Je-LKS
PEER REVIEWED PAPERS
GAMIFICATION AND SERIOUS GAME FOR LEARNING
Vol. 11, n.3, 2015
ISSN: 1826-6223 | eISSN: 1971-8829
Journal of e-Learning and Knowledge Society
The Italian e-Learning Association Journal

EMPLOYING GAME ANALYTICS TECHNIQUES IN THE PSYCHOMETRIC MEASUREMENT OF GAME-BASED ASSESSMENTS WITH DYNAMIC CONTENT

Yasser Jaffal
Dieter Wloka

Department of Technical Computer Science, Universität Kassel
yjaffal@inf.e-technik.uni-kassel.de; dwloka@uni-kassel.de

Keywords: Game-based assessment, psychometric measurement, serious games, simulation, game analytics.

The adaptation Game-Based Assessment (GBA) (Mislevy *et al.*, 2014) has been growing in the last years backed by video games' capability of offering a task model to assess learners' complex knowledge. Since the variables generated from such performances are not directly interpretable, assessment frameworks such as Evidence-Centred Design (ECD) came into play (Mislevy & Almond, 2003). In this work we show our initial findings when using game analytics techniques (Saif El-Nasr *et al.*, 2013) such as play metrics to analyse players' performance in an open world 3D game for traffic education. The results show that play metrics can be used in cases where game has a dynamic user-generated content of unknown structure. Additionally, we discuss how these metrics can form the basis of measuring psychometric principles that ECD uses to evaluate assessments, which are validity, reliability, comparability and fairness (Mislevy & Wilson, 2003).

for citations:
Jaffal Y., Wloka D. (2015), *Employing game analytics techniques in the psychometric measurement of game-based assessments with dynamic content* , Journal of e-Learning and Knowledge Society, v.11, n.3, 101-115. ISSN: 1826-6223, e-ISSN:1971-8829

1 Introduction

Game-based assessment depends on psychometric models such as evidence-centred design ECD, because of their ability to cope with complex performance variables produced by player-game interaction. ECD consists of three basic models: the student model, the task model and the evidence model (Mislevy & Andreas, 2003). The fourth (presentation) model gains additional importance in GBA; because of the nature of games tasks, which usually ask assessment subject to do more than selecting the correct answer from given choices. However, this model is not widely implemented.

The student model of ECD consists of variables that describe specific skills or knowledge measured by the assessment. These can be as simple as the ability to perform addition and subtraction on fractions (Kerr, 2014) or complex 21st century skills like problem solving (Clarke-Midura & Dede, 2010). On the other hand, the task model describes concrete situations and tasks used to measure and report the performance of the assessment subject. Finally, the evidence model works as a bridge between student and task models by specifying how the outcomes of the task model (known as the *observable variables*) can update our beliefs about student knowledge or skills specified in the student model.

GBA can benefit from ECD model by interpreting the variables of student model as test scores or evaluation measurements. Therefore, a stronger relation between what students do in the task model and the inferences about their performance in the student model results in a valid game that can be adopted as an assessment tool without jeopardizing the quality of the assessment process.

To measure the performance of the assessment subject in the task model, GBA uses game *play logs* also known as *click streams* (Halverson & Owen, 2014). These logs record every action the player takes during play sessions, making it possible to track the change in the game state based on the decisions he/she takes. These actions can be starting the game, asking for help, providing correct answer, or, in more complex simulations, taking good, bad or neutral decisions while dealing with an accident (Iseli *el al.*, 2010).

Logging player's actions is not exclusive for GBA, since it has been used in the industry to give game designers and producers insights about in-game player behaviour during alpha or beta testing. For example, (Drachen & Canossa, 2009) show how Geographic Information Systems (GIS) where used within the virtual game world of *Tomb Raider: Underworld* to analyse game maps and identify locations where players die frequently and the reasons of the death. Another research (Canossa & Drachen, 2009) used the same game to identify classes of players based on the frequency of interesting actions such as weapon usage, puzzle solving and navigation speed.

Since game analytics have emerged within commercial game industry, tech-

niques used by this type of data analyses are more game oriented and focus mainly on the game as a product. On the other hand, educational analyses used in GBA handle players' data subjectively in order to draw inferences about their learning behaviour. Moreover, game analytics must handle more complex game data that commercial games with numerous interacting mechanics produce; while educational games usually have simple point-and-click mechanics as well as dialogues (see Buckly *et al.*, 2010; Buschang *et al.*, 2012; Gobert *et al.*, 2012 and Kim & Chung, 2012 for example).

Data produced by the simple mechanics and the limited *decision space* (Klein *et al.*, 2009) given to the player have stronger indication of knowledge, given that they are collected from well-crafted static game content set by domain experts to serve as task model. However, drawing inferences from games with open world set-up granting navigation/interaction freedom for the players to achieve their goals would be more challenging.

In this study we address two main research questions:

1. How valid are real-time 3D games as game-based assessment tools? How can evaluate them in terms of validity, reliability, comparability and fairness?
2. Can play metrics play the rule of observable variables in ECD to aid the automatic scoring in real-time 3D GBA?

We investigate these questions by analysing play data collected from real players as they play our game called *Bicycle World*.

2 Literature Review

GBA is an interestingly growing field, and more research points to the potential of games in assessing advanced skills while saving efforts and costs without compromising assessment quality (Scardamalia *et al.*, 2012; Csapó *et al.*, 2012).

ECD and psychometric data analysis has been widely adopted in game- and simulation- based assessments (deKlerk *et al.*, 2015). A number of games focused on science subjects such as biology and chemistry. *BioLogica* (Buckley *et al.*, 2010) features a set of mini-puzzles to teach about cells and their formation of organisms. *Science Assitment* (Gobert *et al.*, 2012) allows players to hypothesise by filling blanks in a dialogue box with drop down lists then provides them with a virtual lab or an exploration environment where they can practice data collection and test their hypothesis. *Progenitor X* (Halverson & Owen, 2014) employs a story to motivate the player to solve 2D puzzles based on stem cells manipulation methods. A 3D biomedical virtual lab was featured in *Mission Biotech* (Lamb *et al.*, 2014) where players explore the devices of

the lab, do experiments, and interact with NPCs. Interactive assessments using multiple choice questions, click-and-drag matching, and interactive ecosystem simulations were implemented in (Quellmalz *et al.*, 2010). *SimScientists* features interactive experiments for teaching the physics of atomic molecules by selecting materials to add and interaction conditions then running interaction simulation (Quellmalz *et al.*, 2010 and Quellmalz *et al.*, 2013). *True Roots* motivates students to study genetics by using a story of a lost girl trying to identify her parents (Stevens & Casillas, 2006). A similar approach was used in *Hazmat* where players are motivated to investigate and identify chemical materials by performing simulated experiments (Stevens *et al.*, 2004).

In physics, *Newton's Playground* offers the player with an innovative play style in which he/she draws shapes that would then have mass and interact with each other. This game has been studied in (Shute *et al.*, 2013) to explore its validity as a tool for physics assessment. Even the game provided some usage data in log files, the study exploited them only partially and adopted the classical *black box* method (pre- and post-tests).

In addition to science, maths has also been focused on by several GBAs employing ECD. *Save Patch* teaches students mathematical operations of fractions by using pieces of ropes with partial lengths to construct a complete path that game character can walk over to reach its goal (Buschang *et al.*, 2012; Kerr, 2014 and Levy, 2014). *Math Garden* (Klinkenberg *et al.*, 2011) uses gamified context with virtual coins, achievements and badges to motivate students to solve mathematical problems on the game's website.

Analysis of medical cases was the focus of other games and simulations. Interactive dialogue windows were used to simulate the examination of medical cases by physicians in (Margolis & Clauser, 2006). Dental hygiene procedures were implemented in *Dental Interactive* for the purpose of simulation-based formal practice examination (Mislevy *et al.*, 2002).

Serious games have been used for long time in the military field. For example, (Iseli *et al.*, 2010) use ECD along with experts assessment criteria to develop a Bayesian network for automatic scoring in *Navy Damage Control*, a 3D game that asks the player to deal with fire accidents on board, which is done through dialogue boxes.

Among the fields that benefited from educational games and game-based assessment are 21st century skills such as creative problem solving and critical thinking. In *SimCityEDU: Pollution Challenge*, the player assumes the role of a city mayor and has to grow the city economically while causing minimal possible pollution (Mislevy *et al.*, 2014). (Shute *et al.*, 2009) show an example of how commercial games such as *Oblivion* can be used to assess creative problem solving skills by observing player's approach in completing quests. However, neither real data were collected nor practical implementation details

about player's actions logging and scoring were given. Similarly, *World of Goo* was used with screen recording software and human rating to perform an exploratory study to assess problem solving, causal reasoning and static equilibrium skills (Shute & Kim, 2011). In *Taiga Park* (Shute, 2011), data collection by interviewing NPCs and collecting water samples is the main task of the player before he/she draws conclusions about what causes pollution in the park.

3 Case Study

This section discusses the game used in this study, the test sample involved and the data collection procedure followed.

3.1 Bicycle World Game

In this study we used an in-house developed real time 3D game called *Bicycle World*. The game targets 3rd and 4th grade students in Germany; who should, by this age, learn the necessary bicycle traffic rules. The player is given a virtual bicycle observed from first-person view, which can be used to freely navigate in each level and visit predefined places in it (player objectives). Figure 1 shows a screen shot of the game.

Fig. 1 - Screen shot of Bicycle World

All game levels where designed using a custom editor, which also allows users to design their own maps with any structure they need: crossings of different sizes and priorities, narrow and wide streets or curves (visible or hidden by buildings).

Once the player starts playing a specific map, the game begins to track the position and the rotation of the bicycle every 500ms, as well as the rotation of the camera (view direction). The player is not forced to look forward all the time, as he/she can use the mouse to look right or left while driving as shown in Fig. 1. The game tracks also the position and the rotation of all other vehicles controlled by NPCs. Moreover, special events such as accidents, failing to

yield for vehicles coming from higher priority streets, unnecessary stops and driving in the wrong direction are also recorded. The player ID is recorded for each event to guarantee subjective data collection, as well as the time and the location of each event.

3.2 Test Design and Sample

The study was conducted with a sample size of 109 players who played the game for about 45 minutes. 65 (59.6%) of the players where males and 44 (40.4%) were females. Each player was asked to complete 9 scenes: 5 practices, 2 tests with hints, and 2 tests without hints. All players were asked to play the same 5 practices, while each one of the 4 test scenes had two variants: A and B. Players were randomly assigned to group A (N_A=56, 51.4%) or B (N_B=53, 48.6%).

Players come from different backgrounds and play the game for different purposes. Therefore, the sample can be divided into two sub samples that represent categories of players with different characteristics. The first category *Students* ($N_{Students}$=91, 83.5%) represents 4th grade students of two elementary schools with average age of 10 years. The other category *Refugees* ($N_{Refugees}$=18, 16.5%) consisted of foreign refugees (mainly Africans) who played the game as a part of an integration course in order to get familiar with German traffic rules. The average age of this group is 26.

Student tests were performed in two elementary schools in Kassel area. Even these students come from different ethnic and cultural backgrounds, they are all students in German schools and already familiar with the language and the local community. As for refugees, they reside in the small town of Frankenberg (southern to Kassel), and have performed their tests in that town. The municipality of Frankenberg and a public school there cooperated with us in organising these test sessions.

3.3 Data Collection and Pre-Processing

Data used in this analysis are *telemetry data* (Kim *et al.*, 2008; Isbister & Schaffer, 2008) collected remotely over the Internet and stored in a central database. Bicycle World runs on client machines with Windows operating system, and sends the telemetry data over the internet. Collected data is buffered locally on client side and sent once every 10 seconds to the server. Buffering is done on the hard drive to allow the data to be re-sent in case of client crash or connection problems between the client and the server.

All data collected during play sessions are stored in a form of events. For example, player position, rotation, looking direction, accidents, violations of

traffic rules and so forth are all events and each one of them is stored in a single record. To analyse the data, we aggregated all events of a single play session in one record, and generated a table with 59 attributes that summarises the most important features of each session. Only subset of these attributes were used in this initial work, namely a) session length, b) total distance the player travelled, c) total number of accidents and traffic violations committed by the player, d) number of presses on each control key, and e) time spent looking at different directions.

For this initial study, a total number of 1376 play sessions were recorded. Outliers caused by technical difficulties (too long sessions or too many accidents and mistakes) were removed. The total number of these outliers was 100. Since game maps vary in length and difficulty, outlier detection was performed for each map separately using interquartile range.

4 Results

GBA is a multi-stage process that begins with collecting data about player's performance in the task model of ECD. These data need to be validated before they can be interpreted as performance scores that update student model and, eventually, generate test scores. The focus of this research was this initial step, which is the psychometric evaluation of task model's data as performance scores.

This section shows our initial findings regarding the use of game analytics techniques in measuring validity, reliability, comparability and fairness of our game-based assessment.

4.1 Measuring Validity

One of the concerns of the validity is the possible alternative explanations of the measured performance. In our case the main concerns were purely technical and relate to the way the players use the software.

To measure the effect of possible sources of inaccuracy, we formulated two groups of generic metrics using the variables described in section 3.3. The first group describes player's input behaviour in terms of key-press frequency and mouse rotations, in addition to technical details like frame rate. Additionally, outlier sessions were counted per player to generate a *difficulty rank* for that player. We expected that players who are less experienced in dealing with computers and game controls would produce significantly longer sessions with significantly higher number of mistakes, which results in their session being detected as outliers, and eventually have higher difficulty rank.

The second group of metrics (performance metrics) summarises the mi-

stakes the player made per Kilometre in all sessions he/she played. These mistakes include accidents committed, situations in which the player had to stop but did not, situations where the player unnecessarily stopped, and situations where the player drove in the opposite direction. Table 1 shows the measured correlations among the metrics from both groups. These four metrics will be used throughout this paper as the main performance indicators for the players.

The purpose for choosing such simple metrics is their validity for summative scoring of player's performance. Since this study explores the validity of the game as an assessment tool, we had to use performance measurements that are directly interpretable. Generally, players playing the game and travelling the same distance while making fewer mistakes are showing better performance.

Table 1
CORRELATIONS BETWEEN INPUT AND PERFORMANCE METRICS ($p < 0.05$)

Input Metrics	Accidents / KM	Did not stop / KM	Unnecessary stops / KM	Opposite direction / KM
Difficulty rank	0.008	0.161	0.332*	0.270*
Frame rate	0.021	-0.016	0.304*	0.389*
Forward key / KM	0.080	0.113	-0.024	-0.022
Backward key / KM	0.122	-0.020	-0.146	-0.144
Turn keys / KM	0.023	-0.044	-0.124	-0.079
Brakes key / KM	0.160	-0.120	0.608*	0.278*
Look back ratio	-0.070	-0.132	0.259*	0.221*

Apparently, basic mistakes such as unnecessary stops and driving in the opposite direction have significant correlations with computed difficulty rank, in addition to another basic input mistakes such as failure to set looking direction to the front and excessive use of the brakes. Interestingly, a significant positive correlation was found between frame rate and basic mistakes; which suggests that players who performed these proficiency-related mistakes were among those who played the game on better machines with high frame rates. This fact minimises the impact of PC used on the measured performance.

4.2 Measuring Reliability

Reliability is concerned with possible variation in measured variables when the measurement is repeated several times. Since reliability is sensitive to the number of repetitions, two additional data pre-processing steps were necessary before measuring it. The first step was to remove repetitions (sessions where the same player replayed the same map). This step was done by keeping the last replay of each map for each player. In case the player did not complete the

map in the last replay, the first earlier replay in which the map was completed is taken. This method tries to get the best performance in terms of experience with game controls and probability to go through all situations in the map.

The second step was to keep only players who played the latest 6 maps (2 practices and 4 tests), in order to eliminate players with insufficient repetition of the situations, as well as eliminating overly simple tutorials that contain no traffic situations at all. Given that the latter 4 tests vary between groups A and B, measured reliability would be affected by the comparability (section 4.3) among each map pair.

The subset used to measure reliability consisted of 40 players and 240 sessions. For each player, performance metrics where generated from his/her average performance in the two practice maps and the four test maps separately; then these metrics were normalised using z-transformation. The total number of metrics was 8, which are the same metrics in the columns of Table 1. Each one of these metrics was repeated twice: one to represent the value measured from practice maps and one for test maps.

Each metric was used as question and the values of the metrics were taken as answers. Having this question-answer model with 40 participants and 8 questions allowed us to measure reliability using Cronbach's Alpha (Cronbach, 1951). The measured Cronbach's Alpha was 0.71, which is satisfactory for our study.

4.3 Measuring Comparability

Comparability is concerned with how our observations differ when they originate from different sources or the same sources under different circumstances. To measure the comparability of the maps, we used a brute-force algorithm that scans each map and reports the count of maximum possible occurrences for each known traffic situation. Since these situations are the building blocks of the task model and eventually for the scores, maps with coherent number of situations of each type must be interchangeable.

To statistically measure the comparability of players' performance, we computed the means of the four performance metrics from all sessions played by the members of each play group.

Figure 2 depicts performance metrics for play groups A and B. The chart shows no significant difference between the groups in terms of mistakes performed per KM in normal situation. However, outlier analysis showed that 59 outlier sessions came from group A versus 41 only from group B. Further investigation revealed that variant A of the second test map had 13 outliers versus 7 only for variant B, which is not the case for the other three pairs who had their outliers almost evenly distributed between both play groups. Manual analysis

of these sessions revealed that the entrance to the pedestrians' area, where players had to drive, was not clearly notable for some players, causing them to continue driving on the street hence losing their way in the map. However, this was a problem for only 12 players of group A; which had a total of 56 players.

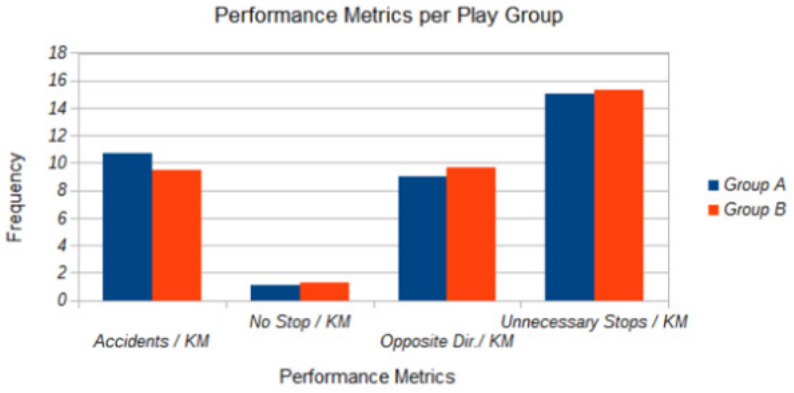

Fig. 2 - Comparing performance metrics between the two play groups

4.4 Measuring Fairness

Fairness considers differences among participants in terms of educational, social and other backgrounds. In this exploratory study, we investigated fairness in two dimensions: gender and cultural background. Since fairness focuses on play patterns regardless of scores, we compared participants on both dimensions using input metrics rather than performance metrics. For each comparison, we computed the average count of key-presses per KM, then normalised these counts using range transformation to be between 0 and 1.

The results are shown in Figure 3 (for gender) and Figure 4 (for cultural background). In general, Figure 3 shows that males keep looking forward for longer time. However, it is not possible to conclude whether this indicates better or worse performance. The point of interest here is the observable gender-dependant play-style that is originated from advanced mechanic of the game, which could result in different test conditions based on whether the player is male or female. However, gamer vs. non-gamer player can also be considered; since it is usually gender dependant (males are more likely to be PC gamers than females). On the other hand, no significant differences in outlier count were found between males and females.

Figure 4 shows that refugees tend to use keys 2 to 3 times less than students, which indicated largely different use pattern that affects test fairness. They

also spend longer time not looking forward, which might indicate usage difficulties. To count for considerable difference between the number of students and the number of refugees participated, the comparison was repeated 5 times by randomly sampling (without replacement) students equal to the count of refugees, then the average of these 5 samples was taken and compared again with refugees. Even with sampling, the differences remained almost the same.

In terms of outliers, 30% of outlier sessions came from refugees who represent 16.5% of the sample, which is an additional concern regarding test fairness.

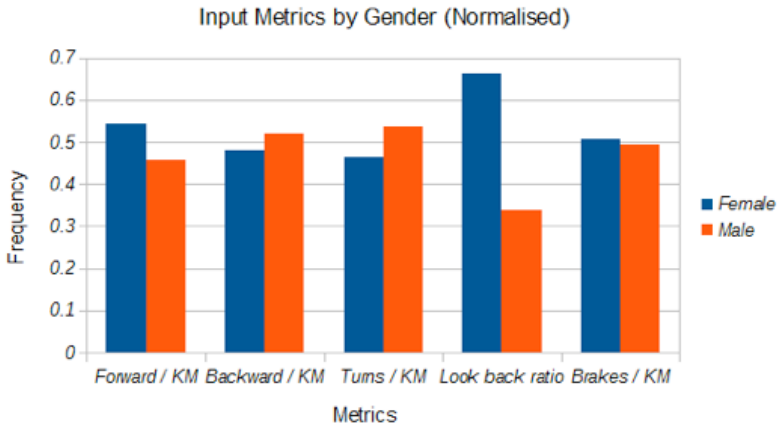

Fig. 3 - Comparing input metrics between male and female participants

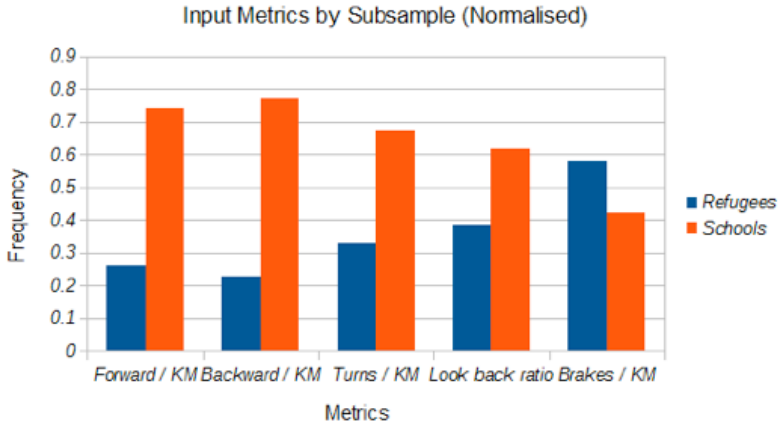

Fig. 4 Comparing input metrics between different sub-samples

Conclusions

Play metrics, and game analytics in general, have potential to serve as observable variables in ECD for game based assessments. In this study we showed how play metrics of Bicycle World players were used to describe and score players' performance. Play metrics were also useful in detecting outliers which represent misuse of the game software. We showed how these metrics were used to compute the psychometric measurements of Bicycle World as game-based assessment. Correlating input metrics with performance metrics showed concerns about the validity of Bicycle World as assessment tool, since problematic sessions (outliers) and excessive use of inappropriate controls had significant correlations with performance metrics that describe basic mistakes. For players who use the software moderately, Bicycle World has good reliability (α=0.71) as assessment tool. Traffic situations, which are the building blocks of the scene, were found to be suitable for computing comparability among different maps; even they cannot detect design pitfalls that are revealed by play testing. Finally, Bicycle World had an input scheme and mechanics that are relatively complex from inexperienced players' perspective, which highly affects its fairness.

The next step would be generating more complex play metrics that describe players' performance at each traffic situation and use these metrics for automatic scoring of the player. These metrics will depend on spatio-temporal game analytics, specifically player trajectory analysis.

REFERENCES

Buckley B., Janice G., Horwitz P. & O'Dwyer L. (2010), *Looking inside the black box: assessing model-based learning and inquiry in BioLogica™*, International Journal of Learning Technology, 5 (2), 166-190.

Buschang R., Kerr D. & Chung G. (2012), *Examining feedback in an instructional video game using process data and error analysis*, in: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Report 817.

Canossa A. & Drachen A. (2009), *Patterns of play: Play-personas in user-centred game development*, in: Breaking New Ground: Innovation in Games, Play, Practice and Theory, London, Brunel University.

Clarke-Midura J. & Dede C. (2010), *Assessment, technology, and change*, Journal of Research on Technology in Education, 42 (3), 309-328.

Cronbach L. (1951), *Coefficient alpha and the internal structure of tests*, psychometrika, 16 (3), 297-334.

Csapó B., Ainley J., Bennett R. E., Latour T. & Law, N. (2012), *Technological issues*

for computer-based assessment, Assessment and teaching of 21st century skills, 143-230.

de Klerk S., Bernard V. & Eggen T. (2015), *Psychometric analysis of the performance data of simulation-based assessment: A systematic review and a Bayesian network example*, Computers & education, 85, 23-34.

Drachen A. & Canossa A. (2009), *Analyzing spatial user behavior in computer games using geographic information systems*, in: Proceedings of the 13th International MindTrek Conference: Everyday Life in the Ubiquitous Era. 182-189, ACM.

Gobert J., Sao Pedro M., Baker R., Toto E. & Montalvo O. (2012), *Leveraging educational data mining for real-time performance assessment of scientific inquiry skills within microworlds*, JEDM-Journal of Educational Data Mining, 4 (1), 11-143.

Halverson R. & Owen E. (2014), *Game-based assessment: an integrated model for capturing evidence of learning in play*, International Journal of Learning Technology, 9 (2), 111-138.

Isbister K. & Schaffer N., eds (2008), *Game usability: Advancing the player experience*, CRC Press.

Iseli M., Koenig A., Lee J. & Wainess R. (2010), *Automated Assessment of Complex Task Performance in Games and Simulations*, in: Proceedings of the Interservice/ Industry Training, Simulation and Education Conference, Orlando, FL.

Kerr D. (2014), *Into the Black Box: Using Data Mining of In-Game Actions to Draw Inferences from Educational Technology about Students' Math Knowledge*, Ph. D. Thesis, University of California.

Kim J. & Chung G. (2012), *Use of a Survival Analysis Technique in Understanding Game Performance in Instructional Games*, in: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Report 812.

Kim H., Gunn D., Schuh E., Phillips B., Pagulayan R. & Wixon D. (2008), *Tracking real-time user experience (TRUE): a comprehensive instrumentation solution for complex systems*, in: Proceedings of the SIGCHI conference on Human Factors in Computing Systems. 443-452, ACM.

Klein G., Pfaff M. & Drury J. (2009), *Supporting a Robust Decision Space, in: AAAI Spring Symposium*: Technosocial Predictive Analytics. 66-71, Stanford University.

Klinkenberg, S., Straatemeier M. & Van der Maas H. (2011), *Computer adaptive practice of maths ability using a new item response model for on the fly ability and difficulty estimation*, Computers & Education, 57 (2), 1813-1824.

Lamb L., Annetta L., Vallett D. & Sadler T. (2014), *Cognitive diagnostic like approaches using neural-network analysis of serious educational video games*, Computers & Education, 70, 92-104.

Levy R. (2014), *Dynamic Bayesian network modeling of game based diagnostic assessments*, in: CRESST Conference: Warp Speed, Mr. Sulu: Integrating Games, Technology, and Assessment to Accelerate Learning in the 21st Century, Redondo Beach, CA.

Levy R. (2013), *Psychometric and evidentiary advances, opportunities, and challenges*

for simulation-based assessment, Educational Assessment, 18 (3), 182-207.

Margolis M. & Clauser B. (2006), *A regression-based procedure for automated scoring of a complex medical performance assessment*. Mislevy R.,Willamson D. and Bejar, I. eds (2006), Automated Scoring of Complex Tasks in Computer Based Testing, Lawrence Erlbaum, Mahwah, NJ.

Mislevy R., Wilson M., Ercikan K. & Chudowsky N. eds (2003), *Psychometric principles in student assessment*, Springer Netherlands.

Mislevy R., Oranje A., Bauer M., von Davier A., Hao J., Corrigan S., Hoffman E., DiCerbo K. & John M. (2014), *Psychometric considerations in game-based assessment*, in: GlassLab Report.

Mislevy R., Steinberg L., Breyer J., Almond R. & Johnson L. (2002), *Making sense of data from complex assessments*, Applied Measurement in Education, 15 (4), 363-389.

Mislevy R., Almond R. & Lukas J. (2003), *A brief introduction to evidence-centered design*, ETS Research Report Series, 3, 1-29.

Quellmalz E., Davenport J., Timms M., DeBoer G., Jordan K., Huang C. & Buckley B. (2013), *Next-generation environments for assessing and promoting complex science learning*, Journal of Educational Psychology, 105 (4), 1100.

Quellmalz E., Timms M. & Buckley B. (2010), *The promise of simulation-based science assessment: The Calipers project*, International Journal of Learning Technology, 5 (3), 243-263.

Quellmalz E., Timms M., Silberglitt M. & Buckley B. (2012), *Science assessments for all: Integrating science simulations into balanced state science assessment systems*, Journal of Research in Science Teaching, 49 (3), 363-393.

Saif El-Nasr, M., Drachen A. & Canossa A., eds (2013), *Game analytics: Maximizing the value of player data*, Springer Science & Business Media.

Scardamalia M., Bransford J., Kozma B. & Quellmalz E. (2012), *New assessments and environments for knowledge building*, in: Assessment and teaching of 21st century skills. 231-300, Springer Netherlands.

Shute V. (2011), *Stealth assessment in computer-based games to support learning*, Computer games and instruction, 55 (2), 503-524.

Shute V. & Kim Y. J. (2011), *Does playing the World of Goo facilitate learning*. Yun Dai D. eds (2012), Design research on learning and thinking in educational settings: Enhancing intellectual growth and functioning, Routledge, New York.

Shute V., 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.

Shute V., Ventura M., Bauer M. & Zapata-Rivera D. (2009), *Melding the power of serious games and embedded assessment to monitor and foster learning*. Ritterfeld U., Cody M. & Vorderer P. eds (2009), Serious games: Mechanisms and effects, Routledge, New York.

Stevens, R. & Casillas A. (2006), *Artificial neural networks*. Mislevy R.,Willamson D. and Bejar, I. eds (2006), Automated Scoring of Complex Tasks in Computer Based

Testing, Lawrence Erlbaum, Mahwah, NJ.
Stevens R., Soller A., Cooper M. & Sprang, M. (2004), *Modeling the development of problem solving skills in chemistry with a web-based tutor*, in: Intelligent tutoring systems. 580-591, Springer Berlin Heidelberg.