EVALUATING TECHNOLOGY-BASED EDUCATIONAL INTERVENTIONS: A REVIEW OF TWO PROJECTS

NICOLETTA ADAMO-VILLANI
HAZAR DIB
Purdue University, West Lafayette, Indiana

ABSTRACT
The article discusses current evaluation methodologies used to assess the usability, user enjoyment, and pedagogical efficacy of virtual learning environments (VLEs) and serious games. It also describes the evaluations of two recently developed projects: a virtual learning environment that employs a fantasy 3D world to engage deaf and hearing children in math- and science-based educational tasks, and a serious game for undergraduate students’ learning of sustainable building concepts and practices. Although a growing body of evidence suggests that serious games and VLEs could be advantageous and beneficial as instructional tools, research focused on their role in educational settings still lacks a unified approach to evaluating these interventions. To address this gap, the article provides an overview of current challenges associated with assessing the effectiveness of serious games and VLEs. In addition, the results from two projects provide lessons learned, and are intended to guide future efforts in developing and evaluating similar educational technologies.

1. INTRODUCTION
Recent reviews have concluded that technology has a great potential to enhance student achievement [1-4]. Technologies, such as Virtual Reality (VR) and serious games, can be used to create interactive environments where learners can
visualize abstract concepts easily and receive feedback to build new knowledge and understanding [5, 6]. Recent research findings show that virtual learning environments can be more effective than traditional teaching tools [7-9]. In general, educational technologies support learning in a nonlinear fashion which has been shown to be effective in teaching students how to be critical and creative thinkers [10].

Although there is a growing body of evidence that suggests that serious games and VLEs could be beneficial as instructional tools, evaluating their effectiveness is still a challenging task. In order for serious games and VLEs to be effective, they need to be usable, enjoyable (e.g., motivating and engaging), and promote learning. Thus, evaluation of these technologies needs to address all three aspects, usability, user enjoyment, and pedagogical efficacy. While methodologies for assessing usability of educational technologies are fairly well established, there is still no clear consensus on evaluation of user enjoyment (also called gameplay experience) and impact upon learning. The challenge of assessing gameplay experience is due to the fact that player enjoyment is determined by a large variety of complex and vaguely defined experiential phenomena. Current research in this area is still trying to establish correlations between psychophysiological responses and questionnaire data in order to reach a scientifically grounded understanding of gameplay experience, and hence a valid evaluation framework.

Evaluating the learning effect is challenging, as serious games and VLEs target many aspects of learning (cognitive, metacognitive, and motivational) that are difficult to measure. Some researchers argue that metrics used to measure learning within games and VLEs are often inadequate and unable to capture important components of learning such as procedural knowledge and metacognitive aspects. Nikolic and Messner raise “concerns regarding both internal and external validity of experimental studies due to inevitable differences in students, classroom settings, treatments and other practical issues that make it impossible to generalize the results” [11]. Jenkinson argues that there is a problem of mismeasure (either qualitative or quantitative) of e-learning, and raises the question of “precisely what it is that we are attempting to measure when we evaluate educational technology?” [12].

This article presents an overview of current evaluation methods of serious games and VLEs, and discusses challenges associated with them. In addition, the article describes the evaluations of two recently developed projects: a serious game for undergraduate students’ learning of building sustainability, and a virtual learning environment for bilingual K-5 math and science education. The results from the two projects provide lessons learned, and are intended to guide future efforts in developing and evaluating similar educational interventions. The article is organized as follows: in section 2 we define VLEs and serious games and describe the underlying pedagogies. In section 3 we discuss current evaluation methodologies and in section 4 we report and discuss two recent projects. Conclusive remarks are included in section 5.
2. VIRTUAL LEARNING ENVIRONMENTS AND SERIOUS GAMES

The name “Virtual Learning Environment (VLE)” is used broadly and because of this broad usage, it can refer to a website with static web pages. At the other end of the continuum, the term VLE can refer to a highly integrated system, including the associated infrastructure, that: a) substantively incorporates 3D virtual reality technology (i.e., an environment that represent objects with a high degree of fidelity); b) is immersive, meaning that educational interactions occur within the environment making the environment a “place”; c) is a designed information space in which the information is explicitly represented and the educational interactions occur so that students are not only active, but actors, that is, they co-construct the information space; d) overlaps and is an extension of the physical environment, thereby possessing the potential to enrich classroom activities; and e) can integrate heterogeneous technologies and multiple pedagogical approaches [13].

In Mike Zyda’s definition, a game is “a physical or mental contest, played according to specific rules, with the goal of amusing or rewarding the participant,” while a serious game is “a mental contest, played with a computer in accordance with specific rules that uses entertainment to further government or corporate training, education, health, public policy, and strategic communication objectives” [14]. Serious gaming repurposes the concepts of videogames and videogame technologies that have been used for commercial entertainment, and uses the gaming approach for training, education, advertising, national defense, general productivity, and more. The founders of the Education Arcade at MIT stated that there are many intrinsic motivations for learning associated with games. “The threat of failure is lowered. Games allow players to try, make mistakes or fail, and then try again without losing face. Discovery and application of learned skills in new contexts encourages exploration and experimentation. A sense of engagement continues during gaming. Computer games allow players to be stakeholders in the events that occur on the screen” [15].

Play, an important contributor to human development, maturation, and learning, is the main component of serious games [16]. Another fundamental element is “fun.” It is the “fun” factor that distinguishes serious games from other pedagogical approaches in that the learner is compelled to learn not necessarily due to the subject matter’s intrinsic appeal, but rather due to the entertainment value of the gaming activity with which the subject matter is associated. The player of a serious game is motivated to play the game, and in so doing continues the lesson much longer and with greater attention than he or she would using traditional learning techniques.

The theoretical framework underlying many VLEs and serious games is guided by the tenets of constructivist learning theories and experiential learning theory [17]. Constructivist theories posit that learning occurs in contexts and that knowledge is grounded in authentic situations, that is, learning and understanding derive from experiences. There are many constructivist theories—e.g., social
constructivism, cognitive constructivism, and situated constructivism—however, they share a perspective that learning is active and self-directed, involves “meaning making,” and “requires the personal interpretation of phenomenon such as the construction of a mental model representing a complex phenomenon” [18]. As stated by Shiratuddin and Hajnal [19], many educational technologies, including serious games and VLEs, offer the opportunity to implement constructivist approaches to teaching and learning in various education settings. For example, serious games and VLEs are very user-centered and engaging [20] and have the potential to transform learners from passive participants to active participants in the learning process. They often incorporate problem-solving activities, thus enhancing real life knowledge application.

Experiential learning theory posits that learning is an active process where “knowledge is created through the transformation of experience” [17, p. 38]. Kolb further posits that people learn best by doing, and that effective learning makes connections between what is learned (i.e., acquired knowledge) and its practical application [21]. Unlike traditional instructional approaches where learning occurs in “one-way information-dispensing methods,” experiential learning approaches (e.g., games and VLEs) foster active learning, address cognitive (i.e., content knowledge) and affective learning issues, enhance collaboration and peer learning, and offer opportunities for more complex and diverse approaches to learning processes and outcomes [22].

3. CURRENT METHODOLOGIES FOR EVALUATING EDUCATIONAL TECHNOLOGIES

Evaluation of serious games and VLEs has focused primarily on three aspects: usability, user enjoyment, and pedagogical efficacy.

Usability is defined as the ease of use and learnability of a software application. Several usability evaluation methods that have been applied to VLEs and serious games can be found in the literature. They include: Formative Evaluation, Heuristic or Guidelines-Based Expert Evaluation, Cognitive Walkthrough, Questionnaires/Surveys, and Interviews [23].

Formative Evaluation is an “observational, empirical evaluation method that assesses user interaction by placing target users in task-based scenarios” [23]. Formative evaluations can collect both quantitative and qualitative data. In general, qualitative data include user comments, critical incidents, and users’ perceptions and reactions; quantitative data include metrics recorded by the application engine such as time spent on a specific task, number of mistakes, completion/non-completion of an activity, mouse trajectories, etc. Heuristic or Guidelines-Based Expert Evaluation [24] is a method in which a panel of experts assesses usability and functionality of an application by comparing it to a set of established usability design guidelines, providing design improvement feedback. A Cognitive Walkthrough [25] is an approach to evaluating a user
interface based on going step-by-step through common tasks that a user would perform, and evaluating the interface’s ability to support each step. This approach is intended especially to help understand how well the interface supports “exploratory learning,” that is, first-time use without formal training. Questionnaire/Surveys [26] are written sets of questions used to obtain comments/perceptions from users after they have used the software application. Interviews [26] are primarily used for gathering information about the users’ experience by talking directly to them. In general, interviews can gather more detailed information than surveys/questionnaires.

While evaluation methods for assessing the usability of educational technologies are fairly well established, evaluation of user enjoyment (or gameplay experience) still lacks a formal, unified approach. User enjoyment is a fundamental component of serious games and VLEs. If users do not enjoy playing the game or interacting with the VLE, they will not continue to play and therefore will not learn. Current methods of evaluating user enjoyment include both subjective and objective techniques. The most common subjective methods include self-reports, usually through questionnaires, surveys, interviews, and think-aloud protocols; objective methods include reports through observational video analysis, psychophysiological testing and behavioral tracking. Subjective reporting through questionnaires and interviews is generalizable and easy to analyze. Some disadvantages of questionnaires and surveys are that they can invade privacy, and subject responses may not correspond to the actual experience. “Think-aloud techniques, while informative, only qualify the experience, rather than providing concrete quantitative data” [27]. Using video recordings to code body language and verbalizations is a rich source of data. However, it is a very time-consuming, rigorous process requiring a high level of attention to details. For this reason, many researchers often rely on subjective data for user preference, rather than video observational analysis.

Recently, new sensor technology and behavioral tracking, such as Eye Tracking and Psychophysiological Player Testing, have made it possible to model and assess user cognition and emotion during gameplay. Eye Tracking measures the saccadic movements (fast motions) and fixations (dwell times) of the human gaze. “Due to the relationship between eye fixations and attentional focus, it is possible to infer and visualize cognitive and attentional processes in gameplay and virtual environment exploration” [28]. Psychophysiological Tests are controlled measures of gameplay experience usually deployed in a laboratory with the benefit of covertly assessing physical reactions of players. They include: electromyography (EMG), a measurement technology for recording the electrical activation of muscles; measurement of electro-dermal activity (EDA), which is directly related to physical arousal; electroencephalography (EEG), the recording of the brain’s spontaneous electrical activity over a short period of time; and near-infrared spectroscopy (fNIR), a non-invasive technique for measuring brain activation [29].
As mentioned earlier, assessing gameplay experience is still a challenging task as player enjoyment is determined by many complex and vaguely defined experiential phenomena. Researchers are still trying to reach a scientifically grounded definition of gameplay experience and develop valid measurement methods.

Assessment of the learning effectiveness is a fundamental issue in serious game and VLE research. Current methods for evaluating the learning effect include: performance measures, self-reports, comparison of pre- and post-test results, and comparison with traditional instructional methods. Performance in a serious game/VLE as a level of success has been considered as an acceptable measurement of learning under the assumption that high performers also learn more. “However, studies on the relationship between performance and learning have demonstrated that the two variables are not correlated and thus performance should not be used to evaluate simulations for learning. Learning involves successful internalization of knowledge and skills which may not manifest in successful performance” [11]. Self-reports on the perception of learning have been commonly used in game/VLE evaluations, however their validity is questionable due to the fact that learners may not always know of what they have learned or may not be able to express it [30]. Pretest-posttest designs are widely used for the purpose of comparing groups and/or measuring change resulting from experimental treatments. A more extensive approach splits up the test group into an experimental group that uses the game/VLE, and a control group that is taught using traditional instruction methods. Various statistical methods are then used in comparing groups with pretest and posttest data, such as analysis of variance (ANOVA) on the gain scores, analysis of covariance, ANOVA on residual scores, and repeated measures ANOVA.

Comparison of serious games/VLEs to traditional teaching methods can be difficult as the two approaches may differ significantly in their instructional objectives, and therefore may not be easily comparable. “Recent studies question the validity of comparing the two teaching methods which are different in nature due to the involved activities and abilities they are intended to support” [11]. Furthermore, the pre-post tests are often unable to capture and measure a broad range of higher-order thinking abilities and applied skills [31].

4. A REVIEW OF TWO PROJECTS: EVALUATION METHODS, FINDINGS, AND LESSONS LEARNED

4.1 SMILE (Science and Math in an Immersive Learning Environment): Evaluation of Usability and “Fun”

SMILE is an immersive learning environment that employs a fantasy 3D virtual world to engage deaf and hearing children in math- and science-based educational tasks. It includes an imaginary town populated by fantasy 3D avatars
that communicate with the participant in written and spoken English and American Sign Language (ASL). The user can explore the town, enter buildings, select and manipulate objects, and interact with the virtual characters. In each building the participant learns specific math/science concepts by performing hands-on activities developed in collaboration with elementary school educators (including deaf educators), and in alignment with standard math/science curriculum. The application is designed primarily for display in stationary projection systems (e.g., FLEX [32]), but can also be viewed on a single screen immersive portable system, or through a stereoscopic head-mounted display unit. Children travel through the virtual world using a 6-dof (degrees of freedom) wand, a chair-based interface (currently under development), or a dance mat, and can grasp and release objects using the wand or a pair of pinch gloves. Signed communication with the virtual characters is accomplished through a basic gesture control system that allows for input and recognition of a limited number of ASL signs. A detailed description of SMILE can be found in [33, 34]; Figure 1 shows one of the subjects interacting with SMILE in the FLEX, and a screenshot of the virtual world.

SMILE was developed using an iterative design and evaluation approach that included two forms of evaluation: formative and summative. The formative evaluation focused on the design features of the application (i.e., usability and fun, quality of the visual representation, and quality of the signing motion), the summative evaluation tested the efficacy of using a computer animated 3D bilingual immersive environment for teaching math and science concepts to deaf and hearing children in grades K-5. In this article we describe the formative evaluation with experts and target users.

To evaluate usability and “fun” (e.g., player enjoyment) of SMILE we used an evaluation approach that draws from research by Gabbard et al. [35] and Read et al. [36]; it included:

Figure 1. Child interacting with SMILE in the FLEX (left); screenshot of the city of “Smileville” (right).
1. user task analysis (with consequent development of representative user tasks);
2. expert guidelines-based evaluation;
3. formative evaluation with target users; and
4. iterative intervention refinement.

The approach used was specific to our application but could be generalized and used for assessment of similar educational VE.

(1) User Task Analysis

A user task analysis is the process of identifying a complete set of tasks and methods required to use a system, following a formal methodology, and generating critical information used throughout the intervention development life cycle. For SMILE, we created a series of complex user tasks scenarios (e.g., construction of new objects—we focused on the cake baking activity), and sub-tasks, such as navigation, object selection, object manipulation, data input, etc.

(2) Expert-Panel Based Evaluation

The expert panel-based evaluation aimed to assess:

1. usability of SMILE;
2. quality of 3D graphics and animation; and
3. quality of the signing motion.

The evaluation was carried out by a panel of experts consisting of seven individuals: three experts in VR application development, two experts in 3D modeling and animation, and two experts in American Sign Language. Each expert was asked to perform an analytical evaluation of the elements of the application that pertained to his/her area of expertise. The goal of the analytical evaluation was to let the experts perform a variety of user tasks in the VE, identify potential problems, and make recommendations to improve the design. The three experts in VR application development assessed the usability of the program by determining what usability design guidelines it violates and supports. The experts in 3D modeling and animation were given a questionnaire with questions focusing on the quality of the visual representation of the virtual world; the experts in ASL were given a similar questionnaire with questions on the quality of the signing motion. The evaluators used a 5-point Likert scale for rating the response to each question and used comment boxes to provide additional feedback.

Overall, the application was found easy to use and all evaluators were able to complete the users’ tasks without difficulty. However two main usability problems were identified:
1. poor selection feedback (violation of Gabbard’s [37] usability guideline “Supply users with appropriate selection feedback”); and
2. absence of interface query support for users).

At the time of the experiment, participants did not receive any feedback upon selection, thus they were not able to ascertain which, if any, objects were selected until they moved them. In the current implementation, selection is identified in the form of highlighting and marquees around objects and different query methods (textual and graphical) have been included.

System performance (i.e., total latency, display update rate, and tracking performance, all elements of the visual representation) and signing motion were given high scores by the experts. In particular, visual design and signed animation received very positive rating and, therefore, recommendations for improvement were not necessary.

(3) Formative Evaluation with Target Users

The goal of the formative evaluation was to assess fun and usability with a group of potential users. The group of subjects included 13 children ages 6½-10; 5 children were deaf ASL signers. The minimum number of participants was determined using the Nielsen and Landauer formula [38] based on the probabilistic Poisson model. Nielsen argues that, for web applications, 15 users would find all usability problems and 5 participants would reveal 80% of the usability findings. Lewis [39] supports Nielsen but notes that, for products with high usability, a sample of 10 or more participants is recommended. For this study it was determined that testing with 12 or more participants should provide meaningful results for usability and fun.

Usability of a game and fun of its use are two closely related concepts. According to the ISO 9241-11 definition, usability is derived from three independent measures: efficiency, effectiveness, and user satisfaction. Even if systems that are designed for children do not neatly fall into this usability paradigm, it has been postulated that fun is one manifestation of what adults call satisfaction [40]. Fun, however, is much more than a child’s expression of satisfaction and, recently, a few researchers have attempted to better define and measure fun. Our evaluation of fun was based on the three dimensions of fun proposed by Read et al. [36]: expectations, engagement, and endurability (or “returnance,” i.e., the desire to do again an activity that has been enjoyable). The evaluation methodology used to measure fun included:

1. ranking and rating exercises;
2. observation; and
3. think aloud protocol.

Ranking and rating exercises were used primarily to measure endurability and expectations, while observation and think aloud protocol were used to assess
engagement and usability. In general, ranking, rating, and observation have been proven to be more reliable than children’s responses to questions on whether or not they liked something. Children are eager to please adults and may tell the evaluators that they liked the program just to make them happy [40]. All ranking and rating exercises were designed to be age appropriate: for instance, children rated activities and elements of the game using a scale with smiling and frowning faces. This scale, represented in Figure 2, is a discrete variation of the “Funometer” proposed by Read et al. [36].

Research shows that children are able to respond more reliably to a pictorial representation with meaningful anchors (smiling and sad faces) rather than to a Likert-type scale [40].

In order to assess engagement, all testing sessions were recorded on video and the footage was scored with reference to a set of positive and negative instantiations. Positive instantiations that were looked for were smiles, laughing, and positive vocalization. Negative instantiations were signs of boredom (yawns), signs of frustration (sighing and shrugs and stomping feet), and negative vocalization.

The usability evaluation included three commonly used evaluation techniques:

1. measurement of key usability factors such as learning time, time to complete a task, number of errors while performing a task, and completion or non-completion of a task;
2. think aloud protocol and “critical incidents,” i.e., problems encountered that affect task flow and performance; and
3. observation.

Since this evaluation focused exclusively on fun and usability, not on learning and knowledge acquisition, all participants were given a pre-test before the beginning of the interactive session. The objective of the pre-test was to ensure that all subjects had the basic mathematics skills necessary to complete the activities. The pre-test also included questions relative to the participant’s expectations of the game, and questions pertaining familiarity with computer games. The evaluation instrument including the pre-test and the rating/ranking exercises is available at http://www2.tech.purdue.edu/cgt/i3/SMILE/demos.html

Figure 2. The pictorial rating scale.
All subjects passed the pre-test (the mean was 23.3 out of 24) and therefore were able to proceed to the interactive testing session. The mean learning time, that is, the mean time necessary to perform a sequence of basic operations (pick up an object, move it and place it inside another object, and then put it on top of another object) was 1:15 seconds. The mean time to complete the cake-baking activity was 5:28 seconds. Eleven subjects were able to complete the activity on their first attempt, one subject completed it on her second attempt, and one subject was unable to bring it to completion and was stopped on the 10-minute mark. Learning time and time to complete the cake were both considered satisfactory. The data comparison between mean learning times and task completion times with video game familiarity was inconclusive. No meaningful correlation between frequency with which children play video games and speed/ability to perform the SMILE activities was identified.

Figure 3 summarizes the results relative to the participants’ expectations. The diagram shows the differences (in mean values) between the predicted and reported experience. This difference is an effective indicator of how enjoyable the experience has been [36]. As mentioned previously, children rated their responses using the scale represented in Figure 2. To calculate the mean values, the happiest face was assigned a value equal to 4 and the saddest face was assigned a value equal to 1. Results show that even if the children had very high

![Figure 3. Diagram showing the differences between the predicted and reported experience.](image-url)
expectations, the reported experience surpassed them. The game was perceived more fun and easier than expected, and slightly more challenging.

Table 1 summarizes the results of the “Again-Again Table” used to measure “returnance.” The table reveals that the activities the children enjoyed the most were the construction of the new object (i.e., the cake) and playing the entire game.

Observation and think aloud protocol showed that other activities the participants found “very fun” were “walking through objects,” “throwing objects,” “opening doors,” and “watching things that move.” As far as usability, children did not appear to have major problems with travel, selection, and manipulation tasks. We noticed a few signs of frustration and comments such as “some of the ingredients are really hard to pick up” and “the scale is hard to read.” One problem to be addressed in future iterations is the size of the 3D shutter glasses. Children kept losing the goggles during interaction and were constantly adjusting them on their noses. The research team is currently researching different solutions such as using small size 3D glasses for children, or coupling the goggles with a head band.

As far as engagement, the majority of the subjects appeared to be very focused on the tasks. Positive comments included: “this is awesome because you feel like you are really in a bakery”; “. . . this game is more exciting than a video game because you don’t see anything around you . . . and you are really inside the building putting a cake in the oven.” Many positive signs were observed, such as laughing, smiling, bouncing in excitement, and “wow” sounds.

<table>
<thead>
<tr>
<th>Question</th>
<th>Mean (1-4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Watch the beginning cartoon again?</td>
<td>3.21</td>
</tr>
<tr>
<td>Watch the ending cartoon again?</td>
<td>3.14</td>
</tr>
<tr>
<td>Make the cake again?</td>
<td>4.0</td>
</tr>
<tr>
<td>Explore the bakery room again?</td>
<td>3.71</td>
</tr>
<tr>
<td>Travel through town again?</td>
<td>3.92</td>
</tr>
<tr>
<td>Watch the amazing baker’s machine in motion again?</td>
<td>3.94</td>
</tr>
<tr>
<td>Play the entire game again?</td>
<td>4.00</td>
</tr>
</tbody>
</table>
4. Challenges and Lessons Learned

Evaluating the usability of SMILE presented a challenge primarily because of the lack of clear heuristics for the ideal design of VEs. Usability evaluation of Virtual Environments (VE) is still a fairly new area of research; only few researchers have recently started to recognize the importance of VE design and evaluation and are keen on developing new effective assessment and design approaches. In order to define a valid set of heuristics, we did an extensive literature search and contacted several VR researchers who were working on similar projects. The set of usability design guidelines used in this study was put together based on various researchers’ recommendations and on previous work by Nielson and Molich [41], Nielson [42], and Gabbard [37]. Although not exhaustive, we believe that this set of guidelines was sufficient to support the expert panel based evaluation of SMILE and we plan to use these guidelines again for future evaluation of similar projects.

Evaluation of user enjoyment (fun) was another challenging task for two reasons:

1. the constraints of the testing environment which limited our ability to observe the subjects; and
2. the nature of the target audience (deaf and hearing young children).

Testing took place in the FLEX, lights were dim, and subjects were wearing shutter glasses while holding a control wand for navigation and object selection/manipulation. Because of the lighting conditions and the goggles, we were not able to observe the subjects’ facial expressions (the eyes in particular) during interaction. In addition, the wand limited the subjects’ body motions and gestures; therefore, we could not rely on the participants’ upper body movements/poses as indicators of excitement or frustration. These constraints are common to any experiment that takes place in a CAVE-like environment. A possible solution to the problem is to encourage the subjects to talk as much as possible during interaction; then vocalizations and mouth expressions (i.e., smiles, pouts, etc.) can be used as primary indicators of emotional state. In the case of SMILE, the problem was particularly challenging, as five subjects were deaf. To compensate for the lack of vocalizations, we conducted brief interviews with the deaf children (with the help of a sign language interpreter) after the experiment. As mentioned in section 3, some of the most common subjective methods for assessing user enjoyment include self-reports, usually through questionnaires and surveys. In the case of children, surveys with open-ended questions do not work well. Often, children are not able to articulate and put in writing how they feel about something. In addition, they are eager to please adults and may write that they liked the program just to make the evaluators happy [40]. We found that using age-appropriate ranking and rating exercises with pictorial rating scales was an effective way to collect valid feedback from the children.
4.2 The Sustainability Challenge—A Serious Game for Learning/Teaching Sustainable Building Design: Evaluation of Students’ Learning

The “Sustainability Challenge” is a serious game for learning sustainable building design principles and practices aimed at undergraduate students enrolled in Civil Engineering, Architecture, and Building Construction Management programs. It is a role-playing serious game (RPG) in which the student plays as the main protagonist. The protagonist’s mission is to travel to various U.S. states and in each state help designers, constructors, and building owners increase their buildings’ environmental and economic performance in the areas of energy conservation, waste reduction, water efficiency, cleaner transportation, etc. The player chooses to play as one of four different characters (two males and two females) and travels through six game levels: Office Buildings, Homes, Schools, Neighborhood Development, Commercial Buildings, Healthcare Buildings. Buildings are located in different states in the United States, and based on their geographical location they present different sustainability challenges.

The game follows the organization of the Leadership in Energy and Environmental Design (LEED) rating system and includes five modules:

1. Homes;
2. Building Design and Construction;
3. Interior Design and Construction;
4. Operation and Maintenance; and
5. Neighborhood development.

To date, the game includes Module # 2, Building Design and Construction, which targets new construction and major renovations, and focuses on seven design criteria: (2.1) Sustainable sites, (2.2) Water Efficiency, (2.3) Energy and Atmosphere, (2.4) Material and Resources, (2.5) Indoor Environmental quality, (2.6) Innovation in Design, and (2.7) Regional Priority. Figures 4 and 5 show screenshots of the game. The game challenges students to identify problems, make decisions and observe the effect of those decisions. To solve problems and complete the game activities, students are encouraged to apply the three sustainability concepts Reduce, Reuse, and Recycle in each one of the specific levels. The game is not meant to replace traditional lectures; it is intended to be used as a review, practice and assessment tool. A detailed description of the game can be found in Dib and Adamo-Villani [43].

Like SMILE, the Sustainability Challenge game was developed using an evaluation methodology that included formative and summative assessments. In this article we describe the summative evaluation with a group of college students.

The goal of the summative evaluation was to provide an indication of whether using the serious game affects students’ content learning. Learning is a complex process that includes cognitive, metacognitive, and motivational components [44].
Figure 4. Two screenshots of the game: site selection criterion (a); Indoor Environmental Quality criterion (b).
Our experimental study focused on the cognitive aspects, for example, descriptive (or declarative) knowledge, or the ability to memorize and recall information; and procedural knowledge, or the ability to apply acquired knowledge to perform specific tasks [45]. In particular, the study aimed to assess the influence of the serious game on student ability to:

1. demonstrate knowledge and understanding of building sustainability principles and practices and LEED categories and rating system (specifically, water efficiency);
2. perform a LEED-based analysis of existing buildings (e.g., calculate existing water consumption and assess adherence to best building sustainability practices); and
3. perform a LEED-based analysis of new designs (e.g., estimate building water consumption and assess adherence to best building sustainability practices).

Figure 5. Screenshots of level 5 (Commercial Buildings—New Design). The player travels to an imaginary city and visits the site where the new restaurant will be located (top right); the player is provided with the design of the new building (top left) and uses the LEED tables and formulas to calculate the building water consumption and improve water efficiency (bottom).
We measured these three learning goals using pre- and post-educational intervention competency testing. The testing instrument was adapted from materials designed by the U.S. Green Building Council (USGBC) to help individuals as they prepare to either register a building for LEED certification or prepare to take the certification exam as a LEED Accredited Professional (LEED AP). The primary resources for these materials are the LEED Online Documentation Requirements and review and practice questions from the USGBC LEED AP Building Design + Construction Study Guide. The pre/posttest used in the study included 35 questions: 26 questions tested the students’ descriptive knowledge, 9 questions tested their procedural knowledge.

**Subjects**

Participants included 42 students from the Building Construction Management (BCM) department at Purdue University.

**Procedure**

All students received a traditional lecture on water efficiency. After the lecture, the students were administered a test to assess their knowledge of the concepts (pre-test). After the pre-test, a randomized complete block design was used to divide the subjects into two groups with similar pre-knowledge. In other words, we used the pre-test scores to group individuals in terms of pre-knowledge and then made sure these groups of individuals were equally assigned to the two intervention groups; for example, group A (control group), traditional readings/assignments, versus group B (experimental group) game.

Two weeks after the pre-test, group A (control) worked in the lab for 1 hour using readings and traditional assignments on water efficiency; group B (experimental) worked in the lab for 1 hour using the serious game in place of the readings and traditional assignments. All subjects in group B were familiar with the game user interface. Both groups were tested for water efficiency competency after the two educational interventions (game versus traditional readings/assignments). The post-test was the same as the pre-test. The difference in scores provided an indication of:

1. whether playing the game improves learning of water efficiency;
2. whether playing the game leads to greater learning gains than traditional learning methods (i.e., readings/traditional assignments); and
3. which aspects of learning (declarative knowledge versus procedural knowledge) benefit the most from using the game.

**Findings**

Results show that playing the game led to an increase in subjects’ content learning by 26%. In particular, students who played the game increased their declarative knowledge by 22% and procedural knowledge by 37%. Table 2 shows
a summary of the results. Findings also show that there are differences in learning gains between Group A and Group B. Figure 6 visualizes these differences. Overall, playing the game led to higher learning gains than traditional learning methods. Three two-sample t-tests were performed to determine if the difference in total learning gains, the difference in procedural knowledge gains, and the difference in declarative knowledge gains between the control and experimental group were statistically significant. Results of the statistical analysis show that the difference in total learning gains between the control and experimental group is not statistically significant ($P$-value = 0.093; $M$ (group A) = 22.13; $SD$ (group A) = 4.87; $M$ (group B) = 26.42; $SD$ (group B) = 3.53); the difference in procedural knowledge gains is statistically significant ($P$-value = 0.025; $M$ (group A) = 26.23; $SD$ (group A) = 7.56; $M$ (group B) = 37.38; $SD$ (group B) = 7.60); the difference in declarative knowledge is not statistically significant ($P$-value = 0.530; $M$ (group A) = 20.98; $SD$ (group A) = 6.41; $M$ (group B) = 22.27; $SD$ (group B) = 5.68).

Overall, these findings suggest that the game has potential for being an effective learning tool. The statistically significant difference in procedural knowledge gains between the two groups confirmed the educational value of serious games in increasing students’ problem solving skills and creative thinking ability, which are fundamental for solving real world challenges. We believe that the difference in total learning gains between the two groups was not statistically significant, although close to significance, due to the small sample size. An ideal sample size, calculated using an online power/sample calculator, would require at least 60 subjects. The intention of this small scale empirical study was to provide preliminary evidence that will lead to a more rigorous evaluation of the effectiveness of the serious game with a larger sample population in future research.

<table>
<thead>
<tr>
<th></th>
<th>Group A (control)</th>
<th>Group B (experimental)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pre-test</td>
<td>Post-test</td>
</tr>
<tr>
<td>Mean score on Declarative Knowledge part</td>
<td>66.92</td>
<td>80.77</td>
</tr>
<tr>
<td>Mean score on Procedural Knowledge part</td>
<td>68.89</td>
<td>86.67</td>
</tr>
<tr>
<td>Mean total score</td>
<td>67.43</td>
<td>82.29</td>
</tr>
</tbody>
</table>
Challenges and Lessons Learned

The main challenge faced during the evaluation of students’ learning was to define clear learning objectives, and subsequently develop a valid test that would reliably measure these objectives. As mentioned previously, serious games target many aspects of learning, especially problem-solving skills, creative thinking abilities, and the ability to apply theoretical knowledge in concrete scenarios. However, the majority of the experimental studies found in the literature have aimed at measuring only descriptive knowledge gains, for example, the ability to recall information. The goal of our evaluation was to measure the impact of the game on both declarative and procedural knowledge. In order to define meaningful learning objectives and a valid evaluation instrument, we worked extensively with content experts, LEED accredited professionals, education researchers, and game developers. Although the process was time consuming and costly, we believe it was worthwhile as it allowed us to gather significant results on the impact of the game on different aspects of students learning. We encourage other researchers to adopt a similar approach and devote the necessary effort and time to developing effective learning goals and valid measurement instruments.

A common flaw of many studies on the effect of games on students’ learning is the small sample size, and therefore the difficulty in generalizing the results. Our study faced the same challenge as we had a relatively small pool of subjects (42). Because of the limited number of participants, we cannot generalize the
results and we can only claim that the game shows promise of being an effective instructional tool. We are conducting additional studies with larger pools of participants and in different settings in order to build stronger evidence of the learning effectiveness of the game.

5. CONCLUSION

In this article we have presented a review of current methodologies for evaluating serious games and VLEs, and reported the findings from the evaluation of two projects, a serious game for teaching/learning building sustainability and a VLE for K-5 math and science education. The contribution of the article is twofold:

1. it informs other researchers about the state of the art in educational technologies evaluation, related challenges, and unsolved problems; and
2. by presenting the results and lessons learned from two recently developed projects, it provides guidelines for future development and evaluation of similar educational technologies.

6. REFERENCES

techniques to measure user experience with entertainment technologies. Behaviour
Information Technology, 25(2), 141-158.
for three-dimensional virtual environments. Proceedings of the 2010 Symposium on
EyeTracking Research Applications (pp. 109-112).
experience in a serious gaming context. International Journal of Computer Science
in Sport (Special Edition), 9 (Edition 2).
of appraisals of students faced with curricular tasks. Educational and Psychological
Measurement, 63(1), 96-111.
Research on Technology in Education, 42(3), 309-328.
32. Fakespace Systems, FLEX. Available at http://vislab.jsums.edu/VIS/Fakespace-
FLEX-reFlex.pdf
ogies for accessible math and science education. IEEE Multimedia—Special Issue on
Accessibility, pp. 38-46.
deaf and hearing children. Proceedings of SIGGRAPH 2007—Educators, August 5-10,
centered design and evaluation of virtual environments. IEEE Computer Graphics
and Applications, pp. 51-59.
expectations: Measuring children’s fun. Interaction Design and Children. Eindhoven,
The Netherlands: Shaker Publishing.
Master’s thesis, Dept. of Computer Science and Applications, Virginia Polytechnic
Institute and State University.
problems. Proceedings of the SIGCHI conference on human factors in computing
systems, Amsterdam (pp. 203-213).
Human Factors, 36(2), 368-378.
with children. Proceeding of the 2004 conference on interaction design and children:
Building a community. Maryland, (pp. 49-56).
of the ACM CHI ’90 Conference, Seattle, WA, April 1-5, 1990 (pp. 249-256).
for teaching and learning building sustainability. ASCE Journal of Computing in
Civil Engineering (in review).

QA: Update?

Direct reprint requests to:
Nicoletta Adamo-Villani
Purdue University
Dept. of Computer Graphics Technology
401 N. Grant St.
West Lafayette, IN 47907
e-mail: nadamovi@purdue.edu