the degree to which a test can consistently measure some attribute over similar conditions. For example, if a bathroom scale indicated that a person weighed 120 pounds in the morning and then 200 pounds later in the afternoon, we would say that the scale is likely an unreliable measure of weight. In assessment, reliability is seen when a person scores really high on an algebra test at one point in time and then scores similarly on a comparable test 1-2 days later. Reliability is relatively easy to establish because it is not concerned with what the assessment is actually measuring or how "true" scores are interpreted in the real world. Validity, on the other hand, addresses this second issue. Broadly, validity refers to the extent to which an assessment actually measures what it is intended to measure. Here are specific validity issues related to traditional assessment.

Face validity. Face validity states that an assessment should intuitively "appear" to measure what it is intended to measure. For example, reading some paragraphs on an uninteresting or irrelevant topic and answering multiple-choice questions about it is not the best measure for reading comprehension (i.e., it lacks good face validity). Students need to be assessed in meaningful environments[2] rather than filling in bubbles on a prepared form in response to decontextualized questions. Digital games can provide such meaningful learning environments by providing students with scenarios that require the application of various competencies, such as reading comprehension. Consider role playing games (e.g., *World of Warcraft*). In these games, players

[2] Meaningful learning environments are complex, authentic, situated, relevant, interactive and information-rich environments that engage students with purposeful problem solving and active interactions.

must read often lengthy and complex quest logs that tell them the goals in the game. Without comprehending these quest instructions, the players would not be able to know how to proceed and succeed in the game. This seemingly simple task in role playing games is, in fact, an authentic assessment of reading comprehension, and thus would have good face validity. Without these situated and meaningful assessments, we cannot determine what students can actually do with the skills and knowledge obtained. Thus new, embedded types of assessment methods are needed to properly assess valued competencies (Shute, 2009).

Predictive validity. Predictive validity refers to an assessment predicting future behavior. Today's large-scale, standardized assessments are lacking in this area. For example, a recent report from the College Board found the SAT only marginally predicted college success beyond high school GPA at around r =.10 (Korbin, Patterson, Shaw, Mattern, & Barbuti, 2008). This means the SAT scores contribute around 1% of unique prediction to college success after controlling for GPA information. However, a single construct like "grit" (i.e., persistence plus passion) has been shown to account for an average of 4% of the variance related to various academic outcomes after controlling for GPA (Duckworth, Peterson, Matthews, & Kelly, 2007). These findings suggest a need for assessments that measure important new constructs *and* predict what kids can do beyond the test.

Consequential validity. Consequential validity refers to the effects of a particular assessment on society and policy decisions. One negative side effect of the No Child Left Behind (NCLB, 2002) initiative, with its heavy focus on accountability, has been teachers “teaching to the test” thereby reducing the face validity of such tests. That is, when teachers instruct content that's relevant to answering items on a test but not particularly relevant for solving real-world problems, this reduces student engagement in school which can lead to increased dropouts (Bridgeland, DiIulio, & Morison, 2006). Moreover, the lack of predictive validity of current assessments can lead to students not getting into college due to the inaccuracy or inappropriateness of the assessments. But the SAT and similar test scores are still being used as the main basis for college admission decisions which can lead to many students missing opportunities at fulfilling careers and lives, particularly disadvantaged youth. There is thus a need for new assessments that lead to positive (not negative) consequences for students.

As can be seen, traditional tests are not fully satisfying various validity requirements. In the next section we describe the workshop and its aim to address some of the aforementioned problems regarding assessment (specifically) and education (broadly).

GLA WORKSHOP

The GLA workshop brought together professionals from diverse fields to the USC campus for two days. Attendees included game designers, assessment and learning experts, game user-researchers, content experts, practitioners, and Common Core Standards experts. Additionally, 10 USC graduate students participated and moderated the Day 2 breakout design sessions.

DAY 1 – WHY STUDY GAMES AND ASSESSMENT?

The morning of the first day of the workshop consisted of presentations and discussions. The presentations suggested various ways that well-designed games can be used for assessment purposes. Speakers included: Ron Chaluisan, Tracy Fullerton, Jim Gee, Zoran Popovic, Katie Salen, Val Shute, Keith Steury, and Will Wright. Below are a set of important topics that arose from these discussions that shed light on reasons why we should explore games further for assessment purposes.

Interaction and Control. Games require interaction between the player and the game environment. This interaction puts the player in control of the learning situation. Games do not allow players to "zone out" or skip challenges since constant player-game interaction is what drives and sustains gameplay. These interaction and control features provide a strong motivation to learn when the student can move at her own pace and test developing hypotheses. As Katie Salen stated, games provide a way for students to be in control of the learning experience while the classroom environment typically does not. Moreover, Jim Gee described how students like to be in charge of the learning process by "doing and creating." Players can often, in one way or another, customize a game to fit with their learning and playing preferences. For example, in the game Little Big Planet, players can create "levels" or obstacle courses that they can then post within the game for others to play online. Currently there are over two million community levels posted in Little Big Planet. Each community level can be rated on a number of dimensions and sorted by players.

Training and Learning. Jim Gee also described how games are excellent at training and scaffolding players until they have automated their skills. Players learn the rules of games readily due to the requirement of "doing" not memorizing. These *processes* of playing (and learning) within the game should be the focus in game-based assessment, which contrasts with the focus on just the outcomes of learning in classrooms. Once mastery of some skill or level is attained, a game may introduce a new class of problems requiring players to rethink their mastery status, learn something new, and integrate this new learning with their old knowledge, skills, and/or dispositions. In turn, new mastery is consolidated through repetition (with variation), only to be challenged again. This cycle has been called the "cycle of expertise" (Bereiter & Scardamalia, 1993), a prominent theory on how expertise is achieved. In school, sometimes the lower-achieving students do not receive enough opportunities to consolidate, and the higher-achieving students do not receive adequate challenges to their school-based mastery.

Challenge and Assessment. Good games tend to stay right at the outer edges of a player's ability. That is, good game designers seem to know just how to strike the right balance to challenge players without making the game too easy (i.e., boring) or too difficult (i.e., frustrating). This is a highly motivating state for learners and it is ignored or missed too often in many learning environments. Will Wright stated that this notion of challenge is predicated on assessments in games. That is, a good game uses player information to guide the natural progression through a

game in varying levels of difficulty. These levels of difficulty may require multiple competencies, and when a player attempts a solution, appropriate feedback is given about performance either explicitly (e.g., scores, dying), or implicitly (e.g., increased difficulty). The notion of challenge also resonates with Jim Gee's observation that games provide an excellent way to integrate learning and assessment, where learning takes place when players work on the cusp of their capabilities, or in the "zone of proximal development" (Vygotsky, 1978). The game then assesses the player's competency level and moves on to a more difficult level, if warranted.

Games Require Multiple Competencies. Val Shute and Jim Gee both pointed out that games often require a player to apply various competencies to succeed, such as thinking about different relationships among variables to solve a complex problem. For instance, in a strategy-based game like *Starcraft*, players must think of the ramifications of their actions not only on aspects of building their own interstellar galaxy, but also in relation to other players' interstellar galaxies (e.g., how to strategically collect resources that will grow faster than those of competing players). In addition, many games require divergent thinking to solve ill-structured problems, encourage players to thoroughly explore a space before moving on, and call for players to work in teams to solve complex problems. This combination of competencies (e.g., creativity, problem solving, persistence, and collaboration) is important inside and outside of games.[3] Furthermore, the competencies required to succeed in many well-designed games also happen to be the same ones that companies are looking for in today's highly competitive economy (Gee, Hull, & Lankshear, 1996).

Enormous Amounts of Data and Graphical Representations. Zoran Popovic and Jim Gee both noted that games provide rich sets of data over long periods of time about players. Furthermore, gameplay information is often visually displayed which helps players see, at any time, the outcomes of their actions. Currently, most of this rich stream of data is only being used for descriptive purposes. However, assessment experts can mine and exploit the data in a myriad of ways to make inferences about player competencies over time using data mining and/or stealth assessment approaches (e.g., Shute at al., 2009; Woolf et al., 2010). These competency estimates can be used to adjust gameplay or provide more explicit support for learning via feedback, hints, and so on. Also, large amounts of data are needed to satisfy the many statistical requirements needed to build reliable and valid assessments.

Games and Learning Communities. Games have the potential to motivate players to create and contribute to learning communities centered around the game. These learning communities are typically Internet-based interest groups which provide another rich context where players can learn and develop expertise. Currently there are hundreds if not thousands of wiki sites that give support on how to play particular digital games. These sites also provide options for players to

[3] Note that persistence is particularly relevant to gaming because learning how to master an aspect of a game or solve a particularly hard problem can be difficult and requires the will to work hard despite repeated failure (Peterson & Seligman, 2004).

voluntarily share their knowledge and skills to support each other's learning. The learning communities are important for a range of players, including novices, as they address many common questions regarding rules of the game as well as more advanced strategies for success. For instance, Zoran Popovic described the game *Foldit*, a digital game that has a large online community. *Foldit* is a biochemistry-related game where the player must combine and/or manipulate strings of shapes (amino acids) to make proteins. Since the debut of *Foldit*, there have been numerous web sites created by users that are used for collaborative problem solving and support for the game. An additional phenomenon associated with *Foldit* is that some players with no background in biochemistry solve problems better than biochemists, suggesting *Foldit* has the potential to teach relevant biochemistry concepts (Copper et al., 2010). Finally, some biochemistry textbooks have incorporated *Foldit* problems in their exercises to help teach important concepts.

Embedded Assessment. Katie Salen noted that embedded assessments are a good way to dynamically measure competencies (i.e., knowledge, skills, and other attributes) in a digital game. There are several advantages to using embedded assessments in games. First, they don't interfere with learning since they operate in real time within the gaming system. Second, because embedded assessments operate in real time, they can be used (a) as the basis for providing timely and helpful feedback to players in order to improve performance, and (b) to adjust the environment to best suit players' current needs. Third, such assessments are able to capture more data than traditional assessments due to their dynamic nature and non-intrusiveness. Again, this advantage of collecting more data can lead to improvements to the reliability and validity of the assessment. Finally, when embedded assessments are invisible, this can remove test anxiety, again leading to a more valid assessment (Shute et al., 2009). These kinds of "stealth assessments" as they're called automatically pull out information directly (and in real time) from log files and makes inferences relative to states of particular competencies using statistical tools such as Bayesian inference networks. For more on the topic, and to see examples of how stealth assessments can operate within digital games (e.g., *Taiga Park*, *The Elder Scrolls IV: Oblivion*), see Shute (2011).

In short, there appear to be many assessment and learning affordances of digital games. The next section describes how the attendees at the workshop were able to work collaboratively to design games intended to support assessment and learning of important new competencies.

DAY 2 – DESIGN TEAMS

When setting out to design a new game, game designers may not initially focus on the particular competencies that may be evidenced during gameplay (e.g., physics understanding, critical thinking skills, systems thinking, empathy). Similarly, when educators set out to create new lesson plans or modules for instruction, they typically do not think about important game design principles and features (e.g., interactivity, ongoing feedback, adaptive levels of challenge, unpredictability). Thus a new collaboration is needed that teams up game designers and

assessment/learning/content experts to produce new game-based assessments that are reliable, valid, and fully engaging as compared to traditional assessments.

Each design team was assigned a subject area or competency that linked to the Common Core Standards (CCS). The reason for linking to the CCS was to show how games can be useful for assessing areas that are important to the nation as a whole. Each design team consisted of around 6-8 people with mixed areas of expertise. Design teams focused on a particular topic/skill with the goal of designing a game prototype that could serve to assess a player's competency level relative to the designated topic. Below are the descriptions of the eight breakout sessions and associated Internet links to the two-minute videos each group presented at the end of the workshop describing their game (login: username=*learning*, password=*assessment*). The first three descriptions include the processes of the team during the design phase as well as details about the ensuing game. The remaining descriptions simply describe the team's game prototype.[4]

Systems thinking. This group began by identifying behaviors that give evidence for systems thinking, defined as the ability to see "the big picture" in social, natural, and technological systems, as well as to understand the constituent parts within systems and their functional interrelationships. The team then began developing a game that required players to apply the skills defined by the systems thinking model. The first idea was a game based on the food cycle in an ecosystem. But the team soon discarded the idea of using the common predator-prey cycle as the basis for assessing systems thinking. They decided to keep some of the hierarchical cause-and-effect relationships from the ecosystem game, but to use more interesting characters that would be more exciting to kids. So the team decided to go with a fantasy horror theme called *Undead Apocalypse*. The game involved the main character (a warrior) who was in charge of maintaining a proper balance of vampires, zombies, and werewolves who were kept in a set of four pits inside a two-by-two space (i.e., four boxes from a bird's eye view).

The game consisted of just a few simple rules. First, vampires eat zombies, and zombies eat werewolves. Second, if vampires or zombies outnumber their prey by 3-to-1, then the prey gets eaten. If, over time, a particular quadrant contains just one type of monster (e.g., all zombies) then the player loses a life. Thus the player's goal is to maintain a careful balance of vampires, zombies, and werewolves in each quadrant. The player must also balance monsters by swapping them in adjacent quadrants (i.e., 1-to-1 swap). Once the player swaps monsters, the rules of predator-to-prey are applied in each quadrant. If there are no 3-to-1 advantages in a quadrant, a random monster dies in that quadrant. Thus there is a cycle of play that starts with a player swapping monsters, and then applying the predator-to-prey rule across all quadrants. This repeats until a player loses all lives. And like in the game *Tetris* (where a player also eventually dies), the goal is to accumulate as many points as possible, described next.

---

[4] The reason that we selected these three breakout sessions to discuss more fully is because each one of the authors of this report participated in each of the three sessions thus were privy to the details of the groups' conversations and thinking.

Systems thinking is assessed in *Undead Apocalypse* by calculating the efficiency of each monster swap. Since there are many monsters to swap, the player must choose the best swap given the conditions of the four quadrants. Like in chess, the game can calculate a rank ordering from best to worst move to minimize player death. When a player chooses a monster swap, the game scores the move and stores it in the player profile. Over time, each player accumulates scores which can be analyzed in a variety of ways (e.g., computing the average of all scores or plotting scores over time to see the player's growth on relevant skills). Variation in scores may be used to predict other competencies outside the game. http://learningparticipation.net/gla2011/group-videos/g-systems-thinking_480x360.html

College Knowledge. The main real world problem this team tackled concerned how to get more disadvantaged students into college. Current state and federal programs working on this issue have limited resources so not many students receive relevant services. As a result, many disadvantaged students have limited understanding about their college options, and often family members are not able to give them necessary guidance.

Preliminary discussions centered on the definition of college knowledge (e.g., financial aid options, different kinds of colleges, different admission requirements). The group then began designing a game that aimed to get kids thinking about choices and their consequences down the road, academically speaking. The conversation quickly shifted to gameplay. The game designer suggested various game mechanics (i.e., rules of the game and gameplay) based on the concept of controlling a large group of "students" moving through a maze, even though it was not clear what competencies were being investigated. On the morning of the second day, the assessment expert created a network of competencies that served as a new point of departure for deriving evidences and hence the game design and mechanics. At the top of the hierarchical network was college knowledge, and underneath it were three first-level variables: financial aid knowledge, causal reasoning skill, and decision making. Evidences for each of these would be certain actions taken in the game.

The goal for players in *Keys to College* (i.e., the working title of the game) is to herd the most people (i.e., students, represented by colored dice in the game) through good decisions within a maze of choices for success in life. The main lesson conveyed by the game is that choices made early (e.g., in middle school) constrain choices available later in life. The game begins with a large group of students starting out in middle school. Important decisions in the game revolve around math courses selected. For instance, pre-algebra was designed as a small door and most of the herd gravitated toward the larger door of "developmental math" (which is also more attractive because it is an easier course). Players getting through pre-algebra receive Key 1. To get through the algebra door later on in the game, one has to have the pre-algebra key. If the player does not have Key 1 upon arriving at the algebra door, the decision would be to either choose another door or go back and get Key 1. There were also "traps" in the game that pulled players toward them. These comprised the usual suspects – drugs, risky sex, and so forth. As the game progressed, doors got smaller, but once through the final door along the path of making good academic

decisions, many options would suddenly become available representing positive futures in professional, high skilled jobs. However, just one door awaited those with a high school diploma. Players would be assessed on their decision-making skills in the game relating to the choices made about doors and paths through the maze. http://learningparticipation.net/gla2011/group-videos/d-college-know_480x360.html

Choicelets. The goal for this group was to provide a game-based assessment relating to young people's ability to seek, filter, and share information in an area of interest. On the first day, the group spent a lot of time trying to articulate the behaviors they wanted to assess, and align theories of learning with the game. One idea that was initially discussed related to online communities (e.g., forums) interested in pet care, where people share information about different types of pets. For example, one community is exclusively for dog owners. Novice and expert dog owners could use the site to collect and/or share information. This theme could also be useful for social aspects of learning (i.e., such as peer groups and reputation), which are critical factors for interest-driven, informal learning (Ito et al., 2010).

The group ended up prototyping a game called *Bad Dog*. In the game, a player owns a puppy that has behavioral problems. The goal is to gather specialized information from various sources such as neighbors, the local veterinarian, and fellow dog owners at a dog park. The underlying assumption for this game is that knowledge is socially mediated and distributed, and a person makes choices in terms of seeking, filtering, and sharing information. For example, players must choose who to listen to for advice regarding dog behavior. If a character in the game has an incessantly-barking dog and wants to offer advice on how to stop your dog from barking, then you may not want to take that character's advice. So the player must decide *who* to go to for advice based on the information given in the game. Once players reach higher levels of the game, they will need to share their specialized information with others in the social network.

The assessment in *Bad Dog* determines the extent that players correctly choose from whom they will (and will not) take advice. In addition, as players progress through the game, they will gain expertise and eventually choose to share their own "dog management" knowledge with others. However, given time constraints, the group wasn't able to design an appropriate assessment for the sharing-knowledge-with-others choice. http://learningparticipation.net/gla2011/group-videos/k-choicelets-BadDog_480x360.html

Literacy. This group created a game called *Alien Time Patrol*. In this game, a player is a time traveling detective who investigates crimes throughout history. Players begin with a case (e.g., the great Chicago fire) and a list of suspects. Players must read texts associated with each suspect and build an argument as to which suspect is guilty by determining where each suspect was at the time of the fire. Players must also cite particular sentences from the text as evidence to support conclusions in the case. As the game proceeds, the cases and text get more complex and require the player to cite multiple sentences to support conclusions. In really difficult cases, players must cite complex information (e.g., motives) rather than just listing the locations of suspects at the time of the crime. The assessment in this game is the extent that players provide correct evidence

to support who committed the crime in each case. http://learningparticipation.net/gla2011/group-videos/a-literacy_480x360.html

Pre-algebra. This group created a game called *Potion Maker*. In this game, the player is a trader who buys and sells goods using ratios and proportional reasoning to solve various problems. For example, the player may find a person whose camel is sick and needs certain ingredients to cure the ill camel. However, the person who owns the camel only has three blueberries. The player would then go around town looking to trade the blueberries to obtain the ingredients needed to cure the camel. To trade the blueberries, the player must use either ratios or proportional reasoning to obtain the ingredients from various vendors. Each vendor had its own "language" (i.e., speaks in decimals, fractions, or ratios). The player has to learn how to communicate with each of the vendors in order to buy and sell goods. Once all the ingredients are obtained, the player combines them to make the potion. The game can get more complicated by making potions more sophisticated (e.g., the person only needs half a recipe). Math reasoning is assessed in the game by measuring the number of successful potions created over a set of problems. In addition to potion making, players need to solve an explicit transfer task in the game related to math reasoning. The group proposed a research study to evaluate if including an explicit assessment vs. an embedded assessment resulted in more or less math reasoning skill. http://learningparticipation.net/gla2011/group-videos/b-pre-algebra_480x360.html

Creativity. This group created a game called *Outside the Box* that required players to take a simple line drawing and create a new line drawing using only the lines and shapes from the original drawing. For example, a player might start with a picture of a stick man. The player's task is to create a new line drawing using only the arms, legs, torso and head from the stick man. New drawings might include a house with one window, a deer, a fishing rod, or a unicycle. Creativity is measured by the novelty of the new picture created. Novelty can be defined as the inverse of the frequency of players who drew the same object. So if a player drew a picture that no one else drew, it would be considered novel and the estimate for being creative would be increased. http://learningparticipation.net/gla2011/group-videos/icreativity4-Outside-the-Box_480x360.html

Civics. This group prototyped a game called *Class/Action*, which was an online game that allowed teenagers to earn achievements for doing activities that make the political and natural environments better. Some activities are fun and easy, like wearing a t-shirt to school that advertises a favorite cause. Some are kind, like mentoring a fellow student, and some things are edgy, like showing public posters that illustrate the effects of cigarettes on the lungs. Each action earns the player a certain number of *achievement points*. Low points are given for simple actions like complaining about peeling paint in a school, while more points are given for holding a rally to protest poor lunch quality in the school. As a player earns points, they get sorted into a specific class based on the types of achievements they have accomplished. These classes are: organizer, maker, enforcer, evangelist, prankster, artist, arbitrator, mentor, and follower. The high-level achievements require players to find other classes and work together with them. Many

achievements for different classes are designed to support each other. At the highest level, a player gets an additional specialization and becomes a "dual class." For example, a Level 3 Social Engineer is part prankster, part organizer, and a little bit artist. Civics is assessed in *Class/Action* by calculating the number of achievement points accrued across a set of relevant actions, per class. http://learningparticipation.net/gla2011/group-videos/f-civics-Class-Action_480x360.html

Digital Literacy. This group developed an online game called *Celebrity Dogwalker*, where the player assumes the role of an assistant to a budding star. The demands of the celebrity are fulfilled by doing tasks on the Internet. The player is required to navigate, manage, and create online resources to support the budding star's ascension to fame. Thus digital literacy is assessed to the extent the player can use the Internet successfully through doing research, shopping, and engaging in social networking. For example, the player might have to post a video on YouTube, purchase something on eBay, set up a Facebook page, or tweet about the celebrity on Twitter. Doing each one of these tasks successfully will enable the player to gain rank and receive positive feedback. Digital Literacy (the core competency) is assessed to the extent that players successfully accomplish their assigned digital tasks and follow the guidelines of appropriate behavior for each digital behavior. For instance, the player must know (a) the rules of bidding on eBay, (b) the rules about posting copyrighted material on Facebook, and (c) the social consequences of slandering someone on Twitter. http://learningparticipation.net/gla2011/group-videos/j-digital-lit2_480x360.html

SUMMARY/SYNTHESIS

The group sessions accomplished a variety of clever game-based assessment designs. Moreover, several of these ideas have been pursued further among members of certain teams. To more clearly understand how the attendees felt about the design sessions, we conducted a short survey about their experiences and perceptions relative to their assigned team. Overall, the attendees thought their groups were successful at making game-based assessments. Here are two quotes on how attendees felt about designing game-based assessments:

> *It was a terrific working team. Interesting to watch the game designer process – they just kept rolling forward with ideas, would hit a bump and just move around it but not slow down. Educators tended to step and go deeper, be more analytical. Together it made quite a strong team for analyzing and solving problems. I thought the process implications went far beyond gaming.*

> *Achievements are a great way to build assessment into a game because they allow us to instrument behavior at a finely-grained level and cluster that information into a more coarsely grained description of high-level activity. Our game not only provides feedback to the student – it helps to align them to other students with complementary skills and also provides a means to analyze not just what sorts of activities students are interested/involved in, but what "types" of students like what types of activities.*

However, there were some lessons to be learned from the breakout sessions. Below are some quotes summarizing issues with regard to the design sessions.

> *The tension of "fun" vs. "learning" was a constant thread. How do you increase the fun level while holding rigorous academic standards?*
>
> *Once the core ideas were in place, more emphasis was placed on the gameplay and less emphasis tended to be placed on fleshing out the assessment ideas and models.*
>
> *It took awhile for people to wrap their head around the assessment goal and the constraints of game design. We were trying very hard to assess learning processes rather than slapping a posttest on a game or simply measuring easy procedural skills that we already know how to measure. I think the thing we were lacking most as a group was a good sense of the very high level of specificity and decomposition needed for both assessment and game design.*

Finally, workshop attendees responded to the question about obstacles related to the design of game-like assessments as follows.

> *Strong preconception that if you design an engaging game and/or high fidelity environment and action, it follows that you will have a good assessment if you can just figure out 'how to score it' (i.e., throw the data over the wall to the psychometrician). What you need to do is include, from the very first iterative design cycles, the constraint that you will need to support a coherent and credible assessment argument as well as the constraints of engagement and fidelity needed in a successful game.*
>
> *Getting clear about the assessment piece – I think we put the game piece first in our group, and we didn't have enough clarity about what we were trying to assess.*

Many of the suggestions called for a research methodology in game-based assessment research. The next section describes research recommendations for moving forward with game-based assessments.

Developing a Research Methodology

Need for a Systematic Approach to Designing Game-Based Assessment

The primary purpose of an assessment is to collect information that will enable the assessor to make inferences about a person's competencies – what she knows, believes, can do, and to what degree. Sound research methods are needed to explore how games can be used as reliable and valid assessments of multiple and diverse competencies. From an assessment standpoint, research is needed to determine the value added of game-based assessment over traditional assessment relative to the quality of the assessment (i.e., validity, reliability, and generalizability) and the impact on learning, engagement, and transfer. This research should also include analyses on how

game-based assessment data predict important external criteria (e.g., high school or college graduation, state test scores, as well as happiness, civic engagement, etc.) better than traditional assessments.

We begin by describing one research approach that is highly suitable for building valid and reliable assessments in games that can be used to support learning. It is called Evidence Centered Design (ECD).

Evidence Centered Design

ECD defines a framework that consists of three main models that work in concert (see Mislevy & Haertel, 2006; Mislevy, Almond, & Lukas, 2004; Mislevy, Steinberg, & Almond, 2003). The ECD framework allows an assessor to: (a) define the claims to be made about learners' competencies, (b) establish what constitutes valid evidence of the claim, and (c) determine the nature and form of problems that will elicit that evidence. A good assessment (which could be a game) elicits behavior that bears evidence about key competencies, and it must also provide principled interpretations of that evidence in terms that suit the purpose of the assessment.

Although ECD can work with simple one-dimensional competency models, its strength comes from treating competency as essentially multidimensional. Variables in the competency model (CM) describe the set of variables on which inferences are based (see Almond & Mislevy, 1999). The term student model is used to denote an instantiated version of the CM – like a profile or report card, only at a more refined grain size. Values in the student model express the assessor's current belief about the level on each variable within the CM for a given student.

An evidence model expresses how the student's interactions with, and responses to a given problem constitute evidence about competency model variables. The evidence model (EM) attempts to answer two questions: (a) What behaviors or performances reveal targeted competencies; and (b) What's the connection between those behaviors and the CM variable(s)? In particular, when faced with a multidimensional CM, the EM identifies which observable outcomes are associated with which competencies (and the strength of the association). Basically, an evidence model lays out the argument about why and how observations in a given situation (i.e., student performance data) constitute evidence about CM variables.

Task models describe features of situations that will be used to elicit performance. A task model (TM) provides a framework for characterizing and constructing situations with which a student will interact to provide evidence about targeted aspects of knowledge related to competencies. Task specifications establish what the learner will be asked to do, what kinds of responses are permitted, what types of formats are available, and so on. The main purpose of tasks is to elicit dynamic evidence (which is observable and empirical) about competencies (which are unobservable and theoretical). Results from the task model in a traditional assessment consist of a set of items or problems; but in the context of game-based assessments, task modeling would

produce a collection of integrated and engaging situations or moments designed to capture particular types of performance data that would then inform the targeted learning goals or competencies.

As a brief example of how the ECD framework can be used to build a game-based assessment, consider the game *Undead Apocalypse* (described earlier in the Systems Thinking breakout session). In designing this game, the group first developed a simple CM by clearly defining what the competency would be (i.e., systems thinking, along with additional skills such as understanding interrelationships among variables and feedback loops). A detailed CM would entail determining the complete theory of all the dimensions of the focal competency. The group then worked on identifying specific gameplay behaviors that would provide valid evidence of systems thinking. For instance, some evidence reflecting "understanding interrelationships among variables" may be the degree to which a player makes good decisions during the game – as required in the monster-swapping task. The design and development of the actual game comes last since it is dependent on what behaviors are identified in the EM, which in turn is dependent on the competencies identified in the CM.

In summary, the ECD framework provides a systematic approach to assessment design and also provides a transparent way to reason about student/player performance. And while there are other ways to develop assessments, they often lack transparency as well as specification about the competencies, and tend to focus on tasks that are too simple or inauthentic. This can result in creating assessments that measure unintended competencies and thus damage the reliability and validity of the assessment. In addition, using an ECD approach for the design of game-based assessments can provide a clear research and communication framework for assessment/learning experts and game designers who want to design and develop new educational and engaging game-based assessments that accurately measure important traditional and new competencies. The final section describes how workshop attendees foresee future work in relation to games, learning, and assessment.

FUTURE STEPS

Bringing together educators, assessment experts, and game designers is a critical part of creating an effective learning ecosystem. - GLA workshop attendee

One overwhelming recommendation from workshop attendees was to hold another workshop (or series of workshops) with more time devoted to discussing assessment and game design. Having a shared understanding of *both* areas is key to moving forward with the design of engaging, educational games. Thus the first recommendation to advance game-based assessment research would be funding future workshops. These workshops should be two or more days and include the right balance of educators (e.g., assessment, learning, and content experts) and game designers to optimize the development of game-based assessment. The workshops could be focused on specific competencies (or even just a single focal competency) that link to the

Common Core Standards. The end result of each team project would be a prototype of a game-based assessment to be considered for future funding.

A second recommendation for the funding agencies to consider involves creating an open solicitation for funding multi-year R&D projects on game-based assessment. This solicitation would seek interdisciplinary research teams to develop assessments embedded with digital games. Proposals should state how the game-based assessment is intended to link to important outcomes (e.g., increased high school graduation rates) as well as to the Common Core Standards. Research projects should clearly specify the competencies being addressed, what game will be used (or developed), why the focal competency is not measured properly in today's education system, and why a game-based approach will be beneficial. Funding opportunities could be focused on one particular competency (e.g., pre-algebra knowledge and skills) or more broadly on a set of rich competencies (e.g., creativity, communication, and collaboration). Proposals should clearly state why the particular game will be well suited to elicit evidence of the focal competency.

A fully outlined validity design should be included in each grant proposal as well. This validity design may be focused on predictive validity (i.e., whether the new assessment predicts important outcomes) since this question shows why the competency is needed. Another validity argument may be centered around why a game-based assessment is needed over a traditional method of assessment that measures the same competency. This argument, often referred to as incremental validity, pertains to how well the game-based assessment predicts an important outcome over and above a traditional assessment. If a game-based assessment does not offer any prediction beyond a traditional assessment, then the value added of the game-based assessment falls into question. However, if a game-based assessment fails the question of incremental validity, there still can be value added from a learning perspective. That is, a game-based assessment can be more engaging than traditional measures, or have diagnostic qualities that can be used to support learning of a competency. For example, a game-based assessment may inform the game about a misconception or other problem related to a competency, then the game could adjust gameplay (e.g., reduce difficulty) to more accurately match the competency level of the player. Thus research proposals would need to specify if learning can occur directly within the game itself as a function of the assessment. If so, how does the game-based assessment provide a better learning experience than just the playing the game without the assessment? How does the game-based assessment complement the game and enhance engagement, learning, and/or transfer?

SUMMARY

The GLA workshop can be seen as the beginning of an exciting, new, and much needed game-based assessment research program that supports both Common Core Standards and important new competencies, and incorporates socio-constructivist learning theories, principles of game and instructional design, as well as embedded assessment. Future research will help us to identify specific game-design elements, their interactions with one another, with the learner(s), and with the content and competencies being supported. For instance, does the type and timing of feedback

differentially affect learners or types of learners? What is the best grain size of competencies to monitor in a game to maximize learning? And similarly, how can we optimally match the level of a challenge to a learner's competency level (i.e., estimate a person's ever-changing zone of proximal development)? For many of these research questions, learning and assessment theory may inform design elements that facilitate the students' in-game experiences and enhance learning and assessment quality in games. A game-based assessment model itself may eventually be developed and used to design and analyze a variety of games answering general and important context questions such as: what works, for whom, to what degree, under which conditions, and for what competencies or domains?

In conclusion, well-designed games are potentially powerful vehicles to support assessment and learning in three major ways. First, game-based assessments can be useful to measure new competencies not currently embraced or supported by our education system. Second, game-based assessments can provide authentic performance assessments situated within realistic and/or meaningful problem solving tasks. This is important because it addresses the problem of face validity in today's assessments (i.e., tests not appearing to be relevant to a competency). Finally, game-based assessments can be useful for learning since they can be woven into the curriculum and also linked to Common Core Standards. This new stream of game-based assessment research addresses a critical problem in education related to the dire state of sequestered, context-devoid assessments which provide limited information about what kids have learned or are able to do. There is much work to be done, and future workshops and research will clarify the challenges ahead and the feasibility of a new game-based assessment research program.

REFERENCES

Almond, R. G., & Mislevy, R. J. (1999). Graphical models and computerized adaptive testing. *Applied Psychological Measurement, 23*(3), 223-237.

Barab, S. A., Zuiker, S., Warren, S., Hickey, D., Ingram-Goble, A., Kwon, E-J., Kouper, I., & Herring, S. C. (2007). Situationally embodied curriculum: Relating formalisms and contexts. *Science Education, 91*(5), 750-782.

Bereiter, C, & Scardamalia, M. (1993). *Surpassing ourselves: An inquiry into the nature and implications of expertise*. Chicago, IL: Open Court.

Bridgeland, J. M., DiIulio, J. J., Jr., & Morison, K. B. (2006). *The silent epidemic: Perspectives of high school dropouts*. Washington, DC: Civic Enterprises Civic Enterprises and Peter D. Hart Research Associates. For the Bill and Melinda Gates Foundation.

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher, 18*(1), 32-41.

Cooper, Khatib, Treuille, Barbero, Lee, Beenen, Leaver-Fay, Baker, Popovic, and Foldit Players. (2010, August). Predicting protein structures with a multiplayer online game. *Nature, 466*(7307), 756–760.

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology, 92*(6), 1087-1101.

Duffy, T. M. & Jonassen, D. (Eds.), (1992). *Constructivism and the technology of instruction: A conversation*. Hillsdale NJ: Lawrence Erlbaum Associates.

Fullerton, T. (2008). *Game design workshop: A playcentric approach to creating innovative games,* 2nd Edition. San Francisco, CA: Morgan Kaufmann.

Gee, J. P. (2003). *What video games have to teach us about learning and literacy*. New York: Palgrave Macmillan.

Gee, J. P. (2008). Learning and games. In K. Salen (Ed.), *The ecology of games: connecting youth, games, and learning* (pp. 21-40). The John D. and Catherine T. MacArthur Foundation Series on Digital Media and Learning. Cambridge, MA: The MIT Press.

Gee, J. P., Hull, G., & Lankshear, C. (1996). *The new work order: Behind the language of the new capitalism*. Boulder, CO: Westview.

Gonzales, P., Williams, T., Jocelyn, L., Roey, S., Kastberg, D., and Brenwald, S. (2008). *Highlights from TIMSS 2007: mathematics and science achievement of U.S. fourth- and eighth-grade students in an international context* (NCES 2009-001). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Howard L. F., Paul J. H., Marisa P. P., & Brooke, E. S. (2010). *Highlights from PISA 2009: Performance of U.S. 15-year-old students in reading, mathematics, and science literacy in an international contex*t (NCES 2011-004). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Ito, M., Baumer, S., Bittanti, M., boyd, d., Cody, R., Herr-Stephenson, B., et al. (2010). *Hanging out, messing around and geeking out: kids living and learning with new media*. The John D. and Catherine T. MacArthur Foundation Series on Digital Media and Learning. Cambridge, MA: The MIT Press.

Korbin, J. L., Patterson, B. F., Shaw, E. J., Mattern, K. D., & Barbuti, S. M. (2008). *Validity of the SAT for predicting first-year college grade point average* (Research Report No. 2008-5). New York: College Board.

Lenhart, A., Kahne, J., Middaugh, E., MacGill, A., Evans, C., & Vitak, J. (2008). *Teens, video games, and civics*. Washington, DC: Pew Internet & American Life Project.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge: Cambridge University Press.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice, 25*(4), 6-20.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2004). *A Brief Introduction to Evidence-Centered Design* (CSE Report 632). CA: Center for Research on Evaluation, Standards, and Student Testing. (ED483399)

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

No Child Left Behind Act (NCLB) (2002). *Title 1: Improving the academic achievement of the disadvantaged. Summary of Final Regulations*. Retrieved from http://www.ed.gov/programs/titleiparta/index.html#reg

Noftle, E. E., & Robins, R. W. (2007). Personality predictors of academic outcomes: Big Five correlates of GPA and SAT scores. *Journal of Personality and Social Psychology, 93*, 116-130.

O'Connor, M., & Paunonen, S. V. (2007). Big Five personality predictors of post-secondary academic performance. *Personality and Individual Differences, 43*, 971-990.

Partnership for 21st Century Skills (2006). *Results that matter: 21st century skills and high school reform*. Retrieved from http://www.p21.org/documents/RTM2006.pdf

Piaget, J. (1950). *The psychology of intelligence*. New York: Routledge.

Prensky, M. (2006). *Don't bother me mom, I'm learning!: How computer and video games are preparing your kids for 21st Century success and how you can help!* St. Paul, MN: Paragon House.

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin, 135*, 322-338.

Salen, K., & Zimmerman, E. (2005). Game design and meaningful play. In J. Raessens & J. Goldstein (Eds.), *Handbook of computer game studies* (pp. 59-80). Cambridge, MA: MIT Press.

Rideout, V., Foehr, U. G., & Roberts, D. F. (2010). *Generation M^2: media in the lives of 8-18 year-olds*. Washington DC: Kaiser Family Foundation.

Shaffer, D. W., Squire, K. A., Halverson, R., & Gee, J. P. (2005).Video games and the future of learning. *Phi Delta Kappan, 87*(2), 104-111.

Shute, V. J. (2007). Tensions, trends, tools, and technologies: Time for an educational sea change. In C. A. Dwyer (Ed.), *The future of assessment: shaping teaching and learning* (pp. 139-187). New York, NY: Lawrence Erlbaum Associates, Taylor & Francis Group.

Shute, V. J. (2009). Simply assessment. *International Journal of Learning, and Media, 1*(2), 1-11.

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. In S. Tobias, & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 503-524). Charlotte, NC: Information Age Publishers.

Shute, V. J. & Torres, R. (in press). Where streams converge: Using evidence-centered design to assess Quest to Learn. To appear in D. Robinson, J. Clarke-Midura, & M. Mayrath (Eds.), *Technology-based assessments for 21st Century skills: Theoretical and practical implications from modern research.* New York, NY: Springer-Verlag.

Shute, V. J., Rieber, L., & Van Eck, R. (2011). Games . . . and . . . Learning. To appear in R. Reiser & J. Dempsey (Eds.), *Trends and issues in instructional design and technology* (3rd ed.) (pp. 321-332). Upper Saddle River, NJ: Pearson Education, Inc.

Shute, V. J., Ventura, M., Bauer, M. I., & 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. Cody, & P. Vorderer (Eds.), *Serious games: Mechanisms and effects* (pp. 295-321). Mahwah, NJ: Routledge, Taylor and Francis.

Sternberg, R. J. (2006). Creating a vision of creativity: The first 25 years. *Psychology of Aesthetics, Creativity, and the Arts, S* (1), 2–12.

Trapmann, S., Hell, B., Hirn, J.W., & Schuler, H. (2007). Meta-analysis of the relationship between the big five and academic success at university, *Journal of Psychology, 215*, 132–151.

Vygotsky, L.S. (1978). *Mind in society: The development of higher mental processes.* Cambridge, MA: Harvard University Press.

Walberg, H.J. & Stariha, W.E. (1992). Productive Human Capital: Learning, Creativity and Eminence. *Creativity Research Journal, 5*, 323-340.

Woolf, B. P., Shute, V. J., VanLehn, K., Burleson, W., King, J., Suthers, D., Bredeweg, B., Luckin, R., & Tonkin, E. (2010). *A roadmap for education technology*. Computing Community Consortium, Washington, DC.