CHAPTER 13 –Modeling Student Competencies in Video Games Using Stealth Assessment

Valerie Shute[1], Matthew Ventura[1], Matthew Small[1], and Benjamin Goldberg[2]
[1]Florida State University, [2]U.S. Army Research Laboratory

Introduction

We have been examining ways to leverage video games to assess and support important student competencies, especially those that are not optimally measured by traditional assessment formats. The term "stealth assessment" refers to the process of embedding assessments directly and invisibly into the learning or gaming environment. Though this approach produces ample real-time data on a player's interactions within the game environment and preserves player engagement, a primary challenge for using stealth assessment in games is taking this stream of data and making inferences about players' competencies that can be examined at various points in time (to see growth) and also at various grain sizes (for diagnostic purposes). In this chapter, we present recent work related to creating and embedding various stealth assessments into *Newton's Playground*, a computer game that emphasizes nonlinear gameplay and puzzle-solving in a two-dimensional (2-D) physics simulation environment. We conclude with a discussion on stealth assessment within GIFT, highlighting research recommendations for enhancing the architecture to support robust methods of evidence-centered assessment.

In this chapter, we examine the possibility of using well-designed games as vehicles to assess and support learning. There are several factors motivating this research. First, our schools have remained virtually unchanged for many decades while our world is changing rapidly. This lack of reform in our schools could be a contributing factor in high dropout rates, especially among Hispanic, Black, and Native American students, which were described as "The Silent Epidemic" in a recent research report for the Bill and Melinda Gates Foundation (Bridgeland, DiIulio & Morison, 2006). According to this report, nearly one third of all public high school students drop out and the rates are higher across minority students. Importantly, when 467 high school dropouts were asked why they left school, 47% of them simply responded, "*The classes were not interesting*." In light of this finding, we need to identify ways (e.g., video games) to get young people engaged in learning the skills needed to succeed in today's competitive economy.

A second reason for using games as assessments is a pressing need for dynamic and ongoing measures of learning processes and outcomes. Interest in alternative forms of assessment is driven by dissatisfaction with and limitations of multiple-choice items. In the 1990s, interest in alternative forms of assessment increased with the popularization of what became known as authentic assessment. Authentic assessment refers to tasks that resemble academic and real-world activities (e,g., Hiebert, Valencia & Afflerbach, 1994). A number of researchers found that multiple-choice and other fixed-response formats substantially narrowed school curricula by emphasizing basic content knowledge and skills within subjects and not assessing higher-order thinking skills (e.g., Kellaghan & Madaus, 1991; Shepard, 1991). However, as Madaus and O'Dwyer (1999) argued, incorporating performance assessments into testing programs is difficult because they are less efficient, more difficult and disruptive to administer, and more time-consuming than multiple-choice testing programs. Consequently, multiple-choice has remained the dominant format in most K–12 assessments in our country. New performance assessments are needed that are valid and reliable, and can be scored automatically.

A third reason for using games as assessment vehicles is that many games typically require a player to apply various competencies (e.g., creativity, critical thinking, problem solving, persistence, and collaboration) to succeed in the game. The competencies required to succeed in many games also happen

to be the same ones that companies are looking for in today's highly competitive economy (Arum & Roska, 2011; Gee, Hull & Lankshear, 1996). Moreover, games are a significant and ubiquitous part of young people's lives. For instance, 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).

In addition to the arguments for using games as assessment devices, there is growing evidence of games supporting learning (e.g., Tobias & Fletcher, 2011; Wilson et al., 2009). However, we need to understand more precisely how and what kinds of knowledge and skills are being acquired. Understanding the relationships between games and learning is complicated by the fact that we don't want to disrupt players' engagement levels during game play. Consequently, learning in games has historically been assessed indirectly and/or in a post hoc manner (Shute & Ke, 2012; Tobias, Fletcher, Dai & Wind, 2011). What's needed instead is real-time assessment and support of learning based on the dynamic needs of players. We need to be able to experimentally ascertain the degree to which games can support learning, and how they achieve this objective.

A challenge with developing a performance-based measure is crafting appropriate situations or problems to elicit a competency of interest. One way to approach this problem is to use video games to simulate problems for performance-based assessment (Dede, 2005; DiCerbo & Behrens, 2012; Quellmalz, Timms, Buckley, Silberglitt & Brenner, 2012). Digital learning environments can provide meaningful assessment environments by providing students with scenarios that require the application of various competencies. In this chapter, we introduce research that explores a variant of this assessment approach by investigating how performance-based assessments can be used in a video game we created called *Newton's Playground.*

Stealth Assessment

Given the goal of using well-designed games to support learning in school settings and elsewhere, we need to ensure that the assessments are valid, reliable, and also unobtrusive (to keep engagement intact). The output from the assessments, however, should be transparent. That is, players should be aware of how they are doing relative to important competencies at any point in time to motivate learning. One way to meet these requirements is to use "stealth assessment" (Shute, 2011; Shute & Ventura, 2013). Stealth assessment refers to Evidence-Centered Design (ECD)-based assessments that are woven directly and invisibly into the fabric of the gaming environment. During game play, students naturally produce rich sequences of actions while performing complex tasks, drawing on the very skills or competencies that we want to assess (e.g., scientific inquiry skills, creativity). Evidence needed to assess the skills is thus provided by the players' interactions with the game itself (i.e., the processes of play), which can be contrasted with a summative score – the norm in educational environments.

Making use of this stream of gameplay evidence to assess students' knowledge, skills, and understanding (as well as beliefs, feelings, and other states and traits) presents problems for traditional measurement models used in assessment. First, in traditional tests, the answer to each question is seen as an independent data point. In contrast, the individual actions within a sequence of events in a game are often highly dependent on one another. For example, what one does in a particular game at one point in time affects subsequent actions later on. Second, in traditional tests, questions are often designed to measure particular, individual pieces of knowledge or skills. Answering the question correctly is evidence that one may know a certain fact: one question – one fact. But by analyzing a sequence of actions within gameplay (where each response or action provides incremental evidence about the current mastery of a specific fact, concept, or skill), stealth assessments can infer what learners know and do not know at any point in time. Now, because we typically want to assess a whole cluster of skills and abilities using evidence coming

from learners' interactions within a game, methods for analyzing the sequence of behaviors to infer these abilities are not as obvious. As suggested above, evidence-based stealth assessments can help address these problems. The next section reviews the game we created called *Newton's Playground* and the associated development of stealth assessments for monitoring learner knowledge and progression.

Stealth Assessment in *Newton's Playground*

Research into what's called "folk" physics demonstrates that many people hold erroneous views about basic physical principles that govern the motion of objects in the world, a world in which people act and behave quite successfully (Reiner, Proffit & Salthouse, 2005). For example, when asked to draw the water level on a picture of a tilted drinking glass, about 40% of young adults draw lines that are not horizontal (McAfee & Proffitt, 1991). When asked to predict the path that a pendulum takes when the string is cut at various points, a large percentage of people make systematically incorrect judgments (Caramazza, McCloskey & Green, 1981). The prevalence of these systematic errors has led some investigators to propose that incorrect performance on these tasks is due to specific "naive" beliefs, rather than to a general inability to reason about mechanical systems (McCloskey & Kohl, 1983). Recognition of the problem has led to interest in the mechanisms by which physics students make the transition from folk physics to more formal physics understanding (diSessa, 1982) and the possibility of using video games to assist in the learning process (Masson, Bub & Lalonde, 2011; White, 1994).

One way to help remove misconceptions in physics is to illustrate physics principles with physical machines (Hewitt, 2009). In physics, a machine refers to a device that is designed to either change the magnitude or direction of a force. Teaching about simple machines (e.g., lever, pulley, and wedge) is widely used as a method to introduce physics concepts (Hewitt, 2009). Recent research on science education also indicates that learners' hands-on experience with such machines (both virtually and physically) support applicable understanding of important physics concepts (Hake, 1998).

We developed a video game called *Newton's Playground* (NP) to help middle school students experience and understand what we call informal physics. We define informal physics as a nonverbal understanding of how the physical world operates. Informal physics is characterized by an implicit understanding of Newton's three laws, balance, mass, conservation of momentum, kinetic energy, and gravity. NP is a 2-D game that requires the player to guide a green ball to a red balloon. The player can nudge the ball to the left and right (if the surface is flat) but the primary way to move the ball is by drawing/creating simple machines (which are called "agents of force and motion" in the game) on the screen that "come to life" once the object is drawn. Everything obeys the basic rules of physics relating to gravity and Newton's three laws of motion. The 74 problems in NP require the player to draw/create four agents: inclined plane/ramps, pendulums, levers, and springboards. All solutions are drawn with colored lines using the mouse.

A ramp is any line drawn that helps to guide a ball in motion. A ramp is useful when a ball must traverse over a gap or obstacle. A lever rotates around a fixed point, that is, a fulcrum or pivot point. Levers are useful when a player wants to move the ball vertically. A swinging pendulum directs an impulse tangent to its direction of motion. The pendulum is useful when the player wants to exert a horizontal force. A springboard (or diving board) stores elastic potential energy provided by a falling weight. Springboards are useful when the player wants to move the ball vertically. For example, in the "golf problem" (see Figure 13-1), the player must draw a golf club on a pin (i.e., little circle on the cloud) to make it swing down to hit the ball. In the depicted solution, the player also drew a ramp to prevent the ball from falling down a pit.

The speed of (and importantly, the impulse delivered by) the swinging golf club is dependent on the size/mass distribution of the club and the angle from which it was released. The ball will then move at a certain speed, length, and trajectory. If drawn properly, the ball will hit the balloon.

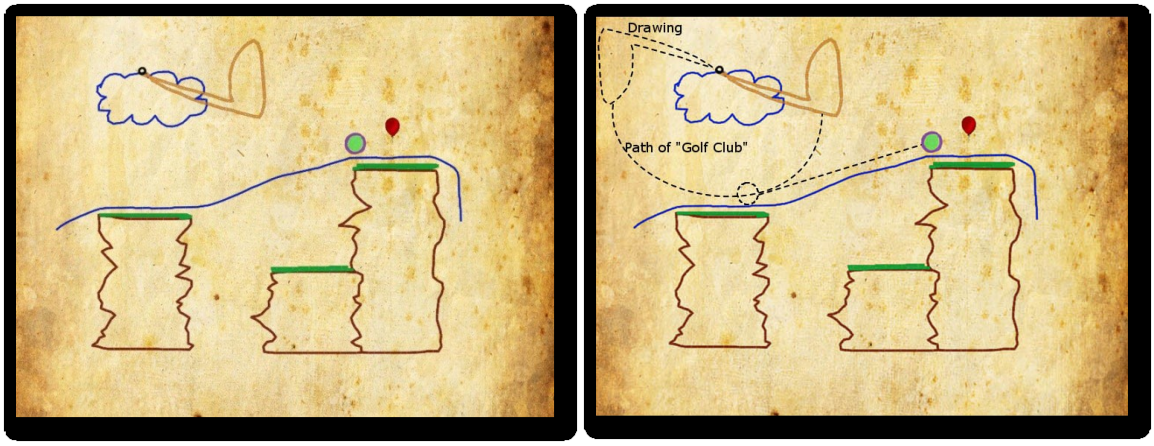

Figure 13-1: Golf problem in NP (left is solution; right is path of motion)

All solutions are drawn with colored markers using the mouse. In a number of cases, the ball must go over a pit. If the ball falls into the pit, the player must start the problem over. Players can replay a problem as often as they like – even after successfully solving it. One motivation to replay a problem is to find even more elegant and creative solutions than were generated before. It is not uncommon for a player to revisit/replay particularly challenging problems multiple times, striving for a better, more elegant solution.

Our system for agent identification (i.e., detecting the creation and use of simple machines) ignores visual data and instead uses information from the underlying physics simulation to classify agents of force and motion. We identify a primary object (PO) for each agent that provides the most salient features for the identification of that agent. That is, the ramp and pendulum agents each use only one object, thus it is the PO. For a springboard, the PO is the one that "springs" up to propel the ball. Finally, the PO in the lever agent is the object that rotates under the load of another object, to lift the ball. Our method continuously monitors all objects in the game for telltale characteristics of the PO in each particular agent of force and motion.

We divide the agent identification process into three stages: *default*, *monitor*, and *identify*. All objects start in the *default* phase for each agent. When an object exhibits characteristics of the PO for any particular agent, it is elevated to the *monitor* stage. Once an object is in the *monitor* stage for a particular agent, detailed data about its movement and interactions with the game world are recorded and the object will inevitably move on to the *identify* stage, where the gathered data are analyzed and a decision is made about whether it is indeed the primary object in a current manifestation of the corresponding agent of motion.

To illustrate our agent identification system, we now describe the process of identifying the pendulum agent of motion (which is the agent used in Figure 13-1 to solve the golf problem). A drawn object begins in the default stage for the pendulum agent. When the object meets the following criteria it is elevated to the monitor stage: (1) the object is attached to a single pin, and (2) the object has rotated more than 20°. The monitor stage will gather physics data for ¾ second and then the agent identification stage will be

triggered. A positive identification is made if, during the monitor phase, the object made contact with the ball (i.e., pendulum strike) and the ball moved more than a preset distance. Regardless of whether the identification is made, the object is then lowered back down to the default stage. The classification of other agents occurs in a similar manner.

Task Modeling for Informal Physics

All NP problems require the player to use one or more agents of force and motion in the solution. Successful solutions thus inform one or more of the competencies that we hope to develop in the student. As an illustration, consider the problem called *ballistic pendulum*, shown in Figure 13-2.

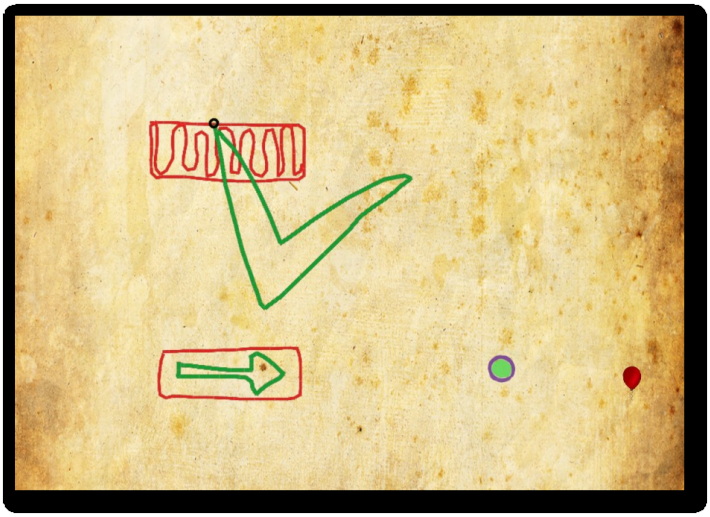

Figure 13-2. **Ballistic pendulum problem**

This problem requires the student to create a pendulum shape with sufficient mass and positioning so that the pendulum will fall down and "kick" the ball into a free-fall trajectory that ends up landing on the red balloon (the figure shows the ball en route to the balloon). Successfully solving this problem suggests that the student has an intuitive grasp of the concepts of torque, linear, and angular momentum since the correct application of each is required to get the ball to the balloon. Incidentally, the ballistic pendulum is also an experiment often done in introductory physics courses in high school or college.

Other Gameplay Features

NP consists of 7 playgrounds (each one containing around 10–11 levels) that progressively get more difficult. The difficulty of a problem is dependent on a number of factors including: relative location of ball to balloon, obstacles, number of agents required to solve the problem, and novelty of the problem. NP also has introductory videos that show how to use the various agents of force and motion. These tutorials illustrate how to draw each agent to solve a simple problem (during gameplay, students have the option to watch any agent-drawing video at any time).

Object Limit

In pilot testing, we discovered that players sometimes opt to draw lines under the ball in order to move it (called line stacking). Repeated use of line stacking can lead players to not learn the agents of force and motion. In order to preclude line stacking to obtain a solution we implemented an "object limit" of 10 objects per level. Once a player draws 10 lines no more lines can be drawn until a line is deleted or the problem is reset (by hitting the space bar). Students can see their line limit in the bottom right hand side of the screen.

Trophies

NP also displays silver and gold trophies in the top left part of the screen, which represent progress in the game. A silver trophy is obtained for any solution to a problem. Players receive a gold trophy if their solution is under a certain number of objects (the threshold varies by problem, but is usually less than three objects). So a player can receive a silver and gold solution for each problem. It is not uncommon for a player to revisit/replay particularly challenging problems multiple times to receive a gold solution.

Log Files

The heart of the stealth assessment lies in the log files generated by NP. NP automatically uploads log files to a server for a session (i.e., log activity between login and log out). Figure 13-3 displays what the log file looks like for one problem.


```
"time_stamp": 12.163,
"level_path": ".\\levels\\p4\\diving board.level",
"game_time": 130.526001,
"pause_time": 1.54,
"restart_count": 2,
"object_count": 14,
"object_limit_count": 1,
"nudge_count": 42,
"erase_count": 13,
"pin_count": 1,
"agent_vector":"61.78 SB, 98.08 SB, 131.60 SB"…
"ball_trajectory": "<0.733, 0.427> <0.766, 0.394>…
"silver": true,
"gold": false,
"solved": true
```


Figure 13-3. Example log file

The session log displays counts and times for several features of gameplay relevant to physics. For example, the "object limit count" reports the number of time a player exceeds the object limit, which can be seen as a lack of knowledge of a particular agent of motion (depending on the problem). Also the "agent vector" reports the agents used in the problem along with the time stamp it was executed (e.g., at timestamp 61.78, a springboard [SB] was created). Finally, the "ball trajectory" reports the 2-D coordinates of the ball over the last few seconds of a solved problem (i.e., the "solution path" of the ball).

A second log file NP reports is called a "replay file." The replay file records all player interactions with the game while attempting to solve a problem. Such interactions include drawing and erasing game objects, creating pins and nudging the ball. NP can read this file to render a visual replay of a problem

attempt in real time. The replay system was integral in tuning and verifying the accuracy of the automatic agent identification system.

Preliminary Results

We recently conducted a study where we had middle school students (n = 165) play NP for around 5 hours (split into six 45-min sessions). Working with a physics professor, we developed a physics test that assesses informal physics knowledge and does not require math for solutions to physics problems. For example, Figure 13-4 shows an item involving a pendulum. The correct answer is "B." We administered the informal physics pretest at the beginning and a post-test at the end of the gameplay sessions.

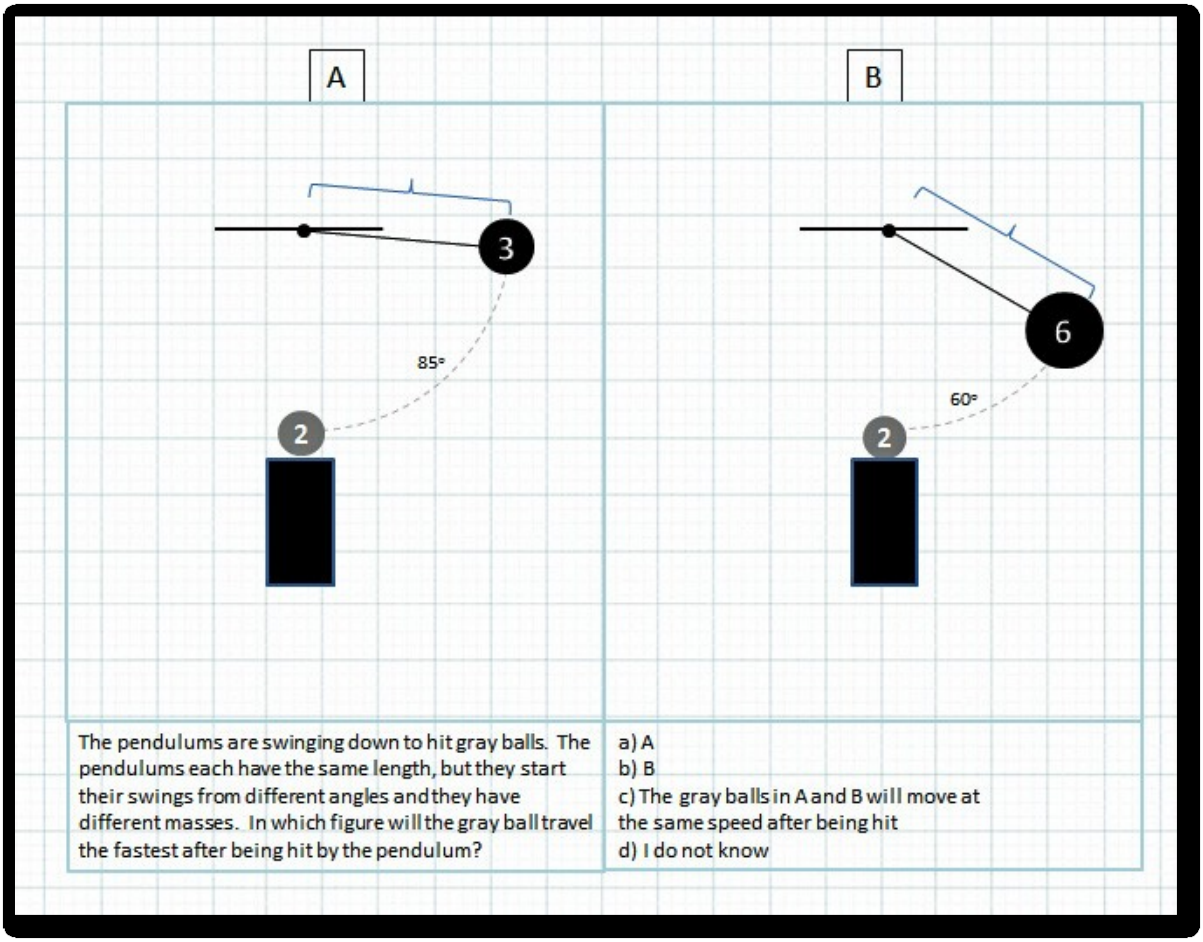

Figure 13-4. Example item from informal physics test

So far, we have found a significant difference between the pre-test and post-test scores (t (154) = 2.12, p < 0.05). Students playing the game improved in their informal, conceptual physics understanding over time. Current analyses are revealing that our stealth assessment indicators correlate to a player's knowledge of physics concepts as measured by the tests.

We now turn our attention to how stealth assessment may be employed within the GIFT architecture.

Stealth Assessment within GIFT

With domain-independency being a major requirement in the development of GIFT, it is important that the architecture supports varying open and dynamic game-based learning environments that apply distinctively different messaging protocols. This involves embedding components and processes within GIFT's domain module to support the authoring of stealth assessments regardless of the game-engine being used. Rules and models built around game interaction must be explicitly linked to concepts defined inside the GIFT architecture. This enables the assessment of concepts within GIFT's domain schema as it relates to evidence captured in a game. Based around this notion, GIFT must support the linkage of activities in a game with defined objectives that denote competent behavior within a given domain, regardless of the data structure being extracted from the game environment. For this purpose, a "Gateway Module" is incorporated that associates an external educational/training system's state data with a domain or competency model built within GIFT. This linkage allows for two disparate systems to communicate with one another. In the case of GIFT, this enables the application of AI tools and methods that facilitate real-time assessment of player actions as they relate to desired learning and performance processes and outcomes.

The challenge is that a majority of gaming platforms apply different messaging protocols. Developing approaches to rapidly pair GIFT with any platform is recommended, which will increase its utility across multiple learning environments. This allows any system to link interaction with GIFT's domain model, where assessments are conducted and progress is communicated to the learner model for determining transitions in performance or competency. This approach to assessment is ideal in game-based environments as tracking interaction data as they relate to objectives can denote comprehension and understanding that is difficult to gauge in traditional assessment techniques. The application of stealth assessment within GIFT potentially provides further diagnosis of game performance, which can be communicated to the pedagogical model for more focused selection of feedback and remediation tactics.

As discussed earlier, stealth assessment is dependent on data streams that can be pulled out of the system. In the example of NP, player interaction is monitored for the purpose of capturing a player's creation and use of the relevant agents and inferring how their actions relate to mastery of informal physics knowledge within a specified problem space. Here, user interaction is monitored to determine the application of agents for solving a problem.

Within the context of game-based systems, the domain model in GIFT must accommodate the inclusion of stealth assessment techniques, such as those implemented in NP, that distinguish competency from interaction within a dynamic environment where learners have free control of their movements. These relationships are currently authored in GIFT's Domain Knowledge File (DKF), where a structured XML schema is used to associate specific domain content with generalized tags that can be communicated to the learner model (i.e., concepts, objects, and assessment logic). In the initial releases of GIFT, each of these components are hand-coded by programmers as they link available gameplay data with defined concepts as they relate to the objectives being instructed. This is not ideal, as the goal of GIFT is to enable instructors and trainers to author intelligent tutoring capabilities without possessing skills across the multiple disciplines (e.g., computer science, instructional design, psychometrics, cognitive psychology) required to build such a system.

To ease this burden, research is necessary to identify and develop tools that intuitively guide a course developer through the authoring process. Ultimately, the overarching goal is for this process to involve minimal to no programming by using natural language and well-developed user interfaces to express policies that can be converted to code and implemented for real-time application. One such emerging technology is InternationalTechnology Alliance (ITA) Controlled English, which is a controlled natural language that is unambiguous for computers and allows for the definition and expression of concepts,

rules, and relationships (Harries, Braines & Gibson, 2012). Another available tool is Engineering and Computer Simulations' (ECS) Student Information Models for Intelligent Learning Environments (SIMILE), which uses a set of standards-based data models and protocols that associate events within a game to learning objectives, tasks conditions, and standards for a given lesson/scenario (ECS, 2013). Researchers should explore the application of these tools and others like it for authoring assessments as they relate to particular activities taken within a gaming environment.

In the current version of GIFT, there is a use case within Virtual Battle Space 2 (VBS2), a game-engine used by the Department of Defense for training purposes. In this instance, assessments are based around Distributed Interactive Simulation (DIS) Protocol Data Units (PDUs) that provide entity and environment state data information, and how they relate to defined objects (e.g., waypoints and time sequences). An example is determining what entrance a player used to enter a building, with waypoints placed at each entrance. When a player crosses one of these objects, the system can determine exactly which doorway a player used and this is then passed to GIFT for assessment purposes. If the intention is for a player to select a specific entrance, formative feedback can be triggered when a non-optimal doorway is selected. Currently, each defined object is identified through the VBS2 mission editor and then translated into the DKF. Research should be conducted looking at approaches to link game mission/scenario editors with the associated DKF. This would enable autonomous populating of fields within the XML schema based on defining attributes among available elements within the application. As related to the building entrance example above, an author could identify a location on a scenario map as an object for use in GIFT assessment, rather than the individual having to pull the data and entering them manually into the DKF.

Another area of interest in game-based intelligent tutoring is being able to predict performance as a player progresses through a scenario so that interventions can be applied before user actions lead to poor performance outcomes. Research related to this function is the application of Markov decision processes (MDPs) that map associated goal objectives to state spaces in a game. Current applications include reinforcement learning and partially observable MDPs (Sigaud & Buffet, 2010; Folsom-Kovarik, Sukthankar & Schatz, in press). MDPs can be applied to determine if a player is heading down a non-optimal path as it relates to defined standards, and can be used to deem what goals are being valued as it relates to meeting scenario objectives. It is recommended that researchers look into how MDPs can be used in GIFT for predicting performance outcomes and providing diagnostics into what elements of their interaction are causing the decrements in performance.

Discussion and Future Research

Well-designed games can provide meaningful assessment environments by providing players with problems that require the application of various competencies, then monitoring their performance. In this chapter, we presented an assessment methodology that enables us to develop tasks in digital games designed to elicit specific performance data that are then statistically linked to our focal competencies.

The research can expand in a number of general directions. That is, we can push the bounds of our stealth assessments relative to implementing the models in additional digital games as well as other digital learning environments to determine the range of environments that may employ the same competency and evidence models for scalable, cost-effective, and engaging solutions to the assessment of complex competencies. In addition, we can examine any added value of including exploratory, data-mining methods to stealth assessment's more theoretically driven approach regarding the quality of the assessment.

As we consider the development of stealth assessments for a diverse range of games and other learning environments, the need for an interface to communicate equally diverse data to GIFT's domain module

becomes apparent. As we alluded to in the previous section, the interface must be portable and flexible, both in terms of technology and content. Thus, it is important to research elements that are common to stealth assessments across learning environments and areas of assessment. The evidence-based models we developed for NP is a good start.

Regarding future research related to learning, stealth assessment has the potential to be quite useful for diagnostic purposes due to the fine-grained analysis of student behavior in situated contexts. In addition, real-time information about player competency states can be useful to support learning through hints and feedback, as well as dynamic matching of game difficulty level to player ability (e.g., providing more challenging problems for those with high levels of various skills). Regarding the example used in this chapter, the indicators linked to the agents of force and motion can serve as the basis for diagnoses. For instance, if a student created a lever that did not successfully solve a problem that could have been solved with a lever, the indicators would inform the most likely reason(s) why. That is, the lever may have failed given (1) the wrong mass of an object that was used on one side of the lever, (2) the fulcrum was positioned inaccurately, and/or (3) the size/length of the lever was too short or too long. Those data (mass, position, and length) are calculated as part of the stealth assessment.

We are excited that researchers are starting to use digital games for learning and assessment. We think stealth assessment is one way to maximize the positive impact digital games can have on students. As a result, the future developmental efforts of GIFT should aim at identifying authoring tools and methods that ease the process of embedding stealth assessment capabilities in game-based learning environments.

Acknowledgements

We would like to sincerely thank the Bill and Melinda Gates Foundation for funding the stealth assessment project described herein.

References

Arum R. and Roska, J. (2011). *Academically adrift: Limited learning on college campuses*. Chicago, IL: The University of Chicago Press.

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

Caramazza, A., McCloskey, M. & Green, B. (1981). Naive beliefs in "sophisticated" subjects: Misconceptions about trajectories of objects. *Cognition, 9,* 117-123.

Crouch, C. A. & Mazur, E. (2001) Peer instruction: Ten years of experience and results. *American Journal of Physics, 69*, 970-977.

Dede, C. (2005). Planning for neomillennial learning styles. *EDUCAUSE Quarterly, 28*(1), 7-12.

DiCerbo, K. E. & Behrens, J. T. (2012). Implications of the digital ocean on current and future assessment. In R. Lissitz & H. Jiao (Eds.) *Computers and their impact on state assessment: Recent history and predictions for the future* (pp. 273-306). Charlotte, NC : Information Age Publishing.

diSessa, A. A. (1982). Unlearning Aristotelian physics: A study of knowledge-based learning. *Cognitive Science, 6,* 37-75.

Feynman, R. P., Leighton, R. B. & Sands, M. (1964). *The Feynman lectures in physics*.City, ST: Addison Wesley.

Feynman, R. P. (1964). *The character of physical law*, Cornell, NY: University Press.

Folsom-Kovarik, J.T., Sukthankar, G. & Schatz, S. (In Press). Tractable POMDP Representations for Intelligent Tutoring Systems. *ACM Transactions on Intelligent Systems and Technologies.*

Gee, J. P., Hull, G. A. & Lankshear, C. (1996). *The new work order: Behind the language of the new capitalism.* St. Leonards Australia: Allen & Unwin.

Gronhaug, K. & Kaufman. G. (Eds.) (1988). *Innovation: A cross-disciplinary perspective*. Oslo, Norway: Norwegian Universities Press/Oxford University Press.

Hake, R. R. (1998). Interactive engagement vs. traditional methods in mechanics instruction. *American Journal of Physics, 66*(1), 64-74.

Halloun, I. (1996) Schematic modeling for meaningful learning of physics, *Journal of Research in Science Teaching, 33,* 407-431.

Halloun, I. & Hestenes, D. (1985). Initial knowledge state of college physics students, *American Journal of Physics, 53,* 1043-1055.

Harries, D., Braines, D. & Gibson, C. (2012). Towards an expression of Policy in Controlled English. *In the Proceedings of the 6th Annual Conference of the International Technology Alliance*. Botley, UK, September 2012.

Hestenes, D. & Wells, M. (1992). A mechanics baseline test. *The Physics Teacher, 30*(3), 159-167.

Hestenes, D., Wells, M. & Swackhamer, G. (1992). Force Concept Inventory. *The Physics Teacher, 30,* 141-151.

Hewitt, P. G. (2009). Conceptual physics (11th ed.). San Francisco, CA: Pearson Education.

Hiebert, E. H., Valencia, S. W. & Afflerbach, P. P. (1994). Understand authentic reading assessment: Definitions and perspectives. In S. W. Valencia, E. H. Hiebert & P. P. Afflerbach (Eds.), *Authentic reading assessment: Practices and possibilities* (pp. 6-21). Newark, DE: International Reading Association.

Kellaghan, T. & Madaus, G. F. (1991). National testing: Lessons for America from Europe. *Educational Leadership, 49*(3), 87-93.

Lenhart, A., Kahne, J., Middaugh, E., Macgill, A. R., Evans, C. & Vitak, J. (2008). *Teens' gaming experiences are diverse and include significant social interaction and civic engagement*. Washington, DC: Pew Internet & American Life Project.

Madaus, G. & O'Dwyer, L. (1999). A short history of performance assessment. *Phi Delta Kappan, 80*(9), 688-695.

Masson, M. E. J., Bub, D. N. & Lalonde, C. E. (2011). Video-game training and naive reasoning about object motion. *Applied Cognitive Psychology, 25,* 166-173.

McCloskey, M. & Kohl, D. (1983). Naive physics: The curvilinear impetus principle and its role in interactions with moving objects. Journal of Experimental Psychology: *Learning, Memory & Cognition, 9,* 146-156.

McDermott, L. (1993). How we teach and how students learn – A mismatch? *American Journal of Physics, 61*, 295-298.

Quellmalz, E.S., Timms, M. J., Buckley, B. C., Silberglitt, M. & Brenner, D. (2012) SimScientists: Measurement in simulation-based science assessments Manuscript submitted for publication.

Reiner, C., Proffit, D. R. & Salthouse, T. (2005). A psychometric approach to intuitive physics. *Psychonomic Bulletin and Review, 12,* 740–745.

Shepard, L.A. (1991). Will national tests improve student learning? *Phi Delta Kappan, 72,* 232-238. Available online at: www.cse.ucla.edu/products/Reports/TECH342.pdf

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. & Ke, F. (2012). Games, learning, and assessment. In D. Ifenthaler, D. Eseryel & Ge, X. (Eds.), *Assessment in game-based learning: Foundations, innovations, and perspectives* (pp. 43-58). New York, NY: Springer.

Shute, V. J. & Ventura, M. (2013). Measuring and supporting learning in games: Stealth assessment. Cambridge, MA: The MIT Press.

Sigaud, O. & Buffet, O. (2010). *Markov Decision Processes in Artificial Intelligence*. Wiley-IEEE Press.

Swann, W. F. G. (1950). The teaching of physics. *American Journal of Physics, 19*(2), 182-187.

Tobias, S. & Fletcher, J. D. (Eds.) (2011). *Computer games and instruction*. Charlotte, NC: Information Age Publishers.

Tobias, S., Fletcher, J. D., Dai, D. Y. & Wind, A. P. (2011). Review of research on computer games. In S. Tobias & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 127-222). Charlotte, NC: Information Age.

White, B. Y. (1994). Designing computer games to help physics students understand Newton's laws of motion. *Cognition and Instruction, 1*(1), 69-108.

Wilson, K. A., Bedwell, W., Lazzara, E. H., Salas, E., Burke, C. S., Estock, J., … Conkey, C. (2009). Relationships between game attributes and learning outcomes: Review and research proposals. *Simulation & Gaming, 40*(2), 217-266.