Assessment in and of serious games. An overview

Francesco Bellotti, Dep.t of Naval, Electric, Electronic and Telecommunications Engineering, University of Genoa; franz@elios.unige.it

Bill Kapralos, Faculty of Business and Information Technology, University of Ontario Institute of Technology, Oshawa, Canada; bill.kapralos@uoit.ca

Kiju Lee, Department of Mechanical and Aerospace Engineering, Case Western Reserve University, Cleveland, OH, USA; kiju.lee@case.edu

Pablo Moreno-Ger, Faculty of Computer Science, Universidad Complutense de Madrid, Spain; pablom@fdi.ucm.es

Riccardo Berta, Dep.t of Naval, Electric, Electronic and Telecommunications Engineering, University of Genoa; berta@elios.unige.it

Abstract

There is a consensus that serious games have a significant potential as a new means to support and enhance the quality of education. However, their effectiveness in terms of learning outcomes is still understudied mainly due to the complexity involved in assessing intangible measures. A systematic approach - based on established principles and guidelines – is necessary to enhance the design of serious games based on established principles and guidelines, because many studies on serious games lack a rigorous assessment, and this is hindering their deployment. An important aspect in the evaluation of serious games, like other educational tools, is user performance assessment. This is an important area of exploration because serious games are intended to evaluate the learning progress as well as the outcomes. This also emphasizes the importance of providing appropriate feedback to the player. Moreover, performance assessment enables adaptivity and personalization to meet individual needs in various aspects, such as learning styles, information provision rates, feedback, etc. This paper first reviews related literature regarding the educational effectiveness of serious games. It then discusses how to assess the learning impact of serious games and methods for competence and skill assessment. Finally, it suggests two major directions for future research: characterization of the player's activity and better integration of assessment in games.

1. Introduction

Serious games are designed to have an impact on the target audience, which is beyond the pure entertainment aspect [37, 39]. One of the most important application domains is in the field of education given the acknowledged potential of serious games to meet the current need for educational enhancement [27, 90].

In this field, the purpose of a serious game is twofold: i) to be fun and entertaining, and ii) to be educational. A serious game is thus designed both to be attractive and appealing to a broad target audience, similar to commercial games, and to meet specific educational goals as well. Therefore, assessment of a serious game must consider both aspects of fun/enjoyment and educational impact.

In addition to considering fun and engagement, thus, serious games' assessment presents additional unique challenges, because learning is the primary goal. Therefore, there is also a need to explore how to evaluate the learning outcomes to identify which serious games are most suited for a given goal or domain, and how to design more effective serious games (e.g., what mechanics are most suited for a given pedagogical goal, etc.). In this sense, the evaluation of serious games should also cover player performance assessment. Performance assessment is important because serious games are designed to support knowledge acquisition and/or skill development. Thus, their underlying system must be able to evaluate the learning progress, since the rewards and the advancement in the game have to be carefully bound to it. This also stresses the importance of feedback to be consequentially provided to the player. Moreover, performance assessment enables adaptability and personalization in various aspects, such as, for instance, definition, presentation and scheduling of the contents to be provided to the player.

In summary, this paper intends to provide an overview of the two major aspects of assessment that concern serious games: i) evaluation of serious games, and ii) evaluation of player performance in serious games. The remainder of this paper is organized as it follows: Section 2 presents a literature review regarding the educational effectiveness of serious games. Section 3 discusses how to assess a serious game's learning impact. Section 4 reviews methods for competence and skill assessment; and Section 5 focuses on in-process assessment, which appears to be well suited for games. Concluding remarks and suggested directions for future research are given in section 6.

2. General Context

Despite the a widespread consensus about the educational potential of video games, there is a shortage of studies that have methodically examined (assessed) learning via gameplay whether considering "entertainment" games or serious games, prompting some to challenge the usefulness of game-based learning (e.g., [16, 84]).

A number of studies that have questioned the effectiveness of game-based learning (see for example, [38, 45, 62]). However, many of those reviews were conducted several years ago and even in the last 10 years, there has been unprecedented development within the videogame field in general and educational games in particular. In contrast, more recently, Blunt [12] gathered evidence from three studies that had unquestionably achieved significantly better test results with students that had learned using games, compared to control groups who received typical instruction.

Furthermore, one cannot ignore the fact that simulations and serious games are a promising means for safely and cost-effectively acquiring skills and attitudes which are hard to get by rote learning [37] and that learning via gameplay may be longer lasting [11]. In addition, there are many examples of studies that have demonstrated that properly designed "learning games" - some examples are provided below - do produce learning, while engaging players [77].

One of the foundational reviews of the effectiveness of gaming was performed by Livingston et al. [65], when they evaluated seven years of research and over 150 studies to examine the effectiveness of gaming. Their results were later on mirrored by Chin et al. [18], and they concluded that "simulation games" are able to teach factual information although they are not necessarily more effective than other methods of instruction [18, 65]. However, it was observed that students preferred games and simulations over other classroom activities and participation in such "gamed simulations" can lead to changes in their the attitudes including attitudes toward education, career, marriage, and children although these effects could be short lived [18, 65].

More recently, Connolly et al. [22] have made an extensive literature study on computer games and serious games, identifying 129 papers reporting empirical evidence about the impacts and outcomes of games with respect to a variety of learning goals, including a critique of those cases where the research methods were not adequate. The findings revealed, however, that playing computer games is linked to a range of perceptual, cognitive, behavioural, affective and motivational impacts and outcomes. The most frequently occurring outcomes and impacts were knowledge acquisition/content understanding and affective and motivational outcomes. Despite the diffused perception that games might be especially useful in promoting higher order thinking and soft and social skills, the literature review provides limited evidence for this, also given the lack of adequate measurement tools for such skills.

There are specific fields were a significant impact from serious gaming have been found. One of the most relevant domains is healthcare, with different experiences that have provided positive results. The effectiveness of virtual reality and games in the treatment of phobias and in distracting patients in the process of burn treatment or chemotherapy has been scientifically validated with the use of functional Magnetic Resonance Imaging (fMRI) which has shown differences in brain activity in patients who were experiencing pain with and without the use of virtual reality and games [11]. An experiment with Re-Mission (a video game developed for adolescents and young adults with cancer), showed that the video-game intervention significantly improved treatment adherence and indicators of cancer-related self-efficacy and knowledge in adolescents and young adults who were undergoing cancer therapy [57]. More recently, Cole and Knutson [21] showed that activation of brain circuits involved in positive motivation during Re-Mission game play appears to be a key ingredient in influencing positive health behavior. Regarding behavioural change, the serious game *The Matrix*, developed to enhance self-esteem, was subject to rigorous scientific evaluation and was shown to increase self-esteem through classical conditioning [26].

Bellotti et al. [9] discuss the results of a lab user test aimed at verifying knowledge acquisition through minigames dedicated to cultural heritage. The implemented minigames were particularly suited for supporting image studying, which can be explained by the visual nature of games. Compared to text reading, the games seem to more strongly force the player to focus on problems, which favors knowledge acquisition and retention.

The above mentioned results show that serious games can be an effective tool to complement the educational instruments available to teachers, in particular for spurring user motivation [7] and for achieving learning goals at the lower levels in the Bloom's taxonomy [22]. The next section is dedicated to analyzing methods for assessing a serious game's learning impact.

3. Assessing a serious game's effectiveness

Learning with serious games remains a goal-directed process aimed at clearly defined and measurable achievements and therefore, must implement assessments to provide an indication of the learning progress and outcomes to both the learner and instructor [10] or as Michael and Chen [70] state "Serious games like every other tool of education, must be able to show that the necessary learning has occurred". For serious games to be considered a viable educational tool, they must provide some means of testing and progress tracking and the testing must be recognizable within the context of the education or training they are attempting to impart.

Assessment describes the process of using data to demonstrate that stated learning goals and objectives are actually being met [18]. Assessment is a complement to purpose, and it is commonly employed by learning institutions, regardless the teaching methods used, whether or not their students actually learn [38]. However, learning is a complex construct making it difficult to measure and determining whether a simulation or serious game is effective at achieving the intended learning goals is a complex, time consuming, expensive, and difficult process [30, 45]. Part of this difficulty stems from the open-ended nature inherent in video games making it difficult to collect data [18]. In other words, how do you show that students are learning what they should learn and how do you know what you are measuring is what you think you are measuring? [10].

Generally speaking, assessment can be described as either i) *summative* whereby it is conducted at the end of a learning process and tests the overall achievements, and ii) *formative* whereby it is implemented and present throughout the entire learning process and continuously monitors progress and failures [13]. With respect to serious games, it has been suggested that formative assessment is particularly useful and should be used particularly given that such assessments can be incorporated into the serious game becoming part of the experience [84], in particular through appropriate user feedback.

Considering the specific serious game domain, Michael and Chen [70] describe three primary types of assessment: i) completion assessment, ii) in-process assessment, and iii) teacher assessment. The first two correspond to summative and formative assessment, respectively. Completion assessment is concerned with whether the player successfully completes the game. In a traditional teaching environment, this is equivalent to asking, "Did the student get the right answer?" and a simple criterion such as this could be the first indicator that the student sufficiently understands the subject taught albeit, there are many problems using this measure alone. For instance, players could cheat and it is hard to determine whether the player actually learned the material or learned to complete the game [70]. Moreover, the game level upgrade barriers and score (as, in general, all the mechanics) must be designed so as to guarantee a proper balance between entertainment, motivation and learning [71]. In-process assessment (we deal with it in detail in Section 5) examines how, when, and why a player made their choices, and can be analogous to observations of the student by the educator as the student performs the task or takes the test in a traditional teaching environment. Teacher assessment focuses on the instructor's observations and judgments of the student "in action" (while they are playing the game) and typically aims at evaluating those factors that the functionalities/logic of the game are not able to capture.

Although various methods and techniques have been used to assess learning in serious games [82] and simulations in general, summative assessment is commonly accomplished with the use of pre- and post-testing, a common approach in educational research [3]. The pre- and post-test design is one of the most widely used experimental designs and is particularly popular in educational studies that aim to measure changes in educational outcomes after modifications to the learning process such as testing the effect of a new teaching method [29]. Within this design, participants are randomly allocated to either a "treatment" group (playing the serious game) or a "control" group (relying on other instructional techniques). Upon completion of the experiment, both groups complete a post-test, and significant differences across the test scores are attributed to the "treatment" (the serious game) [3]. The main problem with the pre- post-test experimental design is that it is impossible to determine whether the act of pre-testing has influenced any of the results. Another problem relates to the fact that it is almost impossible to completely isolate all of the participants (e.g., if two groups of child participants attend the same school, they will probably interact outside of lessons potentially influencing the results while if the child participants are taken from different schools to prevent this, than randomization is not possible) [74].

The most common method of post-assessment currently consists in testing a players' knowledge about what they learned by way of a survey/test/questionnaire or teacher evaluation. This method is frequently employed because it is the simplest to implement, but it relies on the opinions of the player and does not depend on all of the information that can be collected regarding what happened within the game [84]. This method was used by Allen et al. [1] in the form of questionnaires pre- and post-playing their game, Infiniteams Island Game (TPLD). The goal of the game was for the players to learn about their team working abilities and they were able to show through the questionnaires of 240 students that the players gained self-awareness about their skills through the game. ICURA is another example in which pre- and post-testing assessment was used to evaluated the knowledge learned through the game. Specifically, a role-playing game was used whereby students/players learned about Japanese culture in a role playing format. After playing the game, students completed a test to provide confirmation that they did indeed learn the intended material. The information learned about Japanese culture is more factual than for TPLD so the measure of the person's performance through a test is a more objective assessment of the game.

Another summative assessment technique is given by the "level-up" protocol of testing, whereby players are divided into two groups with one of the groups beginning the game at the first level for example and the other beginning at the second level. If the group that started at the first level does significantly better than the other group, this is attributed to a successful game that is capable of imparting the intended instructional material (at least with respect to the first level) [3].

3.1. Indirect measures of learning

In addition to direct measures of learning achievable through targeted assessment, there are also other factors that can indirectly lead to learning. More specifically, serious games captivate and engage players/learners for a specific purpose such as to develop new knowledge or skills" [23] and with respect to students, strong engagement has been associated with academic achievement [84] and thus the level of engagement may also be potentially used as an indicator to the learning a serious game is capable of imparting.

Various tools have been developed to provide a measure of engagement including the *Game Engagement Questionnaire* [14] and the *Game Experience Questionnaire* [51].

Another key characteristic of a game experience is given by flow – a user state characterized by a high level of enjoyment and fulfillment. The theory of flow is based on Csikszentmihalyi's foundational observations and concepts, and consists of eight major components: a challenging activity requiring skill; a merging of action and awareness; clear goals; direct, immediate feedback; concentration on the task at hand; a sense of control; a loss of self-consciousness and an altered sense of time [25]. Incorporating the concept of flow in computer games as a model for evaluating player enjoyment has been a focus of interesting studies [24, 88] and forms the basis of EGameFlow, a scale that was specifically developed to measure a learner's enjoyment of e-learning games [36]. EGameFlow is a questionnaire that contains 42 items allocated into eight dimensions: i) concentration, ii) goal clarity, iii) feedback, iv) challenge, v) control, vi) immersion, vii) social interaction, and viii) knowledge improvement

In addition to subjective assessment, a growing area of assessment includes a branch of neuroscience that is investigating the correlation between user psychological states and the value of physiological signals. Several studies have shown that these measures can provide an indication of player engagement (see [54, 55, 61, 73]) and flow [76]. Common physiological measures include [72, 73]:

- Facial electromyography (EMG) for measuring muscle activity through the detecting of electrical impulses generated by the muscles of the face when they contract. Such muscle contractions can provide an indication of emotional state and mood and can assess positive and negative emotional valence [61].
- Cardiovascular measures such as the inter-beat interval (the time between heart beats), and heart rate. Cardiac activity has been interpreted as an index to valence, arousal, and attention, cognitive effort, stress, and orientation reflex while viewing various media [61]. Although cardiac measure

have been successfully used in a number of game studies, interpreting as described by Kivinkangas et al., [61], interpreting the relevance of the resulting measurements within a game context is difficult and challenging.

- Galvanic skin response (GSR), for measuring the electrical conductance of the skin, which varies with its moisture (sweat) level and since the sweat glands are controlled by the sympathetic nervous system skin can provide an indication of psychological or physiological(emotional) arousal.
- Electroencephalography (EEG) for measuring the electrical activity along the scalp and more specifically, the measuring the voltage fluctuations resulting from current flows within the neurons of the brain. Depending on the actions performed by the player of a game, differences in the EEG can be detected. For example, Salminen and Ravaja [81] describe a study where the EEG of players playing a video game that involved them steering a monkey into a goal while collecting bananas for extra points while avoiding falling off the edge of the game board. They observed that each of the three events evoked differential EEG oscillatory changes leading the authors to suggest that EEG is a valuable tool when examining psychological responses to video game events. That being said, EEG is not widely used due of its complex analysis procedure [73].

Although there have been a large number of studies investigating the use of physiological responses within a game setting, plenty of work remains in providing a meaningful interpretation of the resulting data to facilitate design decisions for developers of serious games and e-learning application [72]. That being said, the area of physiological measurement within a game context is a promising field and although a complete overview of the field is not provided here, excellent reviews are provided by Kivinkangas et al. [61], and Nacke [73].

3.2. Audio/visual technologies to support assessment

In-process and teacher assessments can be accommodated by the use of recent technology. For example, it is now simple and cost-effective to obtain screen recordings of the player's game play, video recordings of the players while they are playing the game, and audio recordings to capture a players voice for example during thinking aloud processes which may happen unexpectedly or may also be encouraged. With today's technology, information from these recordings can also be obtained automatically (without the need for a camera operator, etc.) using a wide variety of available tools. The recordings and the information obtained from the recordings can also be used to facilitate debriefing sessions.

More recent assessment methods include "information trails" that consists of tracking a player's significant actions and events that may aid in analyzing and answering the what, how, when, who and where in the game something happened. Although this cannot necessarily provide the reasons why a player selected a specific action or event as opposed to another one, it is suggested that this information be obtained from the players through debriefing (interview) session after they complete their gameplay session [30, 66, 71].

3.3. Assessing entertainment

As mentioned in the Introduction, a serious game has a twofold aim of entertainment and education, both of which must be considered in the assessment.

With respect to measuring fun and enjoyment, there are two possible directions: i) quantitative approaches, and ii) qualitative approaches [92]. Qualitative approaches for modeling player enjoyment (e.g., the "entertainment" component) rely primarily on psychological observation, where a comprehensive review of the literature leads to the identification of two major lines: Malone's principles of intrinsic qualitative factors for engaging game play [67] - namely challenge, curiosity and fantasy - and the theory of flow, based on Csikszentmihalyi's foundational concepts [25]. Incorporating flow in computer games as a model for evaluating player enjoyment has been proposed and investigated in significant subsequent studies [24, 88].

In contrast, quantitative approaches attempt to formulate entertainment using mathematical models, which yield reliable numerical values for fun, entertainment or excitement. However, such approaches are usually limited in their scope. For instance, Iida et al. [50] focus on variants of chess games, while Yannakakis and Hallam [92] focus on the player-opponent interaction, which they assume to be the most important entertainment feature in a computer game.

Therefore, there are different dimensions on which the player's experiences can be measured. A recent study has investigated the definition of these dimensions based on the actual players' experience [4]. That work exploited the Repertory Grid Technique (RGT) methodology [41], which includes qualitative and quantitative aspects. Within those studies, players were asked to use their own criteria in describing similarities and differences among video games. Analyzing the players' personal constructs, 23 major dimensions for game assessment were identified, among which the most relevant were: i) ability demand, ii) dynamism, iii) style, iv) engagement, v) emotional affect, and vi) likelihood.

4. Techniques and Tools for Student Performance Assessment

Technology-assisted approaches have been employed for years for student performance assessment, thanks to their potential of streamlining the process of standardized tests and simplify scoring and reporting. Recent studies have explored how technologies and tools can improve the quality of assessments by replacing certain tasks previously done by instructors, enabling customization of tests based on students' performance, allowing real-time bidirectional communication between the instructor and students in classrooms, and adopting novel approaches for assessment.

A number of software products are available for online education testing and assessment [94]. Web-based assessments are useful because they decrease class time used for assessment and because multi-media can be integrated into the testing procedure. However, the deployment of such tools requires careful preparation, and the administrator/educator may lose control of the environment in which the test is taken.

Flynn et al. [34] recommend that pedagogic consideration should be given to the choice, variety, and level of difficulty of eAssessments offered to students. Hewson [48] provides preliminary support for the validity of online assessment methods. Guzman et al. [40] conducted empirical studies in a university setting demonstrating reliability for student knowledge diagnosis of a set of tools for constructing and administering adaptive tests via the Internet. In general, most of these tools are answering the growing needs for larger scale education management. However, this approach also raises serious concerns about the quality of the outcomes.

Table 1 summarizes some tools for eAssessment, that we describe below.

Type	Short description	Sample tools
Assessment Management Systems	Tools to support instructors to create, administer, assess and analyse tests	Assessment Tools for Teaching and Learning (e-asTTle), Questionmark Perception (QP), Assess By Computer (ABC)
Tools for natural language answer assessment	Tools that automatically assess answers written in free text	Short Answer Marking Engine (SAME), Intelligentassessment.com
Classroom response system	Interactive student response systems that enable teachers to instantly assess learning in class	CPS Student Response Systems, SMART Response, i>clicker, 2Know!, Audience Response System, Beyond Question

Table 1. Tools for eAssessment.

There are several computer-based systems available for designing tests and analyzing the results. Assessment Tools for Teaching and Learning (e-asTTle) is an online assessment tool, developed to assess students' achievement and progress in reading, mathematics and writing. It was developed for students ages 8-16 in New Zealand schools and utilizes a computer program to create "paper pencil" tests designed to meet individual learning needs in reading, writing, and mathematics [42]. The system compiles a test based on specified entered characteristics as determined by teachers so that students' learning outcomes can be maximized and students can better understand their progress [43, 44]. e-asTTle allows instructors to create tests that are aligned to the teacher's and the classroom's requirements. It allows measuring student progress over time and provides rich interpretations and specific feedback that relate to student performance. e-asTTle presents the results in visual ways making it easier for teachers to discuss performance.

Similarly, Questionmark Perception (QP) is an assessment management system that enables trainers, educators and testing professionals to author, schedule, deliver, and report on surveys, quizzes, tests and exams. QP includes an authoring manager that allows for creation of surveys, quizzes, tests, and exams with a wide variety of question types and options for embedding media [78] and has been shown to be a successful learning and assessment tool [15, 63].

Assess By Computer (ABC) is also designed for flexible computer-based assessment using a variety of question formats [47]. It allows the administrator to design a test via an interactive user interface and then have the student take the test on a stand-alone computer or within a web-browser. ABC has been designed to deliver and stimulate feedback through the mechanisms of formative assessment in a way that encourages self-regulated learning. The designers of ABC promote it as improving the appropriateness, effectiveness, and consistency of assessments [91].

Short Answer Marking Engine (SAME) is a software system that can automatically mark short answers in a free text form [46]. Short answers are responses to questions in the test takers' own words and therefore better reflect how well they understand the material since they have to provide their own response instead of choosing the most plausible of the alternatives, as with multiple choice questions [56]. Noorbehbahani and Kardan [75] have modified the BLEU algorithm so that it is suitable for assessing free text answers. To perform an assessment, it is necessary to establish a repository of reference answers written by course instructors or related experts. The system calculates a similarity score with respect to several reference answers for each question. As a commercial product, Intelligent Assessment Technologies provide technology to deploy online tests, assessments, and examinations. The technological suite also includes a module for automatically assessing short answers written in natural language.

A classroom response system (CRS) allows two-way communication between an instructor and their students using the instructor's computer and students' input devices [2]. CRS have been increasingly accepted in educational environments from K12 to higher education and also in informal learning environments [33]. Using CRS, the instructor poses questions and polls students' answers during the class enabling real-time two-way communications to occur. The system is also used to take class attendance, pace the lecture, provide formative and formal assessment, to enhance peer instruction, allow for just-in-time-teaching, and increase class interactivity [68]. Real-time interaction between students and instructors results in students paying greater attention and provides instructors with instant feedback on the students understanding of the tested subjects. Commercially available systems include the CPS Student Response Systems from *e*Instruction, SMART Response interactive response systems from SMART Technologies, i>clicker, and 2Know! from Renaissance Learning, as well as the Audience Response System from Qwizdom, and Beyond Question from Smartroom Learning Solutions.

The IMS Question & Test Interoperability (QTI) is a standard interoperability format for representing assessment content and results, such as test questions, tests, and reports, so that they can be used by a variety of different development, assessment, and learning systems and be implemented using a variety of programming languages and modeling tools [53]. Specifically, it has a well-documented format for storing quiz and test items, allowing a wide range of systems to call on one bank of items, and reports results in a consistent format. It is marketed as a way for creating a large bank of questions and answers that will be a able to be used with different systems, now and in the future, and a method for information to be easily shared within and across institutions [85]. Applications can be created using XML (extensible markup language) or higher level development tools including virtual learning enviroments (e.g., Blackboard, JLE ESSI and Oracle iLearning), commercial assessment tools (e.g., Can Studios, Clypso from Experient eLearning Technologies, e-Test 3 from RIVA Technologies Inc, QuestionMark Perception and QuizAuthor by Niall Barr), and R&D assessment tools (e.g., Ultimate Assessment Engine at Strathclyde University and E3AN).

An interesting application of web-based assessment is the assessment of the skills of potential hires. The goal here is to make sure that the candidates that the assessor companies choose to interview and hire have the desired skills for the job. For example, Codility Ltd. offers a service that provides online automated assessments of programming skills by having the test taker write snippets of code which are assessed for correctness and performance [20]. They sell their services to companies to test potential recruit's software skills and assess current employees. International Knowledge Measurement (IKM) is another web-based service that produces an objective and comprehensive profile of knowledge and skill of candidates and employees [52]. Both these services and others (KeneXa Prove It!, eSkill Corporation, etc.) have arisen in response to the desire to efficiently find employees that have desired skills for specific jobs. These methods could be adapted and used for testing before, inside and after a serious game.

5. In-game assessment

Assessment of learning and training requires a systematic approach to determine a person's achievements and areas of difficulty. Standardized assessment methods often take less time, are easier to administer, and their results are readily interpretable [19]. However, there are limitations to such approaches including ineffective measurement of complex problem solving, communication, and reasoning skills [79, 64] . There is also a concern regarding whether the practice of "teaching to the test" has the potential to decrease a student's interest in learning and life-long learning [17, 32]. Furthermore, standardized tests lack the flexibility necessary to adjust or modify materials for certain groups, such as very high- or low-performing groups, and therefore may lead to loss of sensitivity for certain groups [32]. Although some standardized tests have added sections that move away from the concerning "fill-in the bubble approach", this decreases the efficiency of standardized tests.

Recent studies have explored how play-based assessment can provide more detailed and reliable assessment and emerging interests reflect the needs for an alternative or supplemental assessment tool to overcome limitations in the standardized approach [83, 58]. Play-based, or in-game, assessment can provide more detailed and reliable information, and the emerging interest in this field reflects the need for alternative and/or supplemental assessment tools to overcome limitations in the standard approaches [83, 58]. Traditionally, play-based assessment refers to analyzing how a person plays in order to assess their cognitive development, but here we focus on how play with supporting technology can be used as a vehicle to assess cognitive skills, or competences involved in the game, but not to assess the play itself. In particular, digital games have the advantage in this type of assessment that they can easily keep track of every move and decision a player makes [70].

As pointed out by Becker and Parker [3], serious games (and games in general) can and generally do contain in-game tests of effectiveness. More specifically, as players progress through the game, they accumulate points and experience, which enables facing new topics and higher difficulties in the next stages and levels. This is a very ecological and effective approach, since it integrates pedagogy and games, thus allowing provision of immediate feedback to the player and implementing user adaptivity [5, 89].

Incorporating in-game assessments takes us away from the predominant, classic form of assessment comprised of questionnaires, questions and answers, etc. that usually interrupts and negatively affects the learning process [10] and is not very suited to verify knowledge transfer. Designing proper in-game assessment is a challenging and time-consuming activity. However, it should be a distinctive feature of any well-designed serious game, where all the mechanics (e.g., score, levels, leaderboards, bonuses, performance indicators, etc.) should be consistent with and inspired by the set pedagogical targets. [10] provides a detailed survey and analysis of serious games, their components and the related design techniques. Still, "*many educational games do not properly translate knowledge, facts, and lessons into the language of games. This results in games that are often neither engaging nor educational*" [28]. The authors suggest that design should combine "*the fantasy elements and game play conventions of the real-time strategy (RTS) genre with numbers, resources and situations based on research about a real-world topic*", such as energy, agriculture, etc. In this way, the player should be able to learn simply by trying to overcome the game's challenges.

In addition, in-game assessment provides the opportunity to take advantage of the medium itself and employ alternative, less intrusive, and less obvious forms of assessment which could (and should) become a game element itself [10]. Integrating the assessment such that the player is unaware of it forms the basis of what Shute et al. [84] describe as stealth assessment. In this way, the player can concentrate solely on the game [35]. This type of assessment incorporates the assessment in to the process of the game by designing it so that knowledge from previous sections will be necessary to move on in the game and the knowledge is not directly measured using a quiz or questionnaire [49].

Immune Attack is an example of a serious game that uses in-process assessment. It was designed with the goal of teaching students about the immune system in a fun environment and while the game does not directly test the player, it does require that the player retains and learn new information about the immune system so that they can progress in the game [59]. In the game, the player must perform tasks such as training macrophages to identify allies versus enemies, identify if a blood vessel is infected, and countering increasingly more difficult attacks from bacteria [59].

CancerSpace is a game format that incorporates aspects of e-learning, adult-learning theory, and behaviorism theory in order to support learning, promote knowledge retention, and encourage behavior change [87]. CancerSpace's design encourages self-directed learning by presenting the players with real-world situations about which they must make decisions similar to those they would make in clinics. The targeted users are professionals working in community health centers. The gameplay is based on role-playing: the user has to help the clinical staff evaluate the clinical literature, integrate the evidence into their clinical decision-making, plan changes to cancer-screening delivery, and accrue points correlating to increased cancer-screening rates. The user takes decisions and observes whether the chosen course of action improves the cancer screening rates, which is the main indicator of performance. The game includes a small number of patient-provider interactions in which the decider must talk with a patient reluctant to get screened. The player's conversation choices are evaluated in pre-programmed decision trees, leading to success (the patient decides to get screened) or failure. Within this educational context, chance is considered an important entertainment and variability feature, which is implemented through wildcard events. To stimulate gameplay, CancerSpace has adapted an award system that motivates players to increase screening rates. The CancerSpace scenarios in which the decider guides the virtual clinical staff are based on research-tested interventions and best practices. Users receive points on the basis of their performance. At each game's conclusion, a summary screen indicates which decisions the player implemented and their effect on the clinic's screening rate.

In a Living World ad-hoc designed for cultural training in Afghanistan [93], the main objective for a player is to successfully interpret the environment and achieve the desired attitude towards him by Non-Player Characters (NPCs) that represent the local population. The entire living-world game space is fueled by the knowledge-engineering process that translates the essential elements of the culture into programmable behaviors and artifacts. For instance, "*In Afghan culture, older men have great influence over younger men, women, and children through local traditions and Islamic law*" or "*Ideologically, the guiding principles of Afghan culture are a sense of familial and tribal honor, gender segregation, and indirect communication*". All the NPCs in the game are modeled accordingly. Winning in the game "simply" requires successfully navigating cultural moves in the game space, thus achieving a good overall attitude of the village toward the player. Another key aspect is serious-ness about assessment. The underlying 3D Asymmetric Domain Analysis and Training (3D ADAT) model, an ad-hoc developed recursive platform for the realization and visualization of dynamic sociocultural models, specifically supports analysis of the cultural behavior exhibited by the player in the game. Conversations and interactions between the NPCs and the player are recorded through a text log to provide game performance analysis. The assessment tool lists all the possible choices for player behavior and conversation, highlighting both the player's choice and the most culturally appropriate response. The tool provides scores on the opinion of the player at the NPC, faction, and village level. Additional comments can be provided that highlight the player's weaknesses, explaining why a particular response is most appropriate. Feedback is thus provided to improve future performance.

Business games, also known as business simulations, are another well-established category of serious games, that are being used for many decades (originally in non-digital form – thus, they were not called serious games) in business schools [60, 86]. In SimVenture, the target of the player is to manage a company, dealing with four major types of issues: production, organization, sales & market and finance. The player has a number of choices to perform in these domains. Their performance is expressed in terms of a parameter called “company value”. But, as in the real-world, the player has to maintain a number of factors, such as profit and loss, a balance sheet, and cash flow. Several other performance figures are also reported in the performance report. Each game session has a simulated time limit, expressed in months. The goal of the game – it can be fixed by the teacher or by the players themselves – can be the maximization of the profit or of cash flow (or any other parameter). Of course, players have to avoid bankruptcy within their time-limit. Several pre-defined scenarios are available and can be loaded by players and classes, so that they can face some common critical cases (e.g., start up a company, managing growth, facing cashflow issues, etc.) at various levels of difficulty. Messages are displayed to the player, at the end of each month’s simulation, highlighting the major issues encountered and to be faced. When defining a new game session, there is the possibility of introducing chance events. In the absence of chance events, the game session is deterministic, thus allowing a straightforward comparison of the performance of various players. Simventure also includes complementary material for teachers and learners.

This material proposes also some additional activities, such as debriefing, answering questions, writing essays, forecasting events and outcomes and business planning that are to be performed under the supervision and with the help of a teacher. This – in particular the presence of a teacher - is important in order to complement the operational knowledge and skills acquired through the gaming (problem-based learning, experiential learning, etc.) with reflection and verbal knowledge and exchange.

PIXELearning’s Enterprise Game is a similar business game, with a major hyphenation on graphic quality and look and feel. Also in this case, defining a product meeting the market demand in terms of quality and price is the most important factor to make the business viable. Definition of a proper marketing strategy is key as well. Here, the performance of competitor companies is also continuously displayed, so that the player is challenged to do better also with respect to them. Both SimVenture and The Enterprise Game, are single player games, while a multi-player web-based environment would probably enhance the playability through online competition and collaboration.

6. Conclusions and directions for future research

For serious games to be considered a viable educational tool, they must provide some means of testing and progress-tracking and the testing must be recognizable within the context of the education or training they are attempting to impart [70]. Various methods and techniques have been used to assess effectiveness of serious games and various comprehensive reviews have been conducted to examine the overall validity of game-based learning. Results of these reviews seem to suggest that game-based learning is effective for motivating and for achieving learning goals at the lower levels in the Bloom’s taxonomy [22].

However, caution is still required with respect to many of the claims that have appeared in the literature about the “revolution” due to the use of serious games in education. Achieving more ambitious learning goals seems to require studying new types of games able to foster more accurate reasoning and reflection, stimulated through proper teacher guidance, allowing the player to efficiently structure the knowledge space. We also believe that comparison studies with other educational technologies should be carried out in order to better understand the serious games’ effectiveness.

Assessing the user learning within a simulation or serious game is not a trivial matter and further work and studies are required. With the advent of cheaper hardware and software, it has been possible to extend and enhance assessment by recording game-play sessions, and keeping track of players’ in-game performance. In-game assessment appears to be particularly suited and useful given that it is is integrated into the game logic, and therefore, does not break the player’s game experience. Furthermore, it enables immediate provision of feedback and implementation of adaptability. In general, for assessment design, it must be stressed that clear goals must be set, followed by techniques to collect data that will be used to verify these goals.

As Kevin Corti of PIXELearning stated, *"[Serious games] will not grow as an industry unless the learning experience is definable, quantifiable and measurable. Assessment is the future of serious games"* [80]. This requires still a lot of research work. We see in particular two major research directions: characterization of the player's activity and better integration of assessment in games.

Characterization of the player's activities involves both task characterization (e.g., in terms of content, difficulty level, type of supported learning style, etc.) and user profiling [6]. It is necessary to identify the dimensions, relevant to learning, along which the users and the tasks are modeled. Then, the matching rules and modalities between users and tasks should be defined. The user profile should be portable across different games and even applications, particularly in the education field. Here, it is particularly important to consider also misconceptions and mistakes. In user profiling, analysis of neuro-physiological signals is particularly promising, as it allows a continuous, in-depth and quantitative monitoring of the user activity and state. Finally, proper user profiling is key to enable adaptability and personalization.

Better integration of assessment in games is essentially a matter of definition of the proper mechanisms and conditions to activate them. It is important that these mechanisms should be general and modular, so to be seamlessly applicable in different games. This will increase efficiency in designing games and authoring contents, which is a key requirement for the serious game industry [7]. A strictly related topic concerns provision of feedback, which is a consequence of assessment and should be properly integrated in the game, in order not to distract the player and also to.

Acknowledgement

We are grateful to the reviewers for their suggestions that have allowed us to significantly improve the quality of the paper. This work has been partially funded by the EC, through the GALA EU Network of Excellence in Serious Games (FP7-ICT-2009-5-258169). The financial support of the *Social Sciences and Humanities Research Council of Canada* (*SSHRC*) in support of the iMMERSE project that B. Kapralos is part of, is gratefully acknowledged.

References

1. Allen L., M. Seeney, L. Boyle, F. Hancock. "The Implementation of Team Based Assessment in Serious Games". 1st IEEE International Conference in Games and Virtual Worlds for Serious Applications, Coventry, England, March 23-24, 2009.

2. Beatty I., W. Gerace (2009). "Technology-Enhanced Formative Assessment: A Research-Based Pedagogy for Teaching Science with Classroom Response Technology" J Sci Educ Technol 2009, v. 18 p. 146-162.

3. Becker K., and J. R. Parker. The Guide to Computer Simulations and Games. John Wiley & Sons Inc., Indianapolis, IN, USA, 2011.Bellotti F., Berta R., De Gloria A., Primavera L., (2009) "Enhancing the Educational Value of Video Games", ACM Computers in Entertainment, Vol. 7, No.2, pp. 23-41

4. Bellotti F., Berta R., De Gloria A., Primavera L., (2009a), "Enhancing the Educational Value of Video Games", ACM Computers in Entertainment, Vol. 7, No.2, pp. 23-41.

5. Bellotti, F., Berta, R., De Gloria, A., & Primavera, L. (2009b). Adaptive Experience Engine for Serious Games. IEEE Transactions on Computational Intelligence and AI in Games, 1(4), 264-280.

6. Bellotti F., Berta R., De Gloria A., Primavera L. (2009c), "A task annotation model for SandBox Serious Games", Proceedings of IEEE Symposium on Computational Intelligence and Games (CIG 2009), pp. 233-240, September 7-10, 2009, Milano, Italy

7. Bellotti F., Berta R. and De Gloria A., (2010a), “Designing Effective Serious Games: Opportunities and Challenges for Research”, Special Issue: Creative Learning with Serious Games, Int.l Journal of Emerging Technologies in Learning (IJET), 5, 2010, pp. 22-35

8. Bellotti F., Berta R., De Gloria A., Primavera L., (2010b), “Supporting authors in the development of Task-Based Learning in Serious Virtual Worlds”, British Journal of Education and Technologies (BJET), Vol. 41, No. 1, January 2010 , pp. 86-107

9. Bellotti F., Berta R., De Gloria A., D'Ursi A., and Fiore V, A serious game model for cultural heritage. ACM *J. Computers and Cultural Heritage,* 5, 4, 2012

10. Bente G., and J. Breuer. Making the implicit explicit: Embedded measurement in serious games. In Serious Games: Mechanisms and Effects. U. Ritterfield, M. J. Cody, P. Vorderer (Eds.), pp. 322-343. 2009.

11. Bergeron B.. Developing serious games. Hingham, MA. USA, Thomson Delmar Learning, 2006.

12. Blunt, R. (2009). Do Serious Games Work? Results from Three Studies. *eLearn Magazine*, 2009(12).

13. Boston C.. The concept of formative assessment. *Practical Assessment, Research & Evaluation*, 8(9), 2002. Retrieved from: http://ericae.net/pare/getvn.asp?v-8 &n=9.

14. Brockmyer J. H., C. M. Fox, K. A. Curtiss, E. McBroom, K. M. Burkhart, and J. N. Pidruzny, “The development of the Game Engagement Questionnaire: A measure of engagement in video game-playing,” *Journal of Experimental Social Psychology*, vol. 45, no. 4, pp. 624-634, 2009.

15. Bull J., R. Stephens (1999). “The use of Question Mark software for formative and summative assessment in two universities”. IETI 1999; v. 36, p. 128–136.

16. Cannon-Bowers J., “The state of gaming and simulation,” in *Proceedings of the Training 2006 Conference and Expo*, Orlando, FL., USA, March 2006.

17. Chicago Tribune, 2010. “Standardized testing will limit students' future” http://articles.chicagotribune.com/2010-04-21/news/chi-100421shafer_briefs_1_standardized-test-scores-teacher-and-principal-evaluations April 21, 2010

18. Chin J. and R. Dukes, and W. Gamson, Assessment in simulation and gaming: “A review of the last 40 years,” *Simulation & Gaming*, vol. 40, no. 4, pp. 553-568, 2009.

19. Clarke M. M., G. F. Madaus, C. L. Horn And M. A. Ramos. 2000 “Retrospective on educational testing and assessment in the 20th century” J. CURRICULUM STUDIES, 2000, VOL. 32, NO. 2, 159-181

20. Codility Ltd. 2009. “codility: WE TEST CODERS”. http://codility.com/

21. Cole SW, Yoo DJ, Knutson B (2012) Interactivity and Reward-Related Neural Activation during a Serious Videogame. PLoS ONE 7(3), 2012

22. Connolly, T. M., E. A. Boyle, E. MacArthur, T. Hainey and J. M. Boyle (2012). "A systematic literature review of the empirical evidence on computer games and serious games." Computers and Education 59(2), pp. 661-686, 2012

23. Corti K.. Game-based learning; a serious business application, PIXELearning. Coventry, UK, 2006.

24. Cowley, B., D. Charles, et al. "Toward an understanding of flow in video games." Computers in Entertainment 66(2): 1-27. 2008.

25. Csikszentmihalyi M., Flow: The Psychology of Optimal Experience, New York: Harper & Row, 1990.

26. Dandeneau S., and M. Baldwin, “The inhibition of socially rejecting information among people with high versus low self-esteem: The role of attentional bias and the effects of bias reduction training,” *Journal of Social and Clinical Psychology*, vol. 23, no. 4, pp. 584-602, 2004.

27. De Grove, F., Mechant, P., Van Looy, J, Uncharted waters?: exploring experts' opinions on the opportunities and limitations of serious games for foreign language learning. In Proceedings of the 3rd international Conference on Fun and Games, Leuven, Belgium, September, 2010.

28. Doucet, L. and Srinivasan, V. 2010. Designing entertaining educational games using procedural rhetoric: a case study. In Proceedings of the 5th ACM SIGGRAPH Symposium on Video Games, Los Angeles, Ca, July, 2010.

29. Dugard P., and Todman J., "Analysis of pre-test-post-test control group designs in educational research," *Educational Psychology*, vol. 15, no. 2, pp. 181-198, 1995.

30. Enfield J., R. D. Myers, M. Lara, and T. W. Frick, "Innovation Diffusion: Assessment of Strategies Within the diffusion simulation game," Simulation & Gaming, to appear 2012.

31. Fairtest, "The Limits of Standardized Tests for Diagnosing and Assisting Student Learning", http://www.fairtest.org/limits-standardized-tests-diagnosing-and-assisting, Aug. 17, 2007

32. Fairtest, "What's Wrong With Standardized Tests?", http://www.fairtest.org/facts/whatwron.htm, May 22, 2012

33. Fies C., J. Marshall (2006). "Classroom Response Systems: A Review of the Literature". Journal of Science Education and Technology, v. 15 issue 1, 2006, p. 101-109.

34. Flynn A., Concannon F., and Campbell M.. An evaluation of undergraduate students' online assessment performances. *Adv. Technol. Learn.* 3, 1, 2006.

35. Froschauer J., I. Seidel, G. Markus, H. Berger, D. Merkl (2010)."Design and Evaluation of a Serious Game for Immersive Cultural Training". 16th International Conference on Virtual Systems and Multimedia, Seoul, Korea. Published in:"International Conference on Virtual Systems and Multimedia", IEEE CS Press, 2010, ISBN: 9781424490257, p. 253 - 260.

36. Fu F. L., Su R. C., and Yu S. C.. "EGameFlow: A scale to measure learners' enjoyment of e-learning games," *Computers and Education*, vol. 52, no. 1, pp. 101-112, 2009.

37. Gee, J.P. What video games have to teach us about learning and literacy. Palgrave MacMillan: New York. 2007

38. Gosen J., and Washbush J., "A review of scholarship on assessing experiential learning effectiveness," *Simulation & Gaming*, vol. 35, no. 2, pp. 270-293, 2004.

39. Greitzer F.L., Kuchar O.A., and Huston K., "Cognitive Science Implications for Enhancing Training Effectiveness in a Serious Gaming Context", ACM J. Educational Resources in Computing, vol. 7, no. 3, August 2007.

40. Guzman E., Conejo R., and Perez-de-la-Cruz, J-L. Improving Student Performance Using Self-Assessment Tests. *IEEE Intelligent Systems* 22, 4, 2007.

41. Hassenzahl, M., Wessler, R., Capturing design space from a user perspective: the repertory grid technique revisited, International Journal of Human—Computer Interaction, 12(3&4), 2000, 441-459.

42. Hattie J., D. Masters (2006). "asTTle – Assessment Tools for Teaching and Learning" HEFCE JISC. <http://www.jisc.ac.uk/media/documents/projects/asttle_casestudy.pdf>

43. Hattie J., G. Brown, P. Keegan, S. Irving, A. MacKay, T. Sutherland, D. Mooyman, P. Patel. "Validation evidence of asTTle reading assessment results: Norms and criteria". Asttle Tech. Rep. 22, University of Auckland/Ministry of Education, November 2003.

44. Hattie J.."Large-scale Assessment of Student Competencies" Symposium: Working in Today's World of Testing and Measurement: Required Knowledge and Skills (Joint ITC/CPTA Symposium); 26th International Congress of Applied Psychology; July 16-21, 2006, Athens, Greece.

45. Hays R. T.. The effectiveness of instructional games: A literature review and discussion. Technical Report 2005-004, Naval Air Warfare Center, Training Systems Division, 2005.

46. HEFCE JISC (2009). “Short answer marking engines”. http://www.jisc.ac.uk/media/documents/projects/shorttext.pdf

47. HEFCE JISC 2010. “Case study 5: Making the most of a computer-assisted assessment system University of Manchester”. http://www.jisc.ac.uk/media/documents/programmes/elearning/digiassess_makingthemost.pdf

48. Hewson. C. Can online course-based assessment methods be fair and equitable? Relationships between students' preferences and performance within online and offline assessments. *J. Comp. Assist. Learn.* 28, 5, 2012.

49. Howell K., E. Glinert, L. Holding, C. Swain. “How to build serious games.” *COMMUNICATIONS OF THE ACM*, v. 50 issue 7, 2007, p. 44-49.

50. Iida, H., N. Takeshita, and J. Yoshimura. A metric for entertainment of boardgames: its implication for evolution of chess variants. In R. Nakatsu and J. Hoshino, editors, IWEC2002 Proceedings, pages 65-72. Kluwer, 2003.

51. IJsselsteijn W. A., W. van de Hoogen, C. Klimmt, Y. de kort, C. Lindley, K. Mathiak, K. Poels, N. Ravaja, M Turpeinen, and P. Vorderer, “Measuring the experience of digital game enjoyment,” in *Proceedings of Measuring Behavior 2008*, Maastricht, The Netherlands, August 2008.

52. IKM 2011. “About IKM: Overview”. http://www.ikmnet.com/about/overview.cfm

53. IMS QTI 2012. “IMS Global Learning Consortium: IMS Question & Test Interoperability Specification (QTI)”. <http://www.imsglobal.org/question/>

54. Janicke S. H., and A. Ellis, “Psychological and physiological differences between the 3D and 2D gaming experience,” in *Proceedings of the 3D Entertainment Summit*, Hollywood, CA., USA, September, 2011.

55. Jelinek H. F., K. August, H. Imam, A. H. Khandoker, A. Koenig, and R. Riener, “Heart rate asymmetry and emotional response to robot assist task challenges in stroke patients,” in *Proceedings of the 2011 Computing in Cardiology Conference*, Hangzhou, China, September 2011.

56. Jordan S., T. Mitchell (2009). “E-assessment for learning? The potential of short free-text questions with tailored feedback”. British Journal of Educational Technology, v. 40 issue 2, p. 371-385.

57. Kato P. M., Cole S. W., Bradlyn A. S., Pollock B. H., A Video Game Improves Behavioral Outcomes in Adolescents and Young Adults With Cancer: A Randomized Trial, Pediatrics Vol. 122 No. 2 August 1, 2008

58. Kaugars, A. S., & Russ, S. W. (2009). Assessing Preschool Children's Pretend Play: Preliminary Validation of the Affect in Play Scale-Preschool Version. Early Education and Development, 20(5), 733-755.

59. Kelly H., K. Howell, E. Glinert, L. Holding, C. Swain, A. Burrowbridge, and M. Roper. 2007. How to build serious games. *Commun. ACM* 50, 7 (July 2007), 44-49.

60. King M., Newman R., (2009) "Evaluating business simulation software: approach, tools and pedagogy", On the Horizon, Vol. 17 Iss: 4, pp.368 – 377

61. Kivikangas J. M., Chanel G., Cowley B., Ekman I., Salminen M., Järvelä S., Ravaja N. “A review of the use of psychophysiological methods in game research”, vol. 3, no. 3, 2011.

62. Kulik A. A.. School mathematics and science programs benefit from instructional technology. United States National Science Foundation (NSF), National Center for Science and Engineering Statistics (NCSES). InfroBrief NSF-03-301, November 2002. Accessed February 22, 2012 from http://www.nsf.gov/statistics/infbrief/nsf03301/

63. Kumar R., M. Dziegielewski, D. Wakefield (2002). “Web-based self-assessments in pathology with Questionmark Perception”. PATHOLOGY, v. 34 issue 3, 2002, p. 282-284.

64. Lipman, M., "Some Thoughts on the Formation of Reflective Education." In Teaching-Thinking Skills: Theory and Practice, pp. 151-161. Edited by J.B. Baron and R. J. Sternberg. New York: W. H. Freeman, 1987.

65. Livingston S., G. Fennessey, J. Coleman, K. Edwards, and S. Kidder. The Hopkins games program: Final report on seven years of research (Report No. 155). Baltimore: Johns Hopkins University, Center for Social Organization of Schools, 1973.

66. Loh, C. S. (2007). Designing Online Games Assessment as Information Trails. In D. Gibson, C. Aldrich, & M. Prensky (Eds.), Games and Simulations in Online Learning: Research and Development Frameworks (pp. 323–348). Hershey, PA: Information Science Publishing.

67. Malone, T. (1981). Toward a Theory of Intrinsically Motivating Instruction. Cognitive Science, 5(4), 333–369.

68. Marshall J. (2008). "The C(3) Framework: Evaluating Classroom Response System Interactions in University Classrooms." Journal of Science Education and technology, v. 17 issue 5, 2008, p. 483-499.

69. Michael D., and S. Chen, S. Serious Games: Games that Educate, Train and Inform. Boston, MA. USA, Thomson Course Technology, 2006.

70. Michael D., and S. Chen. "Proof of learning: Assessment in serious games." http://www.gamasutra.com/view/feature/2433/proof_of_learning_assessment_in_.php 2005. Accessed on February 22, 2012.

71. Moreno-Ger, P., Burgos, D., & Torrente, J.. Digital Games in eLearning Environments: Current Uses and Emerging Trends. Simulation & Gaming, 40(5), 669–687, 2009.

72. Nacke L. E., "Physiological Game Interaction and Psychophysiological Evaluation in Research and Industry", Gamasutra Article, June 28, 2011. http://www.gamasutra.com/blogs/LennartNacke/20110628/7867/Physiological_Game_Interaction_and_Psychophysiological_Evaluation_in_Research_and_Industry.php. Accessed on: December 22, 2012.

73. Nacke L. E. , "Affective Ludology: Scientific Measurement of User Experience in Interactive Entertainment," Ph.D. Thesis. Blekinge Institute of Technology, Karlskrona, Sweden (2009).

74. National Center for Technology Innovation (NCTI), "Experimental Study Design", http://www.nationaltechcenter.org/index.php/products/at-research-matters/experimental-study-design/ Accessed on: December 22, 2012.

75. Noorbehbahani F. and Kardan A. A., The automatic assessment of free text answers using a modified BLEU algorithm. *Computera & Education.* 56, 2, 2011

76. Plotnikov A., Stakheika N., De Gloria A., Schatten C., Bellotti F., Berta R., Fiorini C., Ansovini F., "Exploiting real-time EEG analysis for assessing flow in games", Workshop: "Game based learning for 21st century transferable skills", at iCalt 2012, Rome, Italy, June 2012.

77. Prensky, M., Don't Bother Me Mom—I'm Learning!, Paragon House, 2006.

78. Questionmark Corporation 2012. "Questionmark Perception: Measure Knowledge, Skills and Attitudes Securely for Certification, Regulatory Compliance and successful Learning Outcomes".

79. Resnick, L.B., and Resnick D.P. Assessing the Thinking Curriculum: New Tools for Educational Reform. Pittsburgh, Pennsylvania: Learning Research and Development Center: University of Pittsburgh and Carnegie Mellon University, 1989.

80. Ritterfeld U., M. Cody, & P. Vorderer (Eds.), *Serious games: Mechanisms and effects*. New York: Routledge, 2009, pp. 117-130.

81. Salminen M., Ravaja N. "Oscillatory brain responses evoked by video game events: the case of super monkey ball 2", *CyberPsychology & Behavior*, vol. 10, no. 3, 330-338, 2007.

82. Sebastian C., A. Anantachai, J.H. Byun, J. Lenox. "Assessing What Players Learned in Serious Games: In-situ Data Collection, Information Trails, and Quantitative Analysis". Proceedings of the 10th International Conference on Computer Games: AI, Animation, Mobile, Educational and Serious Games, 2007, p. 10-19.

83. Short EJ, Noeder M, Gorovoy S, Manos MJ and Lewis B. (in press). The Importance of Play in Both the Assessment and Treatment of Young Children. In S. Russ & L. Niec (Eds). An Evidence-Based Approach to Play in Intervention and Prevention: Integrating Developmental and Clinical Science. Guilford.

84. Shute V., M. Ventura, M. Bauer, D. Zapata-Rivera. "Melding the power of serious games and embedded assessment to monitor and foster learning: Flow and grow". In U. Ritterfeld, M. Cody, & P. Vorderer (Eds.), Serious games: Mechanisms and effects (pp. 295-321). Mahwah, NJ: Routledge, Taylor and Francis. 2009.

85. Smythe C., P. Roberts. "An Overview of the IMS Question & Test Interoperability Specification". Computer Aided Assessment, Leicestershire, UK, 2000.

86. Stainton, A.J., Johnson J.E., Borodzicz E. P., (2010). Educational Validity of Business Gaming Simulation: A Research Methodology Framework Simulation & Gaming vol. 41, 5: pp. 705-723

87. Swarz, J., Ousley, A., Magro, A., Rienzo, M., Burns, D., Lindsey, A.M., Wilburn, B., Bolcar, S., , "CancerSpace: A Simulation-Based Game for Improving Cancer-Screening Rates," Computer Graphics and Applications, IEEE , vol.30, no.1, pp.90-94, Jan.-Feb. 2010

88. Sweetser P. and Wyeth P., "GameFlow: A Model for Evaluating Player Enjoyment in Games," ACM Computers in Entertainment, vol. 3, no. 3, July 2005.

89. Tan C. H., K. C. Tan and A. Tay: Dynamic Game Difficulty Scaling Using Adaptive Behavior-Based AI. IEEE Trans. Comput. Intellig. and AI in Games 3(4): 289-301 (2011)

90. Van Eck R., Digital Game-Based Learning: It's Not Just the Digital Natives Who Are Restless, EDUCAUSE Review, vol. 41, no. 2, 2006

91. Wood M., 2009. "Human Computer Collaborative Assessment – Access by Computer (ABC) – University of Manchester" HEFCE JISC 2009.

92. Yannakakis G. N., and Hallam, J. "Evolving opponents for interesting interactive computer games", Proceedings of the International Conference on Computer Games: Artificial Intelligence, Design and Education, 2004.

93. Zielke M. A., Evans M. J., Dufour F., Christopher T. V., Donahue J. K., Johnson P., Jennings E. B., Friedman B. S., Ounekeo P. L., Flores R., "Serious Games for Immersive Cultural Training: Creating a Living World," IEEE Computer Graphics and Applications, vol. 29, no. 2, pp. 49-60, Mar./Apr. 2009

94. Zoanetti N. (2011.). "Software for online testing and quizzes". assessmentfocus.com, <http://www.assessmentfocus.com/online-testing.php>

Conflict of interests

The authors hereby declare that they have no conflict of interests with the companies/commercial products cited in this paper.