Chapter 13

DESIGN OF GAME-BASED STEALTH ASSESSMENT AND LEARNING SUPPORT

AUTHOR(s) – Fengfeng Ke & Valerie Shute, Florida State University

Abstract: In this chapter, we describe the processes of designing and validating game-based learning assessment and/or support in two different games--Portal 2 (by Valve Corporation) and Earthquake Rebuild. The games represent cases of possible game-based learning (i.e., domain-generic and domain-specific), and provide good vehicles for testing the design decisions underlying stealth assessment and learning support. The chapter starts with a critical review of prior research on game-based assessment of competencies and learning support mechanisms in games, and then focuses on the particular design processes and findings from our two game cases. The review and findings suggested that process-oriented data mining and learning analytics methods help to capture the complex and open-ended learning trajectories in a game setting. They also illustrated how the evidence-centered assessment design and the learning context/task design should and can be interweaved in the early phase of game development. We conclude with a discussion relevant to developing and integrating the assessment and support of learning into other learning-game platforms.

Key words: Game-based learning assessment, Learning support, stealth assessment

1. INTRODUCTION

In this chapter, we review and examine two important issues related to the next generation of learning games: (a) the real-time capture and analysis of gameplay performance data (i.e., game-based stealth assessment), and (b) the provision of adaptive learning supports based on the assessment information.

Historically, learning in games has been assessed indirectly and/or in a post hoc manner (Shute, 2011). What's needed instead is real-time and valid assessment of learning based on the dynamic performance of players, which should be seamlessly woven into the game to capture play-based competency development. This assessment information would then provide the basis for targeted and dynamic learner support. Incidental learning is often a consequence of playing well-designed games (MacCallum-Stewart, 2011; Prensky, 2001). However, creating substantive, improvisational learning experiences in games is difficult because knowledge and skill acquisition usually involves conscious elements (e.g., processing information, constructing mental models) in addition to subconscious processes, such as insight. A relevant game design hypothesis is that learning within gameplay will proceed from being improvisational (i.e., acting spontaneously in the environment without pre-planning) to meta-reflective (i.e., considering various points of view), or moving from a tacit experience to an aware, strategic, and reflective application of the target knowledge/skills. The underlying challenge of this design hypothesis is to integrate the learning-analytics-based support (or scaffolding) of meta-reflective learning into the game world and mechanics while not disrupting what is enjoyable about games.

In this chapter, we will describe the processes of designing and validating game-based assessment and/or learning support in two games. The first game, Portal 2 (by Valve Corporation), is an existing, commercial off-the-shelf (COTS) game that we hypothesized would foster spatial skills. The second game, Earthquake Rebuild, is currently under development. It is an architectural game that aims to promote mathematical understanding and math-related problem solving skills. The two games represent typical cases of possible game-based learning (i.e., domain-general and domain-specific), and provide good vehicles for testing the design decisions underlying game-based stealth assessment and learning support. The chapter starts with a critical review of prior research on game-based assessment of domain-relevant competencies and learning support mechanisms in games, and then focuses on the particular design processes and findings from our two game cases. We conclude with a discussion relevant to developing and integrating the assessment and support of learning into other learning-game platforms.

2. LITERATURE REVIEW

There is rapidly-growing interest in data mining and analytics in education, learning sciences, and other academic fields. Research on the automated collection or monitoring of user-generated data has been conducted in multiple fields, such as that on telemetry in computer science (Yairi et al., 2006) and geospatial data mining in Geographic Information System (Miller & Han, 2009). Educational data mining (EDM), highlighted in this chapter, is the process of exploring and extracting descriptive patterns from large amounts of data--"big data"--in educational settings (e.g., logs of student-computer interaction), to provide insights into instructional practices and student learning (Baker & Yacef, 2009; Romero, Ventura, Pechenizkiy, & Baker, 2011; Witten & Frank, 1999). In recent years, EDM has been used to infer students' computer-supported learning engagement and behaviors and hence the development of effective and dynamic learning support (e.g., Baker, Corbett, & Koedinger, 2004; Baker, 2007; Beck & Mostow, 2008; Shute, Ventura, & Kim, 2013).

Closely related to EDM is the learning analytics research that refers to collecting, measuring, analyzing, and reporting data about learners and contexts to understand and optimize learning and the environments in which it occurs (SoLAR, 2011). Similar to EDM, learning analytics (LA) focuses on data-intensive approaches to education, although EDM often uses automated discovery with models while LA leverages more human judgment (Siemens & Baker, 2010).

Prior research has suggested that both EDM and LA can and should be used together to exploit game-based performance data to inform on students' attributes, their on-task or off-task behaviors, competency development related to the targeted subject matter, and hence the effectiveness and design of learning supports. Four recent projects (by Levy, 2014; Shute et al., 2013; Shaffer et al., 2009; Dede, 2012) can exemplify the current state of game-based learning assessment via EDM and/or LA.

2.1 Game-based learning assessment through data mining and analytics

Evaluation of Save Patch: In a recent study, Levy (2014) employed the approach of evidence-centered assessment design (Mislevy, Steinberg, & Almond, 2003) and the method of Bayesian Networks to evaluate student performance in Save Patch, an educational game targeting rational numbers in math.

The process started by a cluster analysis that classifies the gameplay log to extract a list of solution strategies (behaviors) for successful gameplay actions, and that of misconceptions associated with unsuccessful actions. The cluster analysis results then served as categories of values of observable variables. For each targeted math competency in the game, Levy (2014) specified a dichotomous latent variable with its categorical values as mastery (coded as 1) and nonmastery (coded as 0). A dichotomous latent variable was also specified for each of the misconceptions, with its categorical values coded as 1 or 0 based on whether the student possessed that misconception or not. A Bayesian network model was created and calibrated to investigate conditional probabilities of observable categories (values) of each latent variable for individual students (or specific student groups) at different points in time and for each game level. The constructed psychometric model also encompassed transitions from nonmastery to mastery in certain latent variables by specifying the probability that a student is a master at time $t+1$, given they were a nonmaster at time t and had a particular value for the observable at time t.

Game-based stealth assessment: Similar to Levy (2014), Shute and her colleagues (2010, 2013) adopted the approach of *evidence-centered assessment design (ECD)* to design and validate the framework of educational assessments in terms of user-generated in-game (gameplay) data, named stealth assessment, in various game evaluation studies. For example, Shute and her colleagues described the development of competency, evidence, and task models for the assessment of systems thinking in the game Taiga Park (Shute, Masduki, & Domnez, 2010). The ECD-based, stealth assessment framework has also driven the design and validation of Physics Playground (formerly called Newton's Playground), a learning game intended to help secondary school students understand qualitative physics (Shute & Ventura, 2013). The central evidentiary component of stealth assessment for Physics Playground is the game log file that captures multiple gameplay variables (e.g., time spent on a level, number of trials, types of objects created, the trajectory of objects, number of gold trophies obtained). Analyses revealed a significant correlation between in-game assessment indicators (e.g., gold trophies earned) and the external learning measure (qualitative physics test score). Furthermore, students (167 middle school students) significantly improved on the external physics test (administered before and after gameplay) despite no formal instruction in the game. Students also enjoyed playing the game (reporting a mean of 4 on a 5-point scale in where 1 = strongly dislike and 5 = strongly like), and boys and girls equally enjoyed the game.

Epistemic network analysis for epistemic games: Different from the aforementioned studies that emphasized data mining with a quantitative, psychometric modeling approach, Shaffer and his colleagues (2009) adopted a learning analytic approach by collecting and analyzing qualitative data from the game-extended records and interactions in gameplay. Specifically, they performed a systematic coding and aggregation with the qualitative data to identify salient elements of an epistemic frame (i.e., competency), then quantified the coded results by calculating the co-occurrence frequency of each pair of epistemic frame elements. They then created a cumulative network graph that is similar to a social network, "where frame elements (nodes) that are linked more often in the data are closer to each other than those that are linked less often in the data" (p. 7). In the structural network analysis, the unit of analysis is a strip (or segment) of activity "into which ongoing activities are divided for the purpose of analysis" (p. 8). By summing the strips of activities up to a particular time, the trajectory of development of an epistemic frame can be mapped as a dynamic network graph, or series of slices (or phases) over time, with each slice showing the state of the players' epistemic frame at certain time. A descriptive, visual comparative analysis can then be used to examine the trajectories of frame development of a sub-group of players (e.g., novices versus experts), or the knowledge structure of the targeted competency (e.g., by computing the relative weight or centrality of each node in the epistemic network).

EcoMUVE assessment: With an emphasis of data visualization, Dede (2012) described multiple analytical methods relating to learning trajectories in virtual-reality-based, complex inquiry tasks. These methods include:

- Event path analysis and visualization via the heat map. This method involves using the server-side log data to generate event paths and then providing a visual and diagnostic analysis on players' scientific inquiry skills. The path analysis comprises a series of visual slides depicting the relative frequencies of learning events performed by subpopulations of students, aggregated by pre-specified virtual-world location and time unit, for comparative analyses (e.g., high-performing vs. low-performing students). The heat map shows which hotspots the players prefer--where hotspots are highlighted and can be used diagnostically to inform various misconceptions.
- Behavior analysis with the usage of guidance tools and pedagogical agents. This process uses the prediction analysis (e.g., regression and correlation analyses) to examine the effects of various learning support mechanisms and how they relate to student performance. The guidance tool uses individual players' interaction histories to generate real-time, customized support.
- Structured benchmarking task assessments. The last method entails a series of mini modules (or inquiry tasks) in the virtual reality environment. The tasks are created as benchmarking assessments to provide information on skill mastery and promote transfer of learning.

In summary, the aforementioned game-based learning projects illustrate the multi-faceted nature of assessment through data mining and learning analytics. All projects adopt a data-intensive, evidence-based approach, but differ in terms of: (a) the assessment objectives (i.e., to model or predict students' competency development, or to analyze the structure of domain-specific competency or epistemic frame, or to examine the association between the learning trajectory and the learning support and context design), (b) the resources of data (e.g., in-game log data, or game-extended behaviors), (c) analysis methods (e.g., quantitative psychometric modeling, network or structural analysis, and path analysis), and (d) type of visualization (e.g., algorithms, models, network graphs, or spatial and chronical maps).

2.2 In-game learning support

In a recent meta-analysis that synthesized 29 studies on instructional/learning support in game-based learning, Wouters and van Oostendorp (2013) classified learning support features into two major categories – ones that support the selection of relevant information, and ones that facilitate information organization and integration via reflection and explication. Of the articles reviewed in the study, more than half explicitly studied the in-game learning support features. These in-game support features are based on their associations with game-design elements (i.e., game world, game actions, and rules), and can be categorized as: (a) cues and feedback, (b) explicit training or instruction, (c) probes or prompts for self-explanation and reflection, (d) in-game learning tools, (e) incentive structures, and (f) level sequencing or progression.

Adaptive instructional or learning support is emerging as a prominent feature of serious and learning games (Leemkuil & de Jong, 2012; Kickmeier & Albert, 2010; O'Rourke et al., 2014). Adaptive feedback, intelligent pedagogical agents, and adaptive level progression, in particular, feature prominently in such games. For example, O'Rourke et al. (2014) designed four metrics (named “brain points”) to capture and reward players’ novel and incremental content-related game performance. They found that the “brain points” version of the game, in comparison with a control version of the game, increased overall time played, strategy used, and perseverance after challenge. Hwang et al. (2012) examined the role of game level sequencing or navigation in a role-playing science game. They reported that students who learned with the personalized game level sequencing (by matching their learning styles with the game level navigation style – linear or non-linear) showed significantly greater learning achievement, motivation, and acceptance toward game-based learning than those who learned with the game without personalized sequencing. These support tools or mechanisms are typically based on the non-intrusive, stealth assessment of in-game performance via the creation and tracking of evaluation indices and threshold values (Shute et al., 2013; Zapata-Rivera, VanWinkle, Doyle, Buteux, & Bauer, 2009).

We now present two more detailed examples of game-based assessment design. In Example 1, we report a completed, controlled evaluation of domain-generic skills development in a COTS game. In Example 2, we describe how the development of a domain-specific, stealth assessment mechanism is aligned and associated with the design of the game world, game mechanics, and in-game learning support in an underdeveloped math learning game.

3. GAME-BASED LEARNING ASSESSMENT DESIGN FOR PORTAL 2

3.1 Portal 2

Portal 2 is the name of a popular linear first-person puzzle-platform video game developed and published by Valve Corporation. Players take a first-person role of Chell in the game and explore and interact with the environment. The goal of Portal 2 is to get to an exit door by using a series of tools. The primary game mechanic in Portal 2 is the portal gun, which can create two portals. These portals are connected in space, thus entering one portal will exit the player through the other portal. Any forces acting on the player while going through a portal will be applied upon exiting the portal. This allows players to use, for example, gravity and momentum to "fling" themselves far distances through the air. This simple game mechanic is the core basis of Portal 2.

Other tools that may be used to solve puzzles in Portal 2 include Thermal Discouragement Beams (lasers), Excursion Funnels (tractor beams), Hard Light Bridges, and Redirection Cubes (which have prismatic lenses that redirect laser beams). The player must also disable turrets (which shoot deadly lasers) or avoid their line of sight. All of these game elements can help in the player’s quest to open locked doors, and generally help (or hinder) the character from reaching the exit. The initial tutorial levels in Portal 2 guide the player through the general movement controls and illustrate how to interact with the environment. Characters can withstand limited damage but will die after sustained injury. There is no penalty for falling onto a solid surface, but falling into bottomless pits or toxic pools kills the player character immediately.

There are several plausible ways for a person to acquire and hone spatial skills as a function of gameplay in Portal 2.

3.2 Spatial skills

Of particular importance in understanding the role of video gameplay relative to spatial cognition is the distinctions among: (1) figural, (2) vista, and (3) environmental spatial skills (Montello, 1993; Montello & Golledge, 1999). Figural spatial skill is small in scale relative to the body and external to the individual. Accordingly, it can be apprehended from a single viewpoint. It includes both flat pictorial space and 3D space (e.g., small, manipulable objects). It is most commonly associated with tests such as mental rotation and paper folding tasks. Vista spatial skill requires one to imagine an object or oneself in different locations—small spaces without locomotion. Vista spatial skill is useful when trying to image how the arrangement of objects will look from various perspectives (Hegarty & Waller, 2004). Environmental spatial skill is large in scale relative to the body and is useful in navigating around large spaces such as buildings, neighborhoods, and cities, and typically requires locomotion (see Montello, 1993, for a discussion of other scales of space). It usually requires a person to mentally construct a cognitive map, or internal representation of the environment (Montello & Golledge, 1999). Environmental spatial skill depends on an individual's configurational knowledge of specific locations in space, and is acquired by learning specific routes. Configurational knowledge depends on the quality of an individual's cognitive map, or internal representation of an environment. In this map-like representation, all encountered landmarks and their relative positions are accurately represented.

A game like Portal 2 has the potential to improve spatial skills due to its unique 3D environment that requires players to navigate through problems in often complex ways. Over the past 20 years, a growing body of research has shown that playing action video games can improve performance on tests of spatial cognition and selective attention (e.g., Dorval & Pepin, 1986; Feng, Spence, & Pratt, 2007; Green & Bavelier, 2003, Spence, Yu, Feng, & Marshman, 2009; Uttal et al., 2012). Recently, Ventura, Shute, Wright, and Zhao (2013) showed that self-reported ratings of video game use were significantly related to all three facets of spatial cognition, and most highly related to environmental spatial skill. Feng et al. (2007) found that playing an action video game improved performance on a mental rotation task (i.e., small-scale or figural spatial cognition). After only 10 hours of training with an action video game, subjects showed gains in both spatial attention and mental rotation, with women benefiting more than men. Control subjects who played a non-action game showed no improvement.

Recently, Uttal et al. (2012) conducted a meta-analysis of 206 studies investigating the effects of training on spatial cognition. Of these 206 studies, 24 used video games to improve spatial skills. The effect size for video game training was .54 (SE = .12). Findings like these have been explained due to the visual-spatial requirements of 3D action games which may enhance spatial skills (e.g., Feng et al., 2007; Green & Bavelier, 2003; 2007).

3.3 External measures of spatial skills

To both validate the in-game (stealth) measures of spatial skills (e.g., number of portals shot on average, per level) and test for any learning of them from eight hours of gameplay, Shute et al. used three existing, validated assessments for figural, vista, and environmental spatial skills. To measure figural (or small-scale) spatial skill, they used the Mental Rotation test (Vandenberg & Kuse, 1978). To assess vista spatial skill, they administered the Spatial Orientation Test (Hegarty & Waller, 2004). And to measure environmental spatial skill, they developed and validated an assessment called the Virtual Spatial Navigation Assessment. Each is now described.

- *Mental Rotation Test* (MRT). The MRT was adapted from Vandenberg and Kuse (1978). In this test, participants view a three-dimensional target figure and four test figures. Their task is to determine which of the test figure options represent a correct rotation of the target figure. The total score is based on the total number of items where both correct objects are found.
- *Spatial Orientation Test* (SOT). The SOT requires the participant to estimate locations of objects from different perspectives in one picture (Hegarty & Waller, 2004). In each item the participant is told to imagine looking at one object from a particular location in the picture and then point to a second location. Each response is scored as a difference between the participant's angle and the correct angle (scores range from 0-180 degrees). Larger differences between a participant's drawn angle and the correct angle indicate lower vista spatial skill.
- *Virtual Spatial Navigation Assessment* (VSNA). The VSNA (Ventura, Shute, Wright, & Zhao, 2013) was created in Unity. In the VSNA, a person explores a virtual 3D environment using a first person avatar on a computer. Participants are instructed that the goal is to collect all the gems in an environment and return to the starting position. Participants first complete a short familiarization task that requires them to collect colorful gems in a small room. The VSNA consists of an indoor environment consisting of halls in a building (i.e., a maze), and an outdoor environment consisting of trees and hills. In each environment the participant must collect the gems twice--training and testing phases. The VSNA collects data on the time taken to collect all gems and return to the starting position) as well as the distance traveled in the training and testing phase of an environment. The main measure used in the current study consists of the time to collect all gems and return home. Less time suggests greater navigational skill.

3.4 Results from a controlled evaluation of Portal 2

A recent study reported by Shute, Ventura, and Ke (2015) tested 77 undergraduates who were randomly assigned to play either Portal 2 or a control game condition (i.e., the popular brain training game suite called Lumosity) for 8 hours. Before and after gameplay, participants completed a set of online tests related to their spatial skills. Results revealed that participants who were assigned to play Portal 2 showed a statistically significant advantage over Lumosity on the composite measure of spatial skill. Portal 2 players also showed significant increases from pretest to posttest on specific small-scale (MRT) and large-scale (VSNA) spatial tests while those in the Lumosity condition did not show any pretest to posttest differences on any measure. Finally, Portal 2 *in-game performance data* (e.g., number of portals shot on average, per level) significantly correlated to MRT and VSNA after controlling for the respective pretest scores. These findings suggest that performance in Portal 2 predicts outcomes on different (small- and large-scale) spatial measures beyond that predicted by their respective pretest scores.

The improvement of subjects on their spatial skills as a function of playing Portal 2 is likely due to the repeated requirement in Portal 2 to apply and practice their spatial skills to solve problems. This result supports other work investigating video game use and spatial skill (e.g., Feng et al. 2007; Uttal et al., 2012; Ventura, Shute, Wright, & Zhao, 2013). There were no improvements for the Lumosity group on any of the three spatial tests. Overall, the findings of between-group differences on the MRT and VSNA measures, combined with the significant Portal 2 pretest-posttest gains in MRT and VSNA, give strong evidence that playing Portal 2 causes improvements in small- and large-scale spatial skills. Moreover, the fact that a conservative control group was used gives even greater credence to the finding that playing Portal 2 can improve spatial skills over other game-related activities that claim to improve cognitive skills (i.e., Lumosity games). Finally, while video gameplay has been previously shown to improve MRT performance (e.g., Uttal et al., 2012), this is the first research study to provide experimental evidence that video game play can improve performance in large-scale spatial skill.

4. GAME-BASED LEARNING ASSESSMENT AND SUPPORT DESIGN FOR EARTHQUAKE REBUILD

Earthquake Rebuild (E-Rebuild) is a 3D architecture game that intends to promote versatile representation and epistemic practice of mathematics in design and building quests (Ke et al., 2014), and is on the development and user testing phase at the time of writing. The overall goals of E-Rebuild are to plan, design, and rebuild an earthquake-damaged space to fulfill diverse design parameters and needs. The intermediate game goals involve completing each level of the design quest to gain new tools, construction materials, and credits (e.g., game scores in terms of architectural design efficiency, structural soundness, and complexity in structures--which comprise an overall credit that enables a player to perform subsequent game levels).

A learner in E-Rebuild performs multiple types of gameplay (or architectural design) actions: *collection*, *construction*, space and energy *allocation*, and materials *trading*. All four gameplay actions act as both the source and the application of math understanding. The target math topics of E-Rebuild, aligned with the Common Core State Standards (CCSS) for mathematics Grade 6-8 (CCSSI, 2010), are: (a) ratio and proportional relationships; (b) angle measure, area, surface area, and volume; and (c) numeric and algebraic expressions.

Different from most projects in which learning assessment design (via EDM or LA) is a post-hoc practice conducted after game development, E-Rebuild is integrating stealth assessment design directly into the game design process. This section introduces the process of interweaving assessment of learning and game design in this *ongoing*, design-based research project.

4.1 Interweaving the design of game world and that of the game log file

A major component of the game world design in E-Rebuild is to design various game objects, such as constructional materials (e.g., planks, pillars, bricks, prefabricated container houses) and game characters (e.g., victims or residents to be accommodated). The design of the relationship structures and the properties of these objects are aligned with the design of the game log file in terms of the variables and events logged. For example, the key properties of each construction element include its mass (solid vs. hollow, primitive vs. composite), texture, geometric form, size, volume, location, and position or angle. With each object and its element there will be a list of potential actions to be performed, such as clicking, moving, joining, cutting, and scaling. The original state of the objects' properties, the specific actions performed, and hence the state or characteristic change (e.g., increased happiness) following the actions performed (along with the time stamp and occurrence frequency), will all be captured in the game log file for a future sequential analysis.

4.2 Aligning game mechanics with competency-based learning actions

To enable an authentic, performance-based assessment, we align the E-Rebuild game mechanics (i.e., gameplay actions and rules) with math learning actions. Specifically, the integration of two gameplay modes (i.e., the adventure and construction modes) aims to extract integral, multi-stranded math learning actions. That is, in the adventure mode (see Figure 1), players are requested to engage in exploration- and collection-based math concept representation (e.g., identifying a construction item in a specific prism and size) and experience-based reflection (e.g., evaluating their math-specific design performance by seeing how a designed structure collapsed in the earthquake or failed to address the needs). In the construction mode (see Figure 2), players are mainly involved in construction-oriented math calculation and problem solving (e.g., cutting/scaling an item to a desirable size, measuring/rotating the construction site based on a landmark, managing materials).

Figure 1. Collection action in the adventure mode

Figure 2. Construction action in the construction mode

4.3 Designing game tasks based on the competency and evidence models

A *game task library* is being developed based on the target math competencies and the corresponding specifications of the competency and evidence models for the game-based stealth assessment. The competency and evidence models are being explicitly aligned with the Common Core State Standards (CCSS). They follow the structure of a Bayesian network, and have guided the design of specific game tasks and the arrangement of these tasks within and across game levels.

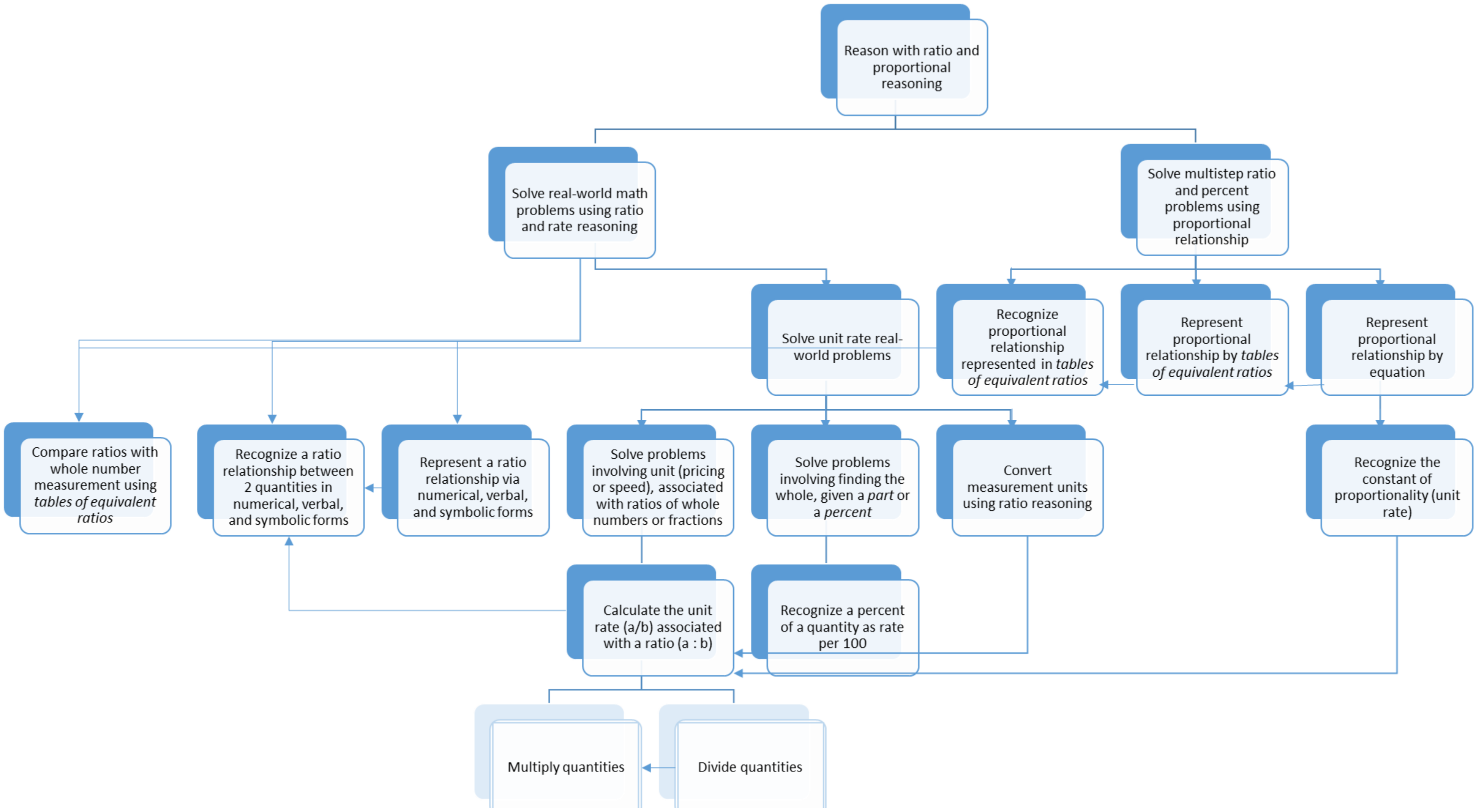

Figure 3. Competency model of the ratio and proportional reasoning defined by CCSS for grades 6-8

4.4 Representing learning in the game-scoring mechanism

The major variables used to evaluate successful knowledge and skill acquisition in E-Rebuild include: (1) *time* taken to complete the current task (e.g., tasks are speeded with a risk/progress bar related to an earthquake-hit), and (2) successful handling of multiple *design constraints* imposed by the needs of the area's residents, the landscape, and the limited construction materials. The first criterion measures the fluency while the second criterion measures the accuracy of math-related architectural problem-solving performance. A composite game score, along with sub-scores embedded in the game reward mechanism (e.g., time credit, material credit, happiness of residents), is then calculated based on the evaluation of the aforementioned evaluation criteria and presented to portray a player's learning profile.

4.5 Learning support design as both the source and the application of data mining

Intuitive interfaces are important to successful human-computer interactions. In E-Rebuild, we design in-game learning supports as an intermediary interface between the player and the game. This interface will support content engagement during gameplay while capturing the processes related to solving a complex math task. For example, a user-testing, comparative analysis with the control-meter interface and the current text-entry box (for feeding numerical values of the x, y, and z coordinates in a scaling tool, see Figure 4) indicates that the text-entry box is obviously associated with less wild guessing or trial-and-error play and more mindful math calculations. Every attempt of using this specific scaling tool, along with the values entered, is captured in the game log file to enable a diagnostic analysis. The results of the diagnosis will then be presented as dynamic feedback in a Scratch Pad screen (see Figure 5). This scratch pad also includes an internal calculator and enables the typing of calculation steps, thus working as the record of mathematical processing performed by the player for the future data mining.

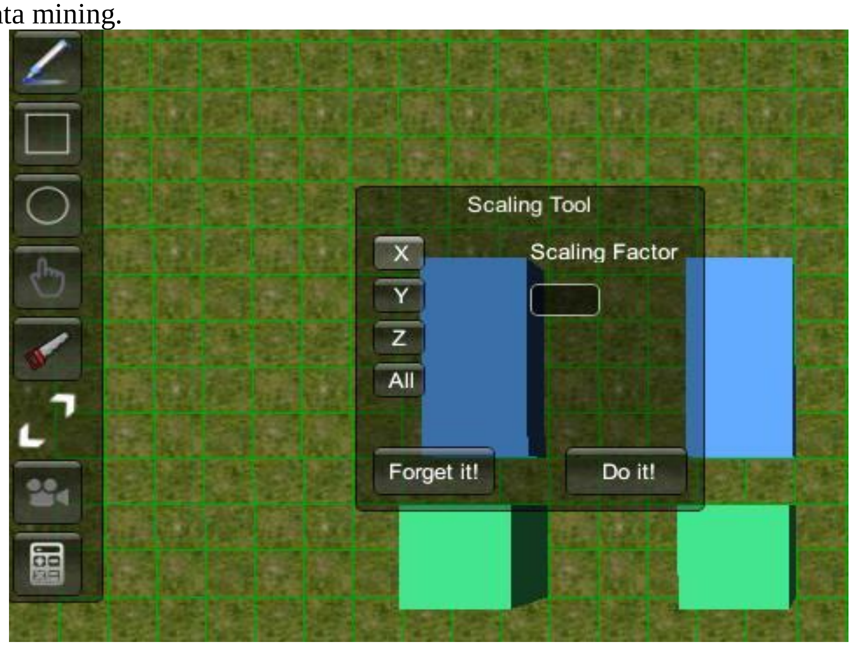

Figure 4. Scaling tool

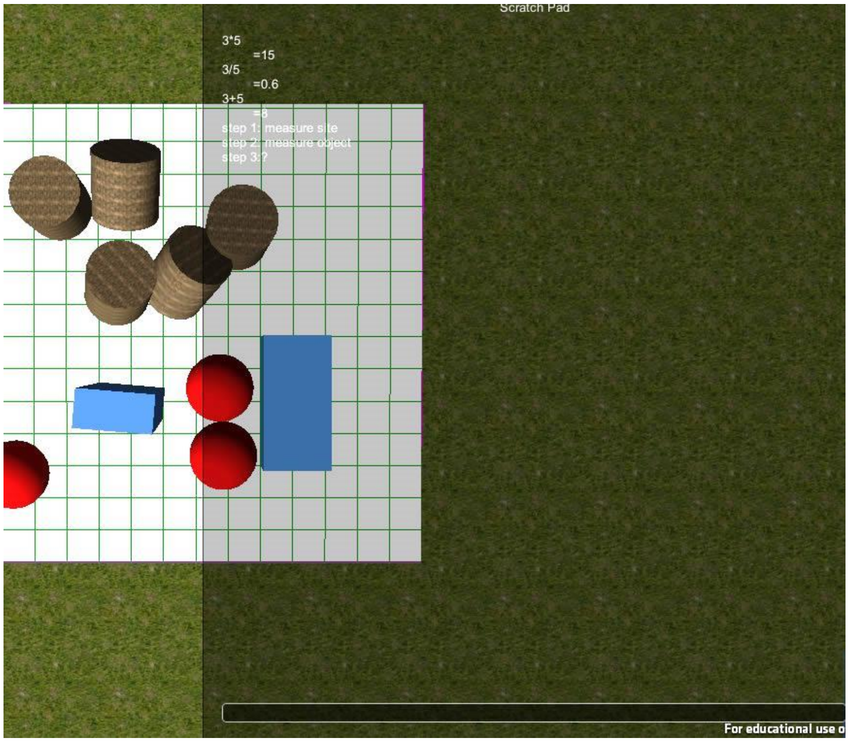

Figure 5. Scratch pad

5. CONCLUSIONS, DISCUSSION, AND FUTURE RESEARCH

5.1 Heuristics of game-based learning assessment design

A salient feature of the aforementioned game-based assessment projects is the diagnostic and formative measurement of multiple domain-relevant steps or cognitive processes underlying each task solution or performance. Process-oriented data mining and learning analytics methods, such as Bayesian networks, social networks or structural analysis, visual or graphical analysis of event paths, and sequential analysis of time series, will capture the complex and open-ended learning trajectories in a game setting.

Game-based assessment should leverage and integrate both quantitative, model-based automatic discovery and qualitative interpretation with human judgment. Frequently, the interpretation and extraction of meaningful patterns from the game log or extended performance data are in need of the perspectives and expertise of stakeholders (e.g., content experts, game designers, student users). The rules for evidence identification and the categorization of observable values also emerge based on the integration of expert decision and data-driven calibration. The sources and products of analytical methods in game-based learning assessment, as the examples in this chapter illustrated, comprise not only numerical values and algorithms but also discourses, descriptive frames, and graphical models.

Notably, prior work in this area, as well as our own research has suggested that performing diagnostic and stealth assessment with game-based learning is especially challenging when the assessment strategies are a post-hoc design decision enforced on an existing game. Both Levy (2014) and Dede (2012) have reported on particular challenges of using data mining or learning analytics techniques to evaluate learning in an existing game or simulation. These challenges include but are not limited to the difficulty of inferring knowledge-mastery transition due to the insufficient and unbalanced task-specific data across game levels, and the difficulty of mapping the event path when a game-based learning task does not involve location exploration. Similarly, it is difficult to collect and analyze all action-based evidences of spatial skills in the Portal 2 study since only part of gameplay actions or object attributes were recorded in the game log file. Moreover, the tagging of variables and events in the current game log file of Portal 2, like that of many commercial games, changes across game levels and makes it extremely difficult to interpret and clean the log data for an automatic pattern discovery. In other words, the strategy and scope of data recording are not well aligned with the method and objective of stealth assessment when the assessment design occurs *after* game development.

A promising solution to the above challenge, as argued by Dede (2012) and Shute and Ventura (2013), is to interweave the evidence-centered assessment design and the learning context/task design in the early phase of game development. The E-Rebuild project illustrated that the development of domain-relevant competency, evidence, and task models should underlie the design and sequencing of tasks within and across game levels. The design of the game log file, in terms of the log's content, structure, and tagging, should be aligned with the game world design and evidence identification rules to enable automatic data cleaning and processing.

5.2 Implications for future game design and evaluation efforts

This chapter has focused on methods for achieving two interrelated goals that we believe can have a significant impact on both formal and informal learning. The first goal is to get more children, particularly females and certain underrepresented minorities (e.g., Black and Hispanic children), excited about and interested in developing STEM-related skills and knowledge—such as spatial skills and understanding ratios and proportional reasoning (which serve to undergird many higher math areas). Recognizing that interest alone is not enough, our second goal is to identify ways to facilitate and deepen learning in immersive, rich, and authentic environments. Well-designed digital games represent a promising vehicle for meeting both goals: capturing children's interest in STEM fields in general, and supporting their learning. More research is needed about the optimal design to be used for valid assessments and real-time learning support. We agree with the conclusion presented by Clark et al. (2011) that more research is needed that provides "supports for students to help them articulate their intuitive understandings from game play with the explicit formal concepts and representations of the discipline" (p. 2192). Our future research will focus on iterative design processes to refine the integration of stealth assessment and learning support in E-Rebuild. Data will be collected via both qualitative and quantitative methods over time to build up a body of evidence on the design generalizations and effectiveness of the learning game and its assessment/support mechanism.

REFERENCES

Baker, R. S. (2007, April). Modeling and understanding students' off-task behavior in intelligent tutoring systems. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 1059-1068). ACM.

Baker, R. S., Corbett, A. T., & Koedinger, K. R. (2004, January). Detecting student misuse of intelligent tutoring systems. In *Intelligent tutoring systems* (pp. 531-540). Springer Berlin Heidelberg.

Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *JEDM-Journal of Educational Data Mining*, *1*(1), 3-17.

Beck, J. E., & Mostow, J. (2008, January). How who should practice: Using learning decomposition to evaluate the efficacy of different types of practice for different types of students. In *Intelligent Tutoring Systems* (pp. 353-362). Springer Berlin Heidelberg.

Clark, D. B., Nelson, B., Chang, H., D'Angelo, C. M., Slack, K. & Martinez-Garza, M. (2011). Exploring Newtonian mechanics in a conceptually-integrated digital game: Comparison of learning and affective outcomes for students in Taiwan and the United States. *Computers and Education, 57*(3), 2178—2195.

Dede, C. (2012, May). Interweaving assessments into immersive authentic simulations: Design strategies for diagnostic and instructional insights. In *Invitational Research Symposium on Technology Enhanced Assessments.* Retrieved from http://www.k12center.org/rsc/pdf/session4-dede-paper-tea2012.pdf.

Dorval, M., & Pepin, M. (1986). Effect of playing a video game on a measure of spatial visualization. *Perceptual and Motor Skills, 62,* 159– 162.

Feng, J., Spence, I., & Pratt, J. (2007). Playing an action video game reduces gender differences in spatial cognition. *Psychological Science,18,* 850–855.

Green, C. S., & Bavelier, D. (2003). Action video game modifies visual selective attention. *Nature, 423* (6939), 534–537.

Green, C. S., & Bavlier, D. (2007). Action-video-game experience alters the spatial resolution of vision. *Psychological Science*, *18* (1), 88-94.

Hegarty, M., & Waller, D. (2004). A dissociation between mental rotation and perspective-taking spatial abilities. *Intelligence*, *32*, 175-191.

Hwang, G., Sung, H., Hung, C., Huang, I., & Tsai, C. (2012). Development of a personalized educational computer game based on students' learning styles. *Educational Technology Research and Development*, *60*(4, Special Issue on Personalized Learning), 623-638.

Ke, F., Shute, V., Erlebacher, G., Clark, K., & Ventura, M. (2014, June 9-10). *Earthquake Rebuild: Math learning through modeling and design.* Poster presented at Cyberlearning Summit 2014, Madison, WI.

Kickmeier-Rust, M. D., & Albert, D. (2010). Micro-adaptivity: protecting immersion in didactically adaptive digital educational games. *Journal of Computer Assisted Learning*, *26*(2), 95-105.

Leemkuil, H., & de Jong, T. (2012). Adaptive advice in learning with a computer-based knowledge management simulation game. *Academy of Management Learning & Education*, *11*(4), 653-665.

Levy, R. (2014). *Dynamic Bayesian Network Modeling of Game Based Diagnostic Assessments* (CRESST Report 837). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

MacCallum-Stewart, E. (2011). Stealth learning in online games. *Digital games and learning*, 107-128.

Miller, H. J., & Han, J. (Eds.). (2009). *Geographic data mining and knowledge discovery*. CRC Press.

Mislevy, R. J., Steinberg, L. S., Almond, R. G., & Lukas, J. F. (2006). Concepts, terminology, and basic models of evidence-centered design. *Automated scoring of complex tasks in computer-based testing*, 15-47.

Montello, D. R. (1993). Scale and multiple psychologies of space. In A. U. Frank, & I. Campari (Eds*.), Spatial information theory: A theoretical basis for GIS. Proceedings of COSIT '93. Lecture notes in Computer Science, vol. 716* (pp. 312– 321). Berlin: Springer-Verlag.9999

Montello, D. R., & Golledge, R. G. (1999). Scale and detail in the cognition of geographic information. *Report of the Specialist Meeting of Project Varenius, Santa Barbara, CA, May 14–16,1998.* Santa Barbara: University of California.

O'Rourke, E., Haimovitz, K., Ballweber, C., Dweck, C., & Popović, Z. (2014, April). Brain points: a growth mindset incentive structure boosts persistence in an educational game. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems* (pp. 3339-3348). ACM.

Prensky, M. (2001). Digital game-based learning. New York: McGraw-Hill.

Romero, C., Ventura, S., Pechenizkiy, M., & Baker, R. S. (Eds.). (2011). *Handbook of educational data mining*. CRC Press.

Shaffer, D. W., Hatfield, D., Svarovsky, G. N., Nash, P., Nulty, A., Bagley, E., Frank, K., Rupp, A. A., & Mislevy, R. (2009). Epistemic network analysis: A prototype for 21st-century assessment of learning, *International Journal of Learning and Media, 1*(2), 33-53.

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

Shute, V. J., Masduki, I., & Donmez, O. (2010). Conceptual framework for modeling, assessing, and supporting competencies within game environments. *Technology, Instruction, Cognition, and Learning, 8*(2), 137-161.

Shute, V. J., Ventura, M., & Ke, F. (in press). The power of play: The effects of Portal 2 and Lumosity on cognitive and noncognitive skills. Submitted to *Computers & Education* (May 9, 2014; revised July 19, 2014).

Shute, V. J., Ventura, M., & Kim, Y. J. (2013). Assessment and learning of qualitative physics in Newton's playground. *The Journal of Educational Research*, *106*(6), 423-430.

Siemens, G., & d Baker, R. S. (2012, April). Learning analytics and educational data mining: towards communication and collaboration. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 252-254). ACM.

SoLAR, (February 27–March 1, 2011), 1st International Conference on Learning Analytics and Knowledge, Banff, Alberta; as cited in George Siemens and Phil Long, "Penetrating the Fog: Analytics in Learning and Education," *EDUCAUSE Review,* vol. 46, no. 5 (September/October 2011).

Spence, I., Yu, J. J. J., Feng, J., & Marshman, J. (2009). Women match men when learning a spatial skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 1097–1103.

Uttal, D. H., Meadow, N. G., Tipton, E., Hand, L. L., Alden, A. R., Warren, C., & Newcombe, N. S. (2012). The malleability of spatial skills: A meta-analysis of training studies. *Psychological Bulletin.* Advance online publication. doi: 10.1037/a0028446

Vandenberg, S. G., & Kuse, A. R. (1978). Mental rotation, a group test of three-dimensional spatial visualization. *Perceptual and Motor Skills, 69,* 915-921.

Ventura, M., Shute, V. J., Wright, T. & Zhao, W. (2013). An investigation of the validity of the virtual spatial navigation assessment. *Frontiers in Psychology, 4,* 1-7.

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Wouters, P., & Van Oostendorp, H. (2013). A meta-analytic review of the role of instructional support in game-based learning. *Computers & Education*, *60*(1), 412-425.

Yairi, T., Kawahara, Y., Fujimaki, R., Sato, Y., & Machida, K. (2006, July). Telemetry-mining: a machine learning approach to anomaly detection and fault diagnosis for space systems. In *Space Mission Challenges for Information Technology, 2006. SMC-IT 2006. Second IEEE International Conference on*(pp. 8-pp). IEEE.

Zapata-Rivera, D., VanWinkle, W., Doyle, B., Buteux, A., & Bauer, M. (2009). Combining learning and assessment in assessment-based gaming environments: A case study from a New York City school. *Interactive Technology and Smart Education*, *6*(3), 173-188.

ACKNOWLEDGEMENTS

The research was supported by the National Science Foundation, Award Number #1318784 and the James T. and Catherine D. MacArthur Foundation #11-99517. We also would like to thanks Matthew Ventura for his work on the Portal 2 project--particularly relative to spatial skill assessment.

AUTHOR INFORMATION

Fengfeng Ke

Institutional affiliation: Florida State University
Institutional address: Tallahassee, Florida, 32306-4453

Complete mailing address:
Educational Psychology and Learning Systems
3205-C Stone Building
Tallahassee, FL 32306-4453
The Florida State University

Telephone number: 850.644.8794
Fax number: 850.644.8776
Email address: fke@fsu.edu

Fengfeng Ke is an associate professor of education in the department of Educational Psychology and Learning Systems at the Florida State University where she works in the areas of game-based learning, virtual reality, computer-supported collaborative learning, and inclusive design of computer-assisted learning. She has published widely in the fields of innovative learning technologies and inclusive pedagogy for e-learning.

Valerie Shute

Institutional affiliation: Florida State University
Institutional address: Tallahassee, Florida, 32306-4453

Complete mailing address:
Educational Psychology and Learning Systems
3205-G Stone Building

Tallahassee, FL 32306-4453
The Florida State University

Telephone number: 850.644.8785
Fax number: 850.644.8776
Email address: vshute@fsu.edu
Short biographical sketch (200 words maximum):

Val Shute is the Mack & Effie Campbell Tyner Endowed Professor in Education in the Department of Educational Psychology and Learning Systems at Florida State University. Her current research involves using games with stealth assessment to support learning—of cognitive and non-cognitive knowledge, skills, and dispositions. Her research has resulted in numerous grants, journal articles, books, chapters in edited books, a patent, and a couple of recent books (e.g., Shute & Ventura (2013), Measuring and supporting learning in games: Stealth assessment, The MIT Press; and Shute & Becker (Eds.) (2010), Innovative assessment for the 21st century: Supporting educational needs, Springer-Verlag).