User Assessment in Serious Games and Technology-Enhanced Learning

Guest Editors: Francesco Bellotti, Bill Kapralos, Kiju Lee, and Pablo Moreno-Ger
User Assessment in Serious Games and Technology-Enhanced Learning
User Assessment in Serious Games and Technology-Enhanced Learning

Guest Editors: Francesco Bellotti, Bill Kapralos, Kiju Lee, and Pablo Moreno-Ger
Editorial Board

Armando Bennet Barreto, USA
Marina Bers, USA
MarkBillinghurst, New Zealand
Frank Biocca, USA
Cathy Bodine, USA
Michael Boronowsky, Germany

Caroline G. L. Cao, USA
Eva Cerezo, Spain
Adrian David Cheok, Singapore
Kerstin S. Eklundh, Sweden
Holger Kenn, Germany
Kiyoshi Kiyokawa, Japan

Richard Kline, USA
Paul Lukowicz, Germany
Ian Oakley, Portugal
Francis Quek, USA
Anthony Savidis, Greece
Manfred Tscheligi, Austria
Contents

**User Assessment in Serious Games and Technology-Enhanced Learning**, Francesco Bellotti, Bill Kapralos, Kiju Lee, and Pablo Moreno-Ger
Volume 2013, Article ID 120791, 2 pages

**Assessment in and of Serious Games: An Overview**, Francesco Bellotti, Bill Kapralos, Kiju Lee, Pablo Moreno-Ger, and Riccardo Berta
Volume 2013, Article ID 136864, 11 pages

**A Game-Based Virtualized Reality Approach for Simultaneous Rehabilitation of Motor Skill and Confidence**, Alasdair G. Thin
Volume 2012, Article ID 213143, 6 pages

**Static and Dynamic User Portraits**, Ko-Hsun Huang, Yi-Shin Deng, and Ming-Chuen Chuang
Volume 2012, Article ID 123725, 16 pages

**Usability Testing for Serious Games: Making Informed Design Decisions with User Data**, Pablo Moreno-Ger, Javier Torrente, Yichuan Grace Hsieh, and William T. Lester
Volume 2012, Article ID 369637, 13 pages

**Assessment of Learners’ Motivation during Interactions with Serious Games: A Study of Some Motivational Strategies in Food-Force**, Lotfi Derbali and Claude Frasson
Volume 2012, Article ID 624538, 15 pages

**Psychophysiology to Assess Impact of Varying Levels of Simulation Fidelity in a Threat Environment**, Thomas D. Parsons, Albert A. Rizzo, Christopher G. Courtney, and Michael E. Dawson
Volume 2012, Article ID 831959, 9 pages
Games have always had a great relevance for human-computer interaction, since they represent a computer application sector which involves a wide enthusiastic audience and has opened the way to new hardware and software tools (e.g., graphic boards, pointing and gesture-based devices, virtual worlds) that have then become common also in other application domains.

Nowadays, an emergent trend concerns serious games (SGs), games that have been designed and are used with a different purpose than pure entertainment [1–3]. SGs are becoming ever more appealing for education and training, and several research coordination initiatives have been undertaken recently, such as the Games and Learning Alliance (http://www.galanoe.eu/) (GaLA) EU FP7 Network of Excellence (NoE) [4] and the Serious Games Network (http://www.seriousgamesnet.eu/) (Segan) EU Lifelong Learning Programme (LLP) NoE. The main goal of these networks is to study SGs so that they can become an effective and reliable tool for education and corporate training. To this end, it is necessary to substantiate the growing interest and appeal of SGs within the scientific and education communities through the definition of proper principles and concrete tools, methodologies, and practices for design and deployment of SGs [5, 6].

Proper assessment of user performance and status is key for SGs with educational goals [7, 8]. An effective application of SGs for education and training demands appropriate metrics, tools, and techniques for user evaluation, in particular by measuring elements such as learning outcomes, engagement or gameplay performance [9, 10]. These are data that are considered as necessary in order to show to the large public of families, teachers, instructors, and stakeholders the validity of SGs in education and training. Devices like stereo cameras, eye trackers, physiological, and neural response sensors, now available at reasonable prices, not only support innovative interactions, but also present opportunities to new user monitoring and evaluation.

Due to the complexity of human nature and individual differences, objective and systematic assessment of human behavior and performance remains highly difficult. In addition, data analysis and evaluation methods for technology-assisted learning and assessment are still underdeveloped because of different perspectives in evaluation. Thus, development of systems and tools able to support provision of effective feedback is a major requirement for a new generation of SGs. Breakthroughs in this area can be made by advancing issues including, but not limited to, an efficient and easy-to-use user interface; effective data management; data analyses...
methods; sensor data fusion and integration; user feedback mechanism.

The goal of this special issue is to investigate user assessment in SGs, showing the state of the art and proposing advancements in some specific topics, with a special perspective on usability and usefulness for learning.

Our review process has led to the selection of six papers by authors of various cultural background and geographical provenance. L. Derbali and C. Frasson investigated motivational strategies (i.e., the use of game elements for providing motivational supports to users) and assessment of learners’ motivation during serious games. They propose the use of physiological sensors—in particular they highlight the importance of monitoring the neural activity—to collect data for a theoretical model of motivation.

P. Moreno-Ger et al. propose a new method of usability assessment adhoc designed for SGs, which exploits large amounts of recorded gameplay data and also provides SG designers with suggestions on how to improve their design.

T. D. Parsons et al. investigated psychophysiological responses of test users in order to understand whether highly immersive virtual environments result in increased sensory arousal, obtaining results that suggest that higher fidelity scenarios have great efficacy related to sensory arousal.

K.-H. Huang et al. propose guidelines and research approaches for developing useful personas for large-scale service design and social interaction design. This is a useful advancement in user modeling and profiling, which is a fundamental step for user behaviors prediction.

A.G. Thin presents a real-world case study concerning the assessment of a virtualized-reality-based game for simultaneous rehabilitation of motor skill and confidence. We believe that this study is an interesting starting point for the development of a new generation of games for health that can be integrated in clinical environments and cases.

The special issue includes also a more general overview paper, written by the guest editors, which discusses the state of the art on assessment of SGs and in SGs. After a review of the literature on the educational effectiveness of SGs, the paper addresses how to assess the learning impact of SGs and methods for competence and skill assessment. Finally, it suggests directions for future research.

The guest editors are proud of presenting a balanced mix of papers, especially in terms of perspectives, approaches, and addressed topics. The selected papers show that the field of user assessment in serious games is advancing especially in the direction of a continuous in-game embedded (stealth) assessment, also exploiting neurophysiological signals. We believe that this, together with development of SG-tailored learning analytics, represents major innovation modules for new generation games, and innovative human-computer interaction solutions are to be developed in order to allow a comfortable yet effective acquisition of inputs and provision of feedback.

References


---

Francesco Bellotti
Bill Kapralos
Kiju Lee
Pablo Moreno-Ger
Assessment in and of Serious Games: An Overview

Francesco Bellotti, 1 Bill Kapralos, 2 Kiju Lee, 3 Pablo Moreno-Ger, 4 and Riccardo Berta 1

1 Department of Naval, Electric, Electronic and Telecommunications Engineering, University of Genoa,
Via all’Opera Pia 11/a, 16145 Genoa, Italy
2 Faculty of Business and Information Technology, University of Ontario Institute of Technology, 2000 Simcoe Street North, Oshawa, Canada L1H 7K4
3 Department of Mechanical and Aerospace Engineering, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, OH 44106, USA
4 Faculty of Computer Science, Universidad Complutense de Madrid, Ciudad Universitaria, 28040 Madrid, Spain

Correspondence should be addressed to Francesco Bellotti; franz@elios.unige.it

Received 18 May 2012; Revised 22 January 2013; Accepted 6 February 2013

Academic Editor: Armando Bennet Barreto

Copyright © 2013 Francesco Bellotti et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

There is a consensus that serious games have a significant potential as a tool for instruction. 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, and many studies lack a rigorous assessment. 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, and so forth. 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 [1, 2]. 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 [3, 4].

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, 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 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., [5, 6]).

A number of studies have questioned the effectiveness of game-based learning (e.g., [7–9]). 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 [10] 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 [1] 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 hereinafter—do produce learning, while engaging players [12].

One of the foundational reviews of the effectiveness of gaming was performed by Livingston et al. [13], when they evaluated seven years of research and over 150 studies to examine the effectiveness of gaming. Their results were later mirrored by Chin et al. [14], and they concluded that “simulation games” are able to teach factual information although they are not necessarily more effective than other methods of instruction [13, 14]. 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 [13, 14].

More recently, Connolly et al. [15] 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.

Serious games look particularly effective in some specific application fields. 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 [16]. More recently, Cole et al. [17] showed that activation of brain circuits involved in positive motivation during Re-Mission gameplay 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 [18].

Bellotti et al. [19] 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 aforementioned results show that serious games can be an effective tool to complement the educational instruments available to teachers, in particular for spurring user motivation [20] and for achieving learning goals at the lower levels in the Bloom’s taxonomy [15]. 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 [21] or as Michael and Chen [22] 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 [14]. Assessment is a complement to purpose, and it is commonly employed by learning institutions, regardless of the teaching methods used, whether or not their students actually learn [7]. 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 [8, 23]. Part of this difficulty stems from the open-ended nature inherent in video games making it difficult to collect data [14]. 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? [21].

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 [24]. 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 [6], in particular through appropriate user feedback.

Considering the specific serious game domain, Michael and Chen [22] 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 assessments, 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 [22]. 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 [25]. 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 [26] and simulations in general, summative assessment is commonly accomplished with the use of pre- and posttesting, a common approach in educational research [27]. The pre- and posttest 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 [28]. 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 posttest, and significant differences across the test scores are attributed to the “treatment” (the serious game) [27]. The main problem with the pre- and posttest experimental design is that it is impossible to determine whether the act of pretesting 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, then randomization is not possible) [29].

The most common method of postassessment 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 [6]. This method was used by Allen et al. [30] in the form of questionnaires before and after playing their game, Infinitimains 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 posttesting assessment was used to evaluate 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) [27].
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 [31], and with respect to students, strong engagement has been associated with academic achievement [6], 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 [32] and the Game Experience Questionnaire [33].

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 [34]. Incorporating the concept of flow in computer games as a model for evaluating player enjoyment has been a focus of interesting studies [35, 36] and forms the basis of EGameFlow, a scale that was specifically developed to measure a learner's enjoyment of e-learning games [37]. 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 [38–41]) and flow [42]. Common physiological measures include the following [41, 43].

(i) 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 [40].

(ii) Cardiovascular measures such as the interbeat 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 [40]. Although cardiac measures have been successfully used in a number of game studies, interpreting as described by Kivikangas et al. [40], interpreting the relevance of the resulting measurements within a game context is difficult and challenging.

(iii) 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.

(iv) Electroencephalography (EEG) for measuring the electrical activity along the scalp and, more specifically, 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 [44] describe a study where the EEG of players plays 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 [41].

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 applications [43]. 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 Kivikangas et al. [40] and Nacke [41].

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 gameplay, 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 consist 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 [23, 25, 45].

3.3. Assessing Entertainment. As mentioned in Section 1, 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,
investigated insignificant subsequent studies [35,36].

Incorporating flow in computer games as a model of flow, based on Csikszentmihalyi’s foundational concepts—challenge, curiosity, and fantasy—and the theory of intrinsic qualitative factors for engaging gameplay [47]—the identification of two major lines: Malone’s principles of component (e.g., entertainment) and (ii) qualitative approaches [46]. 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 gameplay [47]—namely, challenge, curiosity, and fantasy—and the theory of flow, based on Csikszentmihalyi’s foundational concepts [34]. Incorporating flow in computer games as a model for evaluating player enjoyment has been proposed and investigated in significant subsequent studies [35,36].

In contrast, quantitative approaches attempt to formulate entertainment using mathematical models, which yield reliable numerical values for fun, excitement, or achievement. However, such approaches are usually limited in their scope. For instance, Iida et al. [48] focus on variants of chess games, while Yannakakis and Hallam [46] 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 [49]. That work exploited the Repertory Grid Technique (RGT) methodology [50], 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 simplifying 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 [51]. Web-based assessments are useful because they decrease class time used for assessment and because multimedia 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. [52] recommend that pedagogic consideration should be given to the choice, variety, and level of difficulty of e-Assessments offered to students. Hewson [53] provides preliminary support for the validity of online assessment methods. Guzmán et al. [54] 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 e-Assessment, which we describe hereinafter.

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

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 [56, 57]. 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.
Assess By Computer (ABC) is also designed for flexible computer-based assessment using a variety of question formats [61]. 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 [62].

Short Answer Marking Engine (SAME) is a software system that can automatically mark short answers in a free text form [63]. 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 [64]. Noorbehbahani and Kardan [65] 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 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 [66]. CRS has been increasingly accepted in educational environments from K12 to higher education and also in informal learning environments [67]. 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 and 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 [69]. 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 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 [70]. Applications can be created using XML (extensible markup language) or higher level development tools including virtual learning environments (e.g., Blackboard, JLE ESSI, and Oracle iLearning), commercial assessment tools (e.g., Can Studios, Calypso from Experient e-Learning 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 [71]. 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 [72]. 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 and are easier to administer, and their results are readily interpretable [73]. However, there are limitations to such approaches including ineffective measurement of complex problem solving, communication, and reasoning skills [74, 75]. 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 [76, 77]. 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 [77]. 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 [78, 79]. 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 [78, 79]. 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 [22].

As pointed out by Becker and Parker [27], 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 [80, 81].

Incorporating in-game assessments takes us away from the predominant, classic form of assessment comprised of questionnaires, questions and answers, and so forth that usually interrupts and negatively affects the learning process [21] 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. The work of [21] 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” [82]. 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 and agriculture. 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 [21]. Integrating the assessment such that the player is unaware of it forms the basis of what Shute et al. [6] describe as stealth assessment. In this way, the player can concentrate solely on the game [83]. 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 [84].

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 [84]. 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 [84].

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 [85]. 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 preprogrammed 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 [86], the main objective for a player is to successfully interpret the environment and achieve the desired attitude towards him by Nonplayer 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 seriousness 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 nondigital form—thus, they were not called serious games) in business schools [87, 88]. In SimVenture, the target of the player is to manage a company, dealing with four major types of issues: production, organization, sales and 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 predefined 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, and 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 a 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 multiplayer 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 [22]. 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 Bloom's taxonomy [15].

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 gameplay sessions and keeping track of players' in-game performance. In-game assessment appears to be particularly suited and useful given that it 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" [89]. 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 [90]. 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 neurophysiological 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 a 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 [20]. 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 while favoring performance enhancement.

Conflict of Interests

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

Acknowledgments

The authors 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


Advances in Human-Computer Interaction


[58] Questionmark Corporation, “Questionmark Perception Measure Knowledge, Skills and Attitudes Securely for Certification,”
A Game-Based Virtualized Reality Approach for Simultaneous Rehabilitation of Motor Skill and Confidence

Alasdair G. Thin

School of Life Sciences, Heriot-Watt University, Edinburgh EH14 4AS, UK

Correspondence should be addressed to Alasdair G. Thin, a.g.thin@hw.ac.uk

Received 18 May 2012; Revised 24 September 2012; Accepted 8 October 2012

1. Introduction

Functional impairment of human motor function can arise due to a number of different causes including a variety of disease processes, physical trauma, and aging. The best treatment outcomes are seen when rehabilitation exercises are instituted early and in an intensive and repetitive manner in order to promote neural plasticity and muscle hypertrophy [1]. Given the often long and arduous nature of treatment programs requiring many thousands of exercise repetitions over many months, if not years, it is hardly surprising that patients commonly complain that the therapeutic exercises are repetitive and boring and this leads to poor compliance with the prescribed exercises and results in suboptimal treatment outcomes. Not only is this scenario likely to impact on a patient’s quality of life, but also it may ultimately result in a loss of their ability to live independently and necessitate long-term provision of care. Treatment programs are also very often resource intensive in terms of the time a physical therapist needs to devote to an individual patient and also the time spent traveling in connection with treatment. A range of technology-based solutions are therefore currently being actively investigated in terms of their potential for improving the efficiency and effectiveness of rehabilitation programs and to also increase the independence of patients and empower them to take control of their own treatment [2].

Virtual rehabilitation has been defined as “the combination of computers, special interfaces, and simulation exercises used to train patients in an engaging and motivating way.” [3]. A range of different systems developed so far include remote monitoring of therapeutic exercises enhanced by virtual reality (VR) [4, 5], finger strengthening and hand-eye coordination exercises and games using VR combined with a haptic glove [6, 7]. For the purposes of this paper the term “Game-based Virtualized Reality” (GBVR) is considered to include all forms of games that involve players physically interacting with virtual objects that only exist as a digital representation on a screen.

The potential for game-based approaches in rehabilitation practice to provide a more engaging and motivating
experience and that large numbers of game-based rehabilitation scenarios could be developed to provide greater realism and to correspond more closely to a wide range of everyday activities is now being more fully recognised [8–12]. Furthermore, there is now a growing evidence base to support their use in a variety of different rehabilitation applications including mobility and aerobic fitness [13–15], post-stroke rehabilitation of hand-arm function [16–18], balance [14, 19–21], pain distraction while undergoing treatment/therapy [22], and treatment of amblyopia (“lazy eye”) [23].

There is a long standing call to have rehabilitation programs focus simultaneously on improving both physical motor skill and confidence (i.e., self-efficacy) [24] as the importance of a person’s perception of their ability has long been recognized as an important determinant of physical performance [25]. Video games are a highly interactive and engaging form of entertainment [26] and incorporate clear goals, immediate feedback and rich visual and aural information [27]. It would therefore seem plausible that GBVR training could rise to this challenge and go a long way to meeting these dual treatment goals. This study was therefore undertaken to investigate if a rehabilitation training programme using a GBVR approach could simultaneously improve both motor skill and confidence. In order provide a stable and reliable experimental setting for both this study and future basic research into the development and refinement of GBVR, a nondominant hand motor deficit model was devised and will be described in subsequent sections.

2. Materials and Methods

2.1. Design. In order to be able to attribute any increase in motor skill and/or confidence to the GBVR rehabilitation training, a nondominant hand motor skill deficit model was conceived whereby each subject acted as their own control with their dominant and nondominant arms act as the control and intervention conditions, respectively.

2.2. Subjects. Nonambidextrous, young adults with no reported health issues were recruited from the student population at Heriot-Watt University. Potential subjects were asked if they had experience of playing racket sports and any reporting regular recreational participation or any competitive matches were excluded. The study was subject to local ethical committee approval and all subjects gave written, informed consent and underwent health screening. A total of 20 subjects were recruited although three subjects withdrew at various stages from the study. The 17 subjects (8 male) who completed the study ranged in age from 18 to 21 years with a mean (±SD) height 1.74 ± 0.09 m, weight 71.2 ± 10.4 kg, and a body mass index of 23.5 ± 2.6 kg/m². The study comprised a total of six sessions comprising an initial familiarisation session and then five sessions on consecutive days (Monday through to Friday). In order to minimise any effect of bias on the results, the subjects were unaware of the potential rehabilitation applications of the outcome of the study. Furthermore, the subjects were not given any feedback on their performance by the experimenters.

2.3. Familiarisation Session. In the familiarisation session subjects’ height and weight were measured using a portable stadiometer (model 225, Seca Ltd, Birmingham, UK), and weighing scales (model 770, Seca Ltd) respectively. Subjects then completed the Edinburgh Handedness Inventory [28] which comprises a series of questions relating to hand preferences (left or right) for a range of everyday activities. Scoring of the inventory gives a Laterality Quotient (LQ) ranging from −100 for total left-hand dominance to +100 for total right-hand dominance. The LQ obtained for each individual subject was used to designate his or her dominant and nondominant arms. Subjects were then shown a short video clip illustrating how the tennis real-world motor task (RWMT) skill assessment would be conducted. Finally, the subjects performed a basic target shooting game on the game console used in the study (Wii Play, Nintendo of Europe GmbH, Grossostheim, Germany) using the handheld motion sensitive controller in a point-and-shoot manner.

2.4. Baseline Assessment Session. The RWMT skill assessment session took place in indoors on a squash court in order to ensure consistent conditions. The RWMT involved using a tennis racket (Slazenger Smash 2”, Dunlop Slazenger Int. Ltd, Shirebrook, UK) to strike a tennis ball served by a machine (Tennis Twist, Sports Tutor Inc., Burbank, US) over the net and to try and hit a 1.4 m² target on the back wall of the squash court. The tennis RWMT assessment set-up is illustrated in Figure 1. The tennis balls were served to the subjects in a consistent arc trajectory by the serving machine with an interval of 5 seconds between balls. The subjects then had to strike the ball back over the net and aim so as to hit the square target marked out on the back wall. As laid out in Figure 1, with the subject having right-hand dominance, dominant forearm, and nondominant backhand strokes were performed in this configuration. For the other two strokes (nondominant forearm and dominant backhand), the serving machine and the subject were placed at the opposite mirror positions on the court.

Self-efficacy is the term used to refer to task-specific confidence (i.e., the conviction that the behaviour required to achieve a particular outcome can be performed successfully) [25]. Prior to commencement of the RWMT skill assessment, the self-efficacy ratings of the subjects were assessed using a questionnaire designed using published guidelines [29]. Subjects were asked to indicate the number of shots on target they thought they would be able to achieve for each of the four separate tennis strokes.

Subjects had a practice run through the four different strokes (forehand and backhand for both dominant and nondominant arms) comprising 10 trials of each stroke with a short break between strokes allow for adjusting of the position of the tennis serving machine. Performance of RWMT was then assessed with a further 10 trials in turn for each of four strokes with an observer recording the position
of each ball that struck the back wall of the squash court and whether or not it was on target.

2.5. Game-Based Virtualized Reality Training Sessions. During three days following the baseline RWMT assessment, subjects undertook three separate GBVR training sessions each of 30 minutes duration. The sessions involved working progressively through the training drills that formed part of the tennis game (Wii Sports, Nintendo of Europe GmbH, Grossostheim, Germany) and which involved developing the ability to play shots in different directions towards fixed and moving targets and also periods of match play using equally both dominant and nondominant arms. Screen shots of the drills and match play are shown in Figure 2. In order to mimic as closely as possible the RWMT, a commercially available imitation tennis racket (Play On, Toys R Us, Gateshead, UK) was attached to the motion sensitive handheld controller used to control the game and mass was added by means of a solid plastic cylinder machine to fit inside the shaft to make the weights of the two rackets equivalent (Figure 3).

2.6. Post-Training Assessment Session. The post-training RWMT assessment session occurred the day after the last GBVR training session and was an exact repeat of the baseline assessment session.

2.7. Statistical Analysis. All data are reported as mean ± standard error of the mean (SEM) with the exception of the demographic data (±SD). The impact of the GBVR training on self-efficacy ratings and RWMT skill performance (number of shots on target) was assessed using Wilcoxon Signed Rank tests (SPSS 14.0 for Windows, SPSS Inc., Surrey, UK) due to the nonparametric nature of the data with Holm’s sequential Bonferroni adjustment for multiple comparisons. Congruence between actual subjects’ self rating of their ability and their actual number of shots on target was assessed using the method described by Cervone [30] and the confidence interval of the estimate determined by using Bootstrapping.

3. Results

Sixteen of the subjects reported right-hand dominance and the other remaining subject left-hand dominance. The magnitude of the LQs ranged from 50 to 100 and indicated that all subjects had a clear hand preference and that none could be classed as ambidextrous. Prior to the baseline RWMT skill assessment, the subjects’ confidence in their ability to hit the target (self-efficacy ratings) was significantly lower for both nondominant arm forehand and backhand strokes versus the dominant arm (control) strokes (Table 1). As expected, there was a deficit in motor skill performance of both nondominant arm strokes (Table 1). After the GBVR training there were significant increases in the self-efficacy ratings for both nondominant arm strokes (Table 1). With regard to motor performance, there was a significant improvement in the nondominant arm backhand stroke performance (Table 1) while the improvement in nondominant forehand stroke just failed to reach statistical significance ($P = 0.06$). In contrast there were no changes in the dominant arm strokes. Congruence between self-efficacy ratings estimates and actual performance was consistent across all conditions and the grand mean across all subjects and strokes was 81% and out-with the 95% confidence interval for random chance alone.

4. Discussion

The key findings of this study were that a performance deficit in a real-world motor task (RWMT) improved in response to a game-based virtualized reality (GBVR) training programme that closely mimicked the RWMT. Furthermore, the improvement was manifest as concomitant increases in both motor skill and self-efficacy (i.e., increased confidence in ability to perform the task).

As the subjects completed the self-efficacy ratings prior to the RWMT performance assessments, the post-training ratings given by the subjects (Table 1) predicted a differential improvement in their nondominant strokes compared with no change in the performance of their dominant hand strokes. The most likely explanation for the subjects’ change in their perception of their ability to perform the nondominant strokes were due to an improvement as a result of the GBVR training. While care was taken by the experimenters not to give any explicit feedback during training, the subjects would have seen the consequences of their actions in displayed on the screen and through the game scoring mechanisms. Given that the baseline RWMT assessment comprised a total of only 20 shots of each stroke (10 familiarisation and 10 assessment), it seems unlikely that the baseline assessment session could have provided an adequate training stimulus to explain the observed effects. However, were the RWMT to be used for training, it is likely that performance improvements similar to those in the current study would be observed.

The experimental model designed for this study was based on the fact that in non-ambidextrous subjects, the nondominant arm has less well developed motor skill than that of the dominant arm, which by definition has much better developed motor skill due its preferential use in daily
approaches to rehabilitation practice [8–10] and in particular to training. Therefore much greater scope for improvement in response to a treatment program by providing immediate feedback on performance, in-game achievements and rewards, and a sense of accomplishment as they progress through the game. Also, there is evidence to indicate that both the challenge and the immersive potential offered by games controlled by body movement can result in a greater flow experience compared to traditional forms of exercise [31].

In order for GBVR approaches to be incorporated into routine clinical rehabilitation practice, it will require the development of game systems that are fit for purpose [19]. They will need to have sufficient fidelity in their detection of motion that the can keep the patient within the desired movement envelope. The physical exercises that the game play requires will need to correspond the prescribed therapeutic exercises and scenarios and will also need to be capable of providing sufficient stimulus to promote the desired restoration of function. There is also a requirement for the therapist to be able to tailor the game to match the precise rehabilitation requirements of each individual patient and to be able to regularly monitor a patient’s progress and have the ability to adjust the specific game demands over time [8]. The game will also need to appeal to patient and be capable of sustaining their interest. However, not only is there a scope for a wide range of genres of games to appeal to different users, but more fundamentally, if appropriately programmed, a game could adapt and respond to any improvements in a given user’s performance. This would help ensure that the game presents a challenge at an appropriate level for each individual and as the rehabilitation treatment program progressed, the game scenarios would become increasingly demanding.

Other features which could be incorporated into games for rehabilitation include offering helpful tips and coaching, rewarding, and praising the user when they make progress, monitoring game use time, and providing some form of progress chart. It is not necessary and indeed may not be life and therefore has much less potential for improvement in motor performance. In contrast, the relative under-use of subjects’ nondominant arms (again by definition) means that they have a lower baseline level of motor skill and therefore much greater scope for improvement in response to training. There is a growing interest in the use of GBVR approaches to rehabilitation practice [8–10] and in particular its potential for reducing the need for patient travel and reaching out to rural communities [2, 19]. However, the focus of attention has predominantly been on developing new forms of treatment modalities with a view to improving physical outcomes. The results of this study indicate that there is also the potential for significant positive psychological outcomes. The game used in this study provided the subjects with a number of different forms of feedback on their performance. When the user makes “contact” with the virtual tennis ball, a sound is played through a small speaker in the handheld controller. In addition, the trajectory of the ball in the on screen game play reflects the angle, speed, and timing at which the virtual ball is “struck”. The game play environment also provides additional visual (in terms of game score) and aural (cheers of spectators) feedback on performance. The impact of the GBVR training was such that the relative deficit in the RWMT performance in the subjects’ nondominant arms was rehabilitated closer to that of the corresponding dominant arm stroke. Furthermore, it would appear that the subjects were able to sense this motor skill improvement as a result of the GBVR training and that it was reflected in an increase in the self-efficacy ratings for the nondominant arm strokes prior to the post-training RWMT skill assessment. Thus, while the GBVR training did not fully replicate the sensation of striking a physical ball, nor did it directly replicate the geometry of the tennis RWMT, it did provide a training stimulus that was adequate enough to elicit an effect.

As described in the Introduction, rehabilitation treatment programs are usually long and arduous and therefore supporting and sustaining patient motivation is a major challenge. Game-based approaches are of particular interest in this regard due to the motivational appeal of video games [26]. Developing GBVR approaches to rehabilitation therefore require a degree of trade-off to be made, whereby the training stimulus may not quite match the real world equivalent, but that this is more than compensated for by appropriately designed selections of games that promote and maintain patient motivation and long-term adherence to a treatment program by providing immediate feedback on performance, in-game achievements and rewards, and a sense of accomplishment as they progress through the game. The impact of the GBVR training was such that the relative deficit in the RWMT performance in the subjects’ nondominant arms was rehabilitated closer to that of the corresponding dominant arm stroke. Furthermore, it would appear that the subjects were able to sense this motor skill improvement as a result of the GBVR training and that it was reflected in an increase in the self-efficacy ratings for the nondominant arm strokes prior to the post-training RWMT skill assessment. Thus, while the GBVR training did not fully replicate the sensation of striking a physical ball, nor did it directly replicate the geometry of the tennis RWMT, it did provide a training stimulus that was adequate enough to elicit an effect.
desirable that these functions are performed by a virtual representation of a human, but rather they can be subtlety incorporated into the game play experience [32].

4.1. Clinical Rehabilitation Impact. GBVR approaches to rehabilitation have the potential to facilitate simultaneous improvements in motor skill and confidence and may also help increase patient compliance. The availability of low-cost motion sensors and increasingly sophisticated games development tools means significant progress is to be expected in this field over the next few years. The nondominant hand motor deficit model outlined in this study is intended as far as possible to mimic the impact of trauma or disease process on arm motor function, whereby the reduced motor performance in the nondominant arm mimics the functional loss due to trauma or disease process and which therefore is in need of rehabilitation as part of the treatment programme. However, the model does not incorporate the complex nature of different clinical scenarios nor does it reflect the fact that even with specific conditions there is significant variability between patients. While this might potentially be seen as a limitation, there are in fact significant advantages to having a stable and reliable experimental setting in order to undertake basic research into the development and refinement of GBVR approaches including hardware, software, game design, clinical interfaces, and data logging [19]. Furthermore, the convenience of being able to use healthy human subjects and having each subject act as their own control will further reduce potential sources of variability and therefore make experimental testing even more efficient.

With regard to clinical interfaces, it is an essential requirement that therapists are able to precisely control the parameters of the rehabilitation exercises (e.g., specific movements, range of motion, number of repetitions, and frequency of exercise) and also get detailed information via data logging and summary reports on patient performance in order to track rehabilitation progress and adjust their programme when required. It should be noted that existing commercial off-the-shelf interactive games (including the game used in this study) have little or no functionality in this regard. Once solutions that address these issues are developed, it will then be a much more realistic proposition to transition to clinical settings in order to tailor approaches to specific clinical needs and to undertake comprehensive clinical evaluations. Finally, there are also a number of infrastructure issues that will need to be addressed before routine clinical adoption of GBVR rehabilitation is possible including accessibility issues [33], licensing and reimbursement [10], and issues of resource allocation and treatment policies [11].

5. Conclusion

GBVR training has significant potential in the development of rehabilitation practice. However, significant hardware and software design issues still need to be addressed and the nondominant hand motor deficit model described in this study provides useful paradigm for conducting basic research and development. The results of this study indicate that improvements in confidence, which is an important determinant of treatment outcome in terms of daily living and physical and social functioning, should be added to the list of potential benefits of GBVR training in rehabilitation practice.

Conflict of Interests

The author declares that there are no conflict of interests associated with this work.

Acknowledgments

Liam Baird and Graham McCraw who undertook data collection and Annemarie Crozier who provided technical support. This work has been funded in part by the EU under the FP7, in the Games and Learning Alliance (GALA—http://www.galanoe.eu) Network of Excellence, Grant Agreement no. 258169.

References


Research Article

Static and Dynamic User Portraits

Ko-Hsun Huang,1, 2 Yi-Shin Deng,3 and Ming-Chuen Chuang1

1 Institute of Applied Arts, National Chiao-Tung University, 1001 University Road, Hsinchu 300, Taiwan
2 Madeira Interactive Technologies Institute, Caminho da Penteada, 9020-105 Funchal, Portugal
3 Institute of Creative Industrial Design, National Cheng Kung University, 1 University Road, Tainan City 701, Taiwan

Correspondence should be addressed to Ko-Hsun Huang, kohsun.huang@gmail.com

Received 18 May 2012; Revised 12 September 2012; Accepted 13 October 2012

Academic Editor: Bill Kapralos

Copyright © 2012 Ko-Hsun Huang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

User modeling and profiling has been used to evaluate systems and predict user behaviors for a considerable time. Models and profiles are generally constructed based on studies of users’ behavior patterns, cognitive characteristics, or demographic data and provide an efficient way to present users’ preferences and interests. However, such modeling focuses on users’ interactions with a system and cannot support complicated social interaction, which is the emerging focus of serious games, educational hypermedia systems, experience, and service design. On the other hand, personas are used to portray and represent different groups and types of users and help designers propose suitable solutions in iterative design processes. However, clear guidelines and research approaches for developing useful personas for large-scale and complex social networks have not been well established. In this research, we reflect on three different design studies related to social interaction, experience, and cross-platform service design to discuss multiple ways of identifying both direct users and invisible users in design research. In addition, research methods and attributes to portray users are discussed.

1. Introduction

Understanding target users is considered a basic step towards developing good products and services. In traditional industrial design, marketing, city planning, and environmental design, a wide range of research methods, including surveys, field studies, interviews, and focus groups, have been used with the clear purpose of identifying target audiences’ preferences and needs [1–4]. In system and software development, usability evaluation and user-centered design methods, such participatory design, contextual inquiry, or ethnographic techniques, have also been well accepted and applied to better understand end users’ knowledge background, behaviors, cognitive processes, and requirements [5–7].

To support rapid IT development and iterative design processes, it has become important to have clear images and models to represent end users: ideally models which can be reused and reapplied in the development of different products and services [8]. In software engineering, user modeling has focused on having an internal representation of users, which includes information such as background knowledge, preferences, and the ways that users interact with systems [9–11]. This type of user modeling or profiling can be used to design serious games, educational training, and learning systems or to evaluate systems by simulating different types of users [11–13]. Beyond traditional user cognitive models, much work has extended the modeling variables to cover users’ previous computing experiences, personality traits, and background context for educational applications [11, 14]. Hothi and Hall [15] have addressed SaD (static and dynamic) user modeling, which focuses on both static information such as gender and age, and dynamic information such as diverse computing experiences and personality. They considered that such dynamic user data could be used to make a system dynamically adapt to an individual user. Karampiperis and Sampson [14] have highlighted two distinct submodels in adaptive educational hypermedia development: a frequently updated model to represent learners’ knowledge space and a static model to represent learners’ cognitive characteristics and learning preferences [11, 14, 16].
With a purpose similar to that of user modeling, personas have also been used to portray user types in serious games, adaptive training systems, communication, and experience design [17, 18]. The original concept of personas is to create fictional characters to represent different user and consumer types by describing their tastes, perceptions, possible reactions, and attitudes towards a certain product, service, or brand [18–20]. There are several significant benefits to having personas in the design process, such as providing all members of the development team a clear and common image of the target audience, evaluating if design solutions meet users’ needs, and enhancing practitioners empathy on a certain type of user [18, 21]. For instance, Anile [22] has argued that most successful products for children are neither goal- nor task-oriented but meet specific needs of children in a particular age range. She has highlighted that child-personas of more experiential contexts can provide a way to incorporate concepts from developmental psychology into design and allow archetypical users to be presented throughout the design process for technology-enhanced educational systems [22–24]. To design roles and scenarios for serious games and game-based adaptive training systems, Raybourn [17] has also suggested using personas to guide the design process and to help game designers and game writers develop realistic or believable roles for players.

Although both user models and personas can offer various benefits in different phases of iterative design, including initial concept development, user testing, and redesign cycles, there are several problems associated with their application in complex designs of social interactions, experiences, and services. First, user-modeling methodology has been developed with a focus on human-computer interaction and centers on task-oriented analysis of behavior patterns, cognitive processes, and demographic data. The approaches and modeling can reveal detailed information about the ways that a user interacts with a system, but are incapable of taking large-scale social contexts into account [25–27]. Such modeling cannot account for subtle social interactions and communication between people, which is the emerging focus of serious games, adaptive hypermedia, and learning networks [28, 29]. Secondly, the information that a traditional persona offers is also limited to a few attributes. Although a person’s sociocultural background and economic status may be considered, detailed information such as one’s relationships with communities, products, and technologies, are generally absent. In addition, several researchers have criticized persona development for lacking scientific process and clear guidelines, which decrease the method’s reliability and representability [30–33]. Both user modeling and personas restrict their scope by overlooking activity contexts; user modeling only focuses on direct users of systems, and personas mainly represent target and potential consumers, rather than capturing and revealing all participants involved in the activities of interest.

Much work has shown that having social perspectives and understanding nuanced social interaction will be the most challenging but necessary topic in today’s experience and social media design [4, 7, 17]. In addition, much work has shown that enhancing social identity and applying social influence can help practitioners and designers develop systems for learning, behavior changing, and other social purposes [28, 34, 35]. Furthermore, design strategies such as applying CASA (Computer as Social Actor) in technology-enhanced education also require a thorough understanding of learners’ sociocultural backgrounds [36, 37].

This paper discusses current methods, approaches, and frameworks applied in design research, including several user-centered approaches, design frameworks, and models developed for studying cultures. We highlight the strengths and shortcomings of these methods when applied to capture different levels of user information, from detailed interactions to sociocultural backgrounds [3–5]. Examples from user studies in three design projects are used to illustrate that the different levels of information help identify user types and build user portraits. In each case, multiple ethnographical techniques were applied, such as shadowing, in-depth interviews, grounded theory [38, 39], work modeling [40], and social activity modeling [4], to reveal the complicated involvement of different types of participants in social activities and product usage.

We reflect on the three case studies to recommend suitable user study methods for supporting large-scale services, adaptive systems, and development of serious games, as well as necessary information to gather when portraying users. By understanding participants’ relationships and roles within a broader context, we also derive insights for identifying invisible users and potential users and the importance of understanding different users’ motives and concerns.

2. Methods and Frameworks

Emerging areas of interest in IT development, such as serious games, adaptive systems, and experience design, are concerned with large complex communities and diverse generations. Large-scale services such as e-government and social media inevitably have to support people’s social activities in the coming future. According to the central concept of ethnography, people are intelligent and creative, and with ad hoc practices, they can apply their knowledge across domains and act in different contexts [41, 42]. It is the methods used by people to apply knowledge and their values and attitudes, which matter and can help scholars to predict further actions, reactions, and acceptance towards a certain system or service. Instead of looking into what people actually perform, understanding background motives is essential for developing more adaptive and flexible systems and services for greater social purposes in the future.

2.1. Models for Social and Cultural Studies. In social science, the strategies to study a culture include traditional anthropological approaches, such as participant observations with long-term involvement of the fields and interpretative approaches, in which researchers collect, conceptualize, and induce the concepts through diverse methods. For years, numerous discussions and debates have been held about ways to generate more solid and scientific results while
applying these methods and approaches. Many researchers have focused on measurements and standards for improving the validity and reliability of qualitative research [43–45]. For instance, Glaser and Strauss [38] developed a grounded theory in which researchers analyze data and generate a theory repeatedly until they can thoroughly explain and describe the phenomenon. In this iterative process, Strauss placed emphasis on improving reliability and validation criteria in a systematic way [38, 39, 46]. However, in IT and product design practices, while designers and developers apply qualitative approaches, they tend to analyze the situations by using models as guidelines and check lists, and the research results generally only support short-term and inner group usages and have no impact on further related studies.

Applying the above-mentioned positivist or interpretivist approaches to gain knowledge about society and culture is very time consuming and requires experienced researchers to collect and analyze raw data. Therefore, to reduce the cost and to gain insights more efficiently, there are numerous theoretical frameworks and models developed in different domains. For instance, since the 1980s, researchers in management and leadership have discussed numerous models for understanding organizational cultures [47–51]. Most of these works have the same purpose, either to optimize the organizational process or to reduce internal conflicts within organizations.

In the fields of management and leadership, many scholars have taken similar approaches to identify cultures—firstly recognizing a culture's representative characteristics and then categorizing them into types. These researchers generally take an organizational point of view and focus on structures, power distribution, and divisions of labor [48, 52, 53]. The earliest work is Harrison's organization ideologies, in which there are four types of cultures highlighted by their typical features [52, 54]. This classification refers to power-, role-, task-, and person-oriented cultures, respectively, standing for centralized power, hierarchical structure, team support, and individual achievement [52, 55, 56]. Harrison's work has had a great impact on later studies of organizational culture, in which the power distribution and organizational formalization have become the basic criteria to locate a culture [52]. Furthermore, the ways that an organization or a community responds to outside influence and different situations, the flexibility and stability of organizational structures, and the forms of attention have also been widely discussed [48, 53, 57].

In contrast to categorizing cultures into types, many scholars focus on finding cultural concepts and patterns for intercultural studies. The earliest works are Hall's books, in which he identified two dimensions of culture, including the high- and low-context communication and polychronic versus monochronic time orientation [58, 59]. The high-low context concept is used to characterize information transaction and interpersonal communication in cultures, and the second concept is concerned with the way in which people structure their time, tasks, and schedules. Moreover, to deal with cross-national issues in the functioning of organizations, Hofstede [60] indicated the differences between studying national cultures and organizational cultures where national cultures differ primarily in their values, but organizational cultures turn out to differ mainly in their practices. He identified six different dimensions of organizational cultures and claimed five other national cultural dimensions, including power distance, uncertainty avoidance, individualism versus collectivism, masculinity versus femininity, and long-versus short-term orientation [60, 61]. For understanding cultural diversity in global business, Trompenaars and Hampden-Turner [62] also developed a model of culture with seven dimensions, including universalism versus particularism, individualism versus collectivism, neutral versus emotional, and specific versus diffuse.

Instead of classifying existent international or organizational cultures into types, some researchers have focused on developing models for analyzing and understanding the culture of a particular group or organization. For instance, Schein [63] identified three fundamental levels at which culture manifests itself: observable artifacts, values, and basic underlying assumptions. Here, artifacts refer to physical layouts, manners, atmosphere, and phenomena that people can directly feel and observe, and values refer to members' norms, ideologies, and philosophies. The basic assumptions are those taken-for-granted, underlying and usually unconscious aspects that people have and share within groups, and they normally can determine people's perceptions, thought processes, and behaviors [63, 64]. Closely resembling Schein's three levels, Hofstede placed four manifestations of cultures at different levels of depth, namely symbols, heroes, rituals, and values [57, 60]. Rousseau [49] also detailed the nature of the cultural construct and its theoretical roots and layered these cultural elements according to subjectivity and accessibility. His five layers of culture, from outside in, are artifacts, patterns of behavior, behavioral norms, values, and unconscious fundamental assumptions. This type of layered cultural model also appears in many other researchers' works, in which basic assumptions and values are generally placed at the core layer(s), and then are encircled by beliefs, attitudes, rituals, behaviors, and then artifacts [65, 66].

There is one extreme example of the ambitious objective to combine all aforementioned concepts into one single model. Considering cultural impact on the implementation of enterprise resource planning (ERP), Krumbholz and Maiden [67] developed a metaschema for modeling culture. They integrated the surface and the deeper manifestations of culture into common business concepts such as processes, events, and information flows, and then developed a systematic framework to analyze culture for further explanation, prediction, and replanning of different corporate and national cultures. Their framework is based on Schein's three levels of culture and a wide range of social aspects (e.g., group norms, formal philosophy, and linguistic paradigms). Influenced by globalization trends and comparative studies, they also included Hofstede's and Trompenaars and Hampden-Turner's cultural dimensions in their work [16, 50, 62]. This enormous framework has more than twenty components in three main categories.
The first category covers common elements in business processes, such as agent, role, responsibility, and goal. The second category describes the core levels of culture, including hidden assumptions, beliefs, and values. The last category focuses on characteristics such as customs, symbols, and environments. Krumbholz and Maiden tried to cover all the important elements discussed in cultural studies. However, there are many unclear definitions and assumptions in this metaschema. For instance, social interaction is a significant component of the model, which influences and reflects one’s beliefs and represents the types and ways people interact with each other. However, their work does not detail how to explain relationships between social actions, values, and beliefs nor does it suggest methodology to gain such an understanding.

The models of cultural types provide several ways to identify national or organizational cultures and have indicated the significant relationships between power structure and personal behaviors [48, 52–54]. However, there are two major reasons why it is difficult to apply these models directly as research frameworks in other domains. Firstly, the limited cultural types oversimplify the sociocultural issues and avoid the level of depth in cultural significance. In management and leadership cases, the classification approach provides a simple and quick way to identify the cultural characteristics, but due to a lack of clear essential definitions, these models are incapable of revealing the subtle relationships and influences between the cultural components. Second, most human behaviors, especially social activities, do not have a clear objective and can be analyzed into steps and processes. With a strong management and leadership purpose, these models purely focus on work processes and the ways that people achieve their goals and seldom take individuals’ motives, attitudes, and perspectives into account. The similar tendency of work and goal orientation appealed in Hofstede’s global-scale survey [60, 61], which he conducted to evaluate the work values of a specific company. Although Trompenaars and Hampden-Turner [62] discussed personal attitudes towards both leisure and work situations, their results related to underlying assumptions and values were very limited. Krumbholz and Maiden [67] have tried to combine both social science theories and psychological concepts into their metaschema, but their interview results did not reflect the richness and interaction among these cultural levels and components. In addition, the numerous components involved in different perspectives are all compressed into a single model, which makes this metaschema difficult to use as a research framework or as a format for representing the final output of the research.

In social interaction and experience design, issues which are related to individual attitudes and values can often be traced back to and better understood in the light of social norms or culture. However, studies of how society and culture act as constraints for design have very different goals than traditional broad studies of culture itself. They also differ from the above-mentioned models and measurements of culture which focus purely on work and whose purposes are either to enhance organizational performance or to improve management and leadership. Due to their strong task-orientation, the methodologies proposed for these studies focus on practical and behavioral views, environments, and symbols. They overlook several cultural aspects which are important in socially motivated interactions between people, technology, and services, such as people’s motives for actions and their emotional needs.

2.2. Frameworks for Understanding Users and Usage Contexts. In the past decades, design research of information and system development has focused on interactions between humans and machines. Based on the introduction of cognitive psychology, most studies have concentrated on mental processes and information flows, with the scope being achieving a task or solving a problem. Taking an example of Donald Norman’s seven stages of action, which have had a lasting impact on usability engineering and industrial design, the analysis of human action is concerned with a loop of forming the goal, forming the intention, specifying an action, executing the action, perceiving the state of the world, interpreting the state of the world, and evaluating the outcome [68]. Comparing these types of mental models with microsociological theories (e.g., symbolic interactionism and dramaturgy), they all focus on the interaction between individuals and environments and discuss how people perceive and interpret the outside signals, symbols and then take an action [69].

The scope of design research in technology development expanded during the 1990s from task analysis to more complex activity analysis. For understanding larger-scale usage contexts and supporting user experience design, many ethnographical research methods were introduced in design practices, such as interviews and long-term involvement observations [70–72]. In addition, to speed up design cycles, many analytic frameworks and models were introduced and developed to make user research more efficient, cheaper, and deeper. These frameworks, such as AEIOU (activity, environment, interaction, object, and user), POEMS (people, objects, environments, message, and services), Ax4 (atmosphere, actors, activities, and artifacts), activity theory, and contextual design methodology, have been discussed and applied in various domains with great success [40, 70, 73–75]. Most of the frameworks provide clear guidelines and dimensions for investigating the entire activity context, including practical behaviors and actions, related objects and environment settings, and information content, as well as taking into account relationships among people.

Among these research frameworks, contextual design methodology [40] and activity theory have their specific advantages for both system development and other design practices. Contextual design methodology was developed based on research techniques of ethnography and was influenced by the development of participatory design techniques in the 1980s and 1990s. It has a special purpose to help researchers and designers identify domain problems in rapid design cycles, especially for software and hardware redesign and usability evaluation. To help researchers and designers convey their domain knowledge, thoughts, and ideas, Beyer and Holtzblatt [40] developed five work models as a tangible
representation for issues in different dimensions, including the flow model, cultural model, sequence model, artifact model, and physical model. Furthermore, for developing a successful system, which can “fit with the customer’s culture, make conforming to policy easy and reduce friction and irritation in the workplace,” Beyer and Holtzblatt [40] addressed the importance of understanding organizational culture. Their cultural model highlights cultural influences among individuals, groups, and the organization and also helps researchers identify invisible power, preferences, values, and emotions. Although the cultural model concentrates on representing organizational culture, it still shows individuals’ opinions and attitudes to some extent.

Activity theory, with its roots in the 1930s Soviet cultural-historical psychology, was introduced and adapted into HCI and CSCW as a lens in ethnographic research [73, 76, 77]. To understand the mental capability of a single individual, activity theory considers a “goal-directed” activity as the unit of analysis and provides an analytical framework to describe activities with three hierarchical constructs: subject, object, and tool. In the theory, activities are described by how a single individual (subject) achieves a goal (object) through tools, and using tools reveals the details of both physical interactions and mental processes. Under this framework, a complex activity can be broken down into action or operation levels and can be analyzed from both behavioral and psychological viewpoints. To deal with multi-user systems and collaborative work, Engeström [73, 78] later proposed an extended schema for activity theory with additional constructs of community (people who share the same goal), rules, and division of labor. Engeström's schema makes activity theory very useful in groupware and social media design [79, 80].

With a strong intention to bridge the gap between subjective-to-objective and macro-to-micro concerns in design study, MultiLevel Social Activity Model (MLSAM) was proposed by Huang and Deng [4, 81]. There are two basic arguments behind the concepts. First, they argued that social behaviors are deeply localized and historical on the account of cultural background. They showed that the traditional customs perform a social function by creating cohesiveness in families and by offering a habitual practice that can be passed on from one generation to another. The social activity not only reflects norms and common values of the society, but also presents various inherent characteristics of the cultural context. Secondly, they argued that people have great agency and creativity to fulfill needs and achieve purposes in a variety of ways. Traditional field observations and user studies, which focus on a limited scope and settings, can only reveal partial results of how people perform to achieve their goals. Hence, their model emphasizes the necessity to identify people’s motives and attitudes, which comparing to actual actions, are more permanent and static, and could be considered in further design for different platforms and services.

From an integrated sociological viewpoint, individuals’ behaviors are not only encouraged, but also constrained by norms, religions, and sociocultural backgrounds. To well support hypermedia and adaptive systems for social purposes in the future, design research needs to extend from traditional usability evaluation and task-oriented studies to a larger-scale sociocultural scope. Therefore, an in-depth research approach is needed to answer to the complexity of social interactions. The following section presents three case studies of applying both contextual design [40] and multilevel social activity model [4, 81] in design practices and shows the benefits of applying multiple research methods to portray users.

3. Design Studies

To better support product and service design in the future, we applied multiple user study methods to uncover the complexity of participants’ perspectives, interactions, and attitudes within different types of activities: a traditional ritual of a Taiwanese tea ceremony, Taiwanese teenagers’ social activities, and technology use, and sports watching in Southern Europe. The first study is selected to show how well-accepted software design methods can be used together with the multilevel social activity model [4, 81], to capture more user information within a social event, a cultural tradition. The second case study is selected to discuss if this integrated approach can also be applied to investigate a complex physical and virtual social context. The third case study is a service design, in which the same approach is applied to identify different user types and help designers and practitioners make design decisions on developing adaptive and cross-platform systems. Contrary to what is common in traditional IT development, we identify different user types by using both the contextual design flow model [40] and the multilevel social activity model [4], to highlight participants’ roles, relationships, motives, and sociocultural background.

3.1. A Traditional Social Ritual. The first case is a study of traditional tea ceremonies in Taiwan. These ceremonies have deep cultural roots and also contain complex forms of social interaction, which are typically ignored in most IT development and for which established research methodology is lacking [4]. We reflect on this case study because of its rich sociocultural backgrounds, which represent a good example to show complex social relationships and different user motives within an activity.

In this case study, we applied multiple user experience research methods, which include observations [3], contextual design work modeling [40], social activity modeling [4, 81], and in-depth interviews with grounded theory analysis [38, 39], to reveal different aspects of personal perceptions toward the overall sociocultural context of Taiwan's tea ceremony. In order to identify the value of the tea ceremony to participants from different generations, three subjects of different ages were recruited, together with their regular tea ceremony groups. The three subjects were all experts of tea ceremonies. The first subject was a forty-year-old parent with a twenty-year experience of tea ceremonies. The second subject was a retired senior citizen, with a seventeen-year experience and with an interest in learning techniques and
knowledge of brewing tea from magazines and other connoisseurs. The third subject was a female graduate student, with a fourteen-year experience of tea ceremonies and with familiarity of the Internet and associated technologies. Each had taken part in tea ceremonies at least once per week in recent months. To observe the detailed social interaction within the activity, their family members and friends, who are regular members of the ceremonies, were also invited to participate in the sessions of observations and contextual inquiries. In total, we had ten participants.

The flow model in Figure 1(a) represents the different roles in the ceremony and their interactions. On the basis of the in-depth interviews and contextual inquiries, this flow model has been extended to portray all groups who directly or indirectly engage in a ceremony. Direct participants are the host (tea server), family members, and other friends, while indirect participants include hobbyists, tea sellers, tea producers (farmers in direct marketing), and tea connoisseurs. Although indirect participants do not actually attend ceremonies, they significantly impact the act through their close ties to the host. According to the study, the host of a ceremony is most likely a person of the middle-aged generation or the head of household. This person tends to actively exchange information and sentiments with other hobbyists and tea sellers, as well as seek out information from books, magazines, and newspapers.

While the flow model captures roles and interactions, the multilevel social activity model further reveals how participants’ underlying motives and attitudes toward tea ceremony differ between generations. It also shows that attitudes and resulting behaviors are all strongly rooted in a cultural context. Both the younger generation and the elderly/middle-age generations (both groups in Figure 1(b)) are direct participants in tea ceremonies (marked in orange in Figure 1(a)). However, a comparison of the two models reveal that although the younger generation is interested in tea ceremonies and has inherited tea-drinking habits from their families, they are generally not consumers of the tea industry. They have neither contacted any other participants (i.e., tea sellers and tea farmers in direct marketing), nor do they receive information through magazines or newspapers. Instead, their primary sources of information about tea ceremonies are parents or grandparents.

The multilevel social activity model identifies different user groups by their attitudes towards activities and motives for participation. In addition to people’s direct and immediate requirements, we consider that there are many hidden reasons for people to engage in a social activity that may be overlooked by methodologies that do not go deep enough. For instance, the social activity model reveals that the elderly have positive attitudes toward tea ceremonies potentially reflecting the awareness of Westernization, while the younger generation is more attracted by the health benefits associated with drinking tea.

An understanding of differences in user groups’ cultural backgrounds and underlying motives can help designers make better decisions, in particular in design for social activities and communication. For instance, knowledge of cultural features, for example, pouring a cup of tea to convey esteem and respect during a tea ceremony, can further lead to more accurate portraits of users and better predictions of user behaviors. However, it is difficult to gain this knowledge using only activity- and usage-centered design methods. Instead, the case studies indicate that the knowledge can be gained from a historical and broader-context approach.

3.2. Virtual and Physical Social Networks. Crazy Vote was a social website in Taiwan that provided users with personal web space, such as weblogs and a message board. Due to its unique interface and features for voting on users’ portraits, it became the biggest social website for Taiwanese teenagers in 2008, with more than 20,000 users of ages 15 to 19. To guide future application development, the company supported a two-month research project to fully understand their users’ online activities and expectations of social media [82]. In the study, seven highly active users, who have their own fan clubs and hold social events, and two regular users, participated in both in-depth interviews and contextual inquiries, and the online logs of another 40 highly active users and 40 regular users were sampled at random to understand behavioral patterns in the platform. All qualitative data were analyzed by following grounded theory with Nvivo [83, 84]. At the same time, contextual design work models and the multilevel social activity model (MLSAM) were used as design research guidelines to capture information and present results [82].

By extending the research scope with both contextual inquiries and in-depth interviews, the study shows that teenagers’ common processes of making friends are complex, but flexible. In addition, the boundary between online social interaction and actual relationships is very blurred to Taiwanese teenaged online users. Their reason for making new friends on the Internet is simply to expand their interpersonal relationships in the real world. For instance, the interviewees mentioned that they preferred to make friends who live nearby to increase chances of meeting up in person, as society does not encourage teenagers to travel alone. Social issues that are associated with meeting online friends also make teenagers form unique networks to ensure that all members are using their real identities and to later develop real-life confidence in each other.

As shown in Figure 2, the multilevel social activity model also helped us identify three different user types among the website users. First, activity promoters, who were extremely confident and familiar with most social norms and manners on social media, voluntarily held gathering events, established clubs and recruited other users to join their own clubs. Second, followers were willing to participate in social events but had less interest to be a group leader or to organize activities. Third and finally, self-oriented users made up 90% of the user base, and their activities on the platform were more self-oriented, such as maintaining and updating blogs and photo albums. They seldom visited others’ blogs or left messages to others. According to the in-depth interviews, we found that most self-oriented users were either introvert or lacked experience of interaction with unfamiliar people on the Internet. Therefore, we further separated these three
Advances in Human-Computer Interaction

People who actually participate in tea ceremonies
- Friends/relatives
  - Have tea snacks/cakes
  - Chat with others
- Host
  - Introduce tea ceremony
  - Make tea repeatedly
  - Serve tea repeatedly
  - Introduce and taste tea
- Connoisseurs (ceremony experts)
  - Introduce tea ceremony
  - Guide how to make tea
- Tea producer (tea farmer)
  - Introduce new types of tea
  - Record customers' preferences
  - Provide and send samples
  - Sell tea
- Tea sellers (teahouse/tea shop)
  - Introduce tea ceremony
  - Introduce different types of tea
  - Provide tea ceremony service
  - Make/serve tea
- Tea snacks
- Make/serve tea

People who are involved in tea ceremonies
- Friends/relatives
  - Taste tea
  - Have tea snacks/cakes
  - Help to clean teaware after the ceremony
- Colleagues/business partners
  - Taste tea
  - Have tea snacks/cakes
  - Talk about business
- Fellow tea drinkers and other hobbyists
  - Taste tea
  - Exchange information
  - Recommend teas
- Tea producer
  - Introduce new types of tea
  - Record customers' preferences
  - Provide and send samples
  - Sell tea
- Tea sellers
  - Introduce tea ceremony
  - Introduce different types of tea
  - Provide tea ceremony service
  - Make/serve tea
- Books, magazines, and newspaper
  - Provide (medical) information of tea drinking
  - Introduce tea ceremony
- Host
  - Introduce tea ceremony
  - Make tea repeatedly
  - Serve tea repeatedly
  - Introduce and taste tea
- Connoisseurs (ceremony experts)
  - Introduce tea ceremony
  - Guide how to make tea
- Tea producer (tea farmer)
  - Introduce new types of tea
  - Record customers' preferences
  - Provide and send samples
  - Sell tea
- Tea sellers (teahouse/tea shop)
  - Introduce tea ceremony
  - Introduce different types of tea
  - Provide tea ceremony service
  - Make/serve tea
- Tea snacks
- Make/serve tea

People who are direct consumers of teas and teaware
- Tea ceremony services
- Tea ceremonies enhance the sense of belonging of family members
- Enhancing social relationships
- Improve family gathering
- Heritage family tradition
- Family tradition heritage
- Serve tea and chat
- Serve tea and chat

People who participate in tea ceremonies, but not consumers
- Convey esteem and respect during a tea ceremony with Chinese traditional manners
- Motives
  - Have a balanced lifestyle
  - Have mental relaxation
  - Enjoy family gathering
  - Improve family gathering
  - Heritage family tradition
  - Serve tea and chat
  - Artifacts and environments with traditional Chinese features provide a comfortable ceremony context
- Attitudes
  - Work-life balance
  - Awareness of increasing Westernization
  - Enhancing social relationships
  - Enhancing social relationships
  - Zen Buddhism/philosophical Taoism
  - Serve tea and chat
  - Serve tea and chat

Sociocultural context
- Popular lifestyle in Taiwan
- Significant traditional customs
- Chinese social values
- Philosophy of Zen/Taoism
- Significant traditional customs
- Chinese social values
- Philosophy of Zen/Taoism
- Significant traditional customs
- Chinese social values
- Philosophy of Zen/Taoism
- Significant traditional customs
- Chinese social values
- Philosophy of Zen/Taoism

(b) The flow model of tea ceremonies

Figure 1: Different user types in tea ceremonies.
type users, activity promoters, social-oriented followers, and self-oriented users, into two groups, in which the first group users are more social-oriented (marked in orange in Figure 2) and the second group users are more self-oriented (marked in blue in Figure 2).

Some teen users had a common and well-defined procedure for making friends successfully and efficiently in the Crazy Vote platform, and most users in the social-oriented group were aware of and applied this process. First, nonverbal introductions would take place through the voting system or by sending emoticons to others. Communication would then be initiated by leaving a private message or by visiting and leaving public comments on each other’s blogs. People who share similar interests and habits may then exchange other online contact information, such as MSN Messenger or Yahoo! Messenger accounts, and start communicating electronically outside of the Crazy Vote platform. In the end, these online friends may end up talking on mobile phones and meeting up face-to-face.

Rapid expansion of information and communication technology has made young people comfortable with using a wide range of communication platforms. Although taking place in an online environment, the observed process among Taiwanese teenagers for making friends is natural, mature and matches traditional Taiwanese social norms. For instance, the initial use of nonverbal emoticons and “likes” to make others aware of their presence was described in the interviews as a type of “reserved” introduction, similar to a head nod or eye gaze. Young users considered it too aggressive and impolite to suddenly show up and introduce themselves in front of strangers. However, according to both interviews and online tracing, users in the self-oriented group were unaware of this process and fell back on expanding their presentation of themselves in the system. Although both socially and self-oriented users initially shared the common goal of making new friends through the Crazy Vote website, the self-oriented users perceived a difference in the way they made friends, which later caused them to focus on their own blogs.

This study illustrates how more in-depth user research can lead to detailed interaction issues as well as an understanding of sociocultural contexts and their influences on users’ motives and behaviors. Such knowledge can also be applied and reused in many different design projects. In addition, understanding users’ expectations and abilities helps development teams make better decisions and predict user engagement. For instance, the social-oriented group’s capabilities and successful strategies of making friends can
be applied in social media design to help and guide the other types of users. However, current design research generally focuses on a single platform or a particular environment, which narrows down the research scope and overlooks people’s great ability to manipulate different resources to achieve their goals.

3.3. Large-Scale Services and Cross-Platform Experiences. The third case study investigates people’s sport watching experiences in Portugal. The design process began with interviews and observations of 20 active sports fans and people who had participated regularly in football watching activities, followed by modeling according to MLSAM [4] and contextual design [40] to understand their experiences, motives, and behaviors in sports watching. On the basis of the models, three personas and 30 different design concepts were then generated in a workshop, and 15 subjects were asked to do card sorting to rank the design concepts. In the design phase, several user experience and service design techniques were applied to define the details of the service system, including scenarios and storyboards, customer journey, use-case analysis, and service design blueprint [3, 17, 85, 86].

To gain overall feedback and to improve the systems, formal usability evaluation was conducted with five active football fans, by following collaborative usability inspection, rapid iterative testing, and evaluation and single-subject testing [87, 88]. The flow model in Figure 3(a) shows that football fans use laptops and other high-tech devices to enhance their football watching experience at home. Examples include accessing high quality streaming, using a projector, receiving statistical information from websites, and discussing referees and penalties with friends on the Internet. People watching the game at home do so either together with family members or with close friends. Figure 3(b) shows that a football watching activity in a sports bar involves several different types of people in addition to the football fans. This includes peddlers who sell team scarves and jerseys to sport bar customers, the staff, and owner of the bar, and the fans’ friends, who despite not being present in the bar interact with bar customers through digital devices.

The multilevel social activity model (Figure 4) distinguishes several types of people by their different motives for participating in sport watching. Primary supporters enjoy watching games, with strong interests in details of the game and high-quality game play. Potential supporters consider football watching primarily as a social activity and their motives for participation are generally derived from primary supporters’ interest (e.g., most of our male interviewees mentioned that their girlfriends, wives, and children try to understand the game rules so that they can be more involved in discussions of details together during the game). The other user types, such as bar owners and environment providers, participate in the events because of expectations of business opportunities.

Using the users types identified in the multilevel social activity model, we marked the primary supporters (football fans and community) in orange in Figures 3(a) and 3(b). Entertainment and media companies have traditionally always targeted their products purely towards primary supporters. Groups marked in blue are not primarily attending the gathering to watch the game, but their greater interest in social interaction nonetheless makes them potential customers of a service. Although they are not the active sports fans, it is possible to include these groups as stakeholders in current or future service design processes to develop better services of greater scope. For instance, bar owners can be clients of media companies, if a service design aims to provide a social entertainment space with complete sport channels and other facilities.

Due to resource and time limitations in this academic project, only primary supporters were selected for the continued development. On the basis of the research findings, two personas were developed who, while both being primary supporters, have very different expectations of the IT product used. As shown in Figure 5, two user portraits were introduced to the development team in the session of brainstorming and decision-making. All project participants and sponsors agreed that the two user portraits were very representative of typical football fans in Portugal. The further customer activity journey, service architecture, scenario, and storyboard were also developed according to our personas’ motives, usage of technology, and lifestyles.

Having a clear image of two user types also helped designers and engineers generate correct use cases [89] and prioritize important features in the later stages of prototyping and system development. For instance, the development team discussed different use cases and designed detailed interaction based on service design blueprints, which require correct selections of platforms (channels) for a certain type of users. During implementation, we noticed that our user portraits helped the development team focus on supporting the primary usage situations and helped reduce the complexity and conflicts while supporting cross-platform interactions. Moreover, the two user portraits were also used to plan the usability evaluation and to discuss our service values in user test sessions [87, 88].

4. Static and Dynamic User Portraits

In this section, we reflect on the three case studies and argue that two types of user information, the dynamic and static attributes, should be captured and included in design research. The suitable methods and techniques to capture this information will also be discussed.

4.1. Dynamic User Information. The cases indicate that there are two types of information that have not been highlighted in general user modeling, profiling, or persona description, yet are important for identifying potential users and predicting their behaviors. The first type of user information is dynamic attributes, such as knowledge space [14, 16], age, lifestyles, and IT consumption. The second type is permanent information, such as attitudes, preferences, customs, and appropriate behaviors.
People who are considered as primary users/consumers of most sports related services
People who participate in sport watching, but not fully concentrate on the game

(a) The flow model of football watching at home

(b) The flow model of football watching at bars

Figure 3: The flow models of sport watching.

The importance of the first type of information emerged from both the case studies of tea ceremonies and of teenagers' social activities. As shown in Figure 1, tea ceremonies are a typical type of traditional social activity that enhances the sense of belonging of family members, and which is passed on from one generation to another. However, from marketing and usage-centered viewpoints, the youth involved in the ceremony are neither consumers nor end users. Due to differences in media usage, they generally receive no information from tea producers or direct-marketing farmers...
In the case study of teenagers’ social activities, we found that although teenagers are willing to be active users of technology in their pursuit of new friends, their available modes of communication are often limited by their economic status. As teenagers generally have limited financial resources, most
IT products and media producers are unwilling to develop services for them. However, the older teenagers will see great improvements in their personal economy in only one or two years, and their great interest and reliance on the Internet will make them the target audience of smart phones, tablets, and other IT products in the near future. This type of information, predicting future audiences, is basic in marketing, but it is often overlooked in design research.

Different from most work in adaptive system development [15], we consider age and IT consumption to be dynamic attributes for user portraits. Traditional systems normally serve a specific type of users. For instance, most educational applications target a user group in a narrowed age range. However, for lifelong learning networks, large-scale services, and games with social purposes, the goal is to satisfy a large, diverse audience with long-run usage. Much work has highlighted several dynamic variables in development of technology enhanced learning systems and adaptive games, such as learners’ knowledge space, skills, and capabilities [16, 90, 91]. However, we argue that user portraits, profiles, and models should also reflect how people’s hypermedia experience, interests, and social behavior change with demographic transitions. Therefore, design research needs to consider predictable shifts in age, lifestyle, economic status, and IT consumption. To capture the dynamic information, surveys, self-documents, and focus groups can provide effective results.

4.2. Static User Information. More permanent and static information, including individuals’ preferences, attitudes, and values, has been highlighted in some work of developing personas and user models [11, 14, 17, 28]. However, higher-level information such as norms, interpersonal relationships, and sociocultural backgrounds, are generally ignored in IT development.

According to multilevel social activity models and flow models of all the case studies, participants in a certain activity can have very similar behaviors, but their concerns, attitudes, and motives may still differ greatly. For instance, in the case study of football watching, we identified that participants’ requirements and underlying motives varied, including seeking high-quality watching experiences (primary supporters), seeking detailed information (primary supporters), wanting to discuss the game with others, and enjoying each others’ company (potential supporters). However, customer journey maps [92], which are used to describe an activity route of a user and to plan different touch points that characterize user’s interactions within the service, cannot fully support development teams when deciding suitable platforms (e.g., touch points and channels in service design). On the basis of our data gathered through multiple methods, we consider that the customer journey map has to be extended into an activity journey map (Figure 6), to reflect users’ differing expectations. The map contains three types of customer journeys, representing different users’ types and their requirements. The awareness of the differences allows designers to develop reflective and reasonable scenarios and also helps development team prioritize use cases, decide appropriate platforms (channels), and plan user testing.

In addition, understanding of high-level information such as values, attitudes, and sociocultural background is easily taken for granted and therefore ignored in design research. Mulwa et al. [91] have listed twenty-one different user features for developing adaptive educational systems in the literature from 1996 to 2008, but none of these variables reflects values, social behaviors, and activity contexts. Most user experience design and HCI research methods can reveal detailed usage situations and users’ cognitive characteristics, but have difficulty identifying meanings and norms behind activities. For instance, teenagers’ friend-making processes on the Internet still follow the cultural manners that apply in their daily life. Additionally, in the case study of tea ceremonies some participants considered the ceremony to be part of East Asian culture and feel responsible to maintain this tradition.

4.3. Identify Invisible Users and Portray Users. On the basis of the studies of different social activities in different countries, we have discussed the strengths and weaknesses of different research methods. To better support rapid design research in most IT development, common user-centered design methods and frameworks provide a cheaper and more efficient way to account for the behaviors and interests of target users. However, in design for supporting communication, social purpose, and larger-scale services, we recommend to apply multiple research methods to gain deeper insights into the contexts and to identify potential and invisible users. Therefore, we propose the following process.

1. Focus on a certain activity: as a pilot study, researchers can choose a representative activity and apply common context-, user-, or usage-centered research methods to capture events in great detail. In this stage, people who actually participated in the activity, objects, media, environmental circumstances, and interactions are identified.

2. Extend the context: the goal of this stage is to identify the flow of information and find the social network that is related to the activity in the pilot study. Through an iterative process, researchers can identify how both central and peripheral events and people are connected with each other and associated with the main activity. To avoid missing important details, we suggest applying long-term ethnographic approaches, such as interviews, shadowing, behavior tracing, and self-documents.

3. Recognize participants: in this stage, all people’s goals, roles, interests, and participation in the target activity should be clarified. This information can help researchers to further distinguish different user types and identify potential users.

4. Selection and focus: a workshop or a focus group in this stage can help development teams decide which
types of users should become the target audience and to set clear priorities for design development. Once the user types are selected, surveys and in-depth interviews can be used to capture in-depth information about these selected users.

To make the research results serve the same purpose as user models and personas, both dynamic and static information need to be highlighted while portraying the users. First, the portrait (it can be presented as a model or a document) should contain basic demographic information, such as gender, interests, and preferences. However, different from normal user profiling, we highly recommend that researchers also include dynamic user attributes that help predict how users’ behaviors will change with shifts in their demographic data. For example, predictable increases in income are associated with transitions from a teenage life to maturity and result in changes in IT consumption and use. The second type, static information, is about higher-level concerns, including users’ attitudes, motives, beliefs, and their sociocultural background. As mentioned above, this information is permanent and can enrich design solutions and can help evaluate designs in different phases of design cycles and in different projects.

5. Discussion
Through contextual design flow modeling, multilevel social activity modeling, and traditional ethnographic techniques, we have shown that there are different types of invisible users, who are involved in social activities, but are neither direct product users nor customers. In the case of a tea ceremony, the young generation, who has inherited tea-drinking habits from their parents, is not considered as a target audience of the traditional tea industry and can only access very limited information. In the case of football watching, potential supporters represent a large group of people who participate in the activities regularly because their children, parents, friends, or colleagues are football fans, and they are willing to share these interests. However, most entertainment and media companies have only considered providing services for football fans. On the basis of what we learn from the three studies, it is argued that discovering all user types and presenting static and dynamic user portraits will bring many long-term benefits for systems and services.

(i) Identify potential users: in the three cases, we have shown that extending the research scope can help gain the necessary understanding of complex social interactions and social contexts and discover potential and indirect end users. We consider this to be significant understanding for serious games and technology enhanced education systems.

(ii) Support diverse and large user groups: development teams can predict new needs and expectations of future target audiences by identifying predictable changes in users’ IT-product usage and economic status. Furthermore, identifying user types and motives can help designers propose more reasonable design solutions, satisfy users’ different needs, and develop more adaptive systems.

(iii) Support multiple projects: compared to the outputs of usability and activity research, sociocultural information such as attitudes, values, and norms is more permanent in time and is valid across different media and physical spaces. Therefore, this type of knowledge can continuously be reused and reapplied in different projects. For instance, both the first and second studies have shown people’s daily life and online activities are driven by norms and
sociocultural values, which can always be considered in social media development.

As discussed in Section 2, most design research frameworks developed in HCI and related IT development generally do not focus on identifying user types or portraying users with the important dynamic and permanent attributes highlighted in the previous section. The design case studies have shown that ethnographic approaches, in-depth interviews, flow modeling, and multilevel social activity modeling can help development teams gain deeper insights, but that these methods are also time consuming and may require highly experienced researchers. Complex social media, service, and serious games development is challenging due to involving diverse user types, uncertain activity goals, and complex social interactions. Therefore, we argue that more comprehensive design research with a broader scope is needed to gain deeper insights, which can continue benefiting different phases in an iterative design cycle and help develop more adaptive and thoughtful systems.

6. Conclusion

Traditional user modeling mainly focuses on human-computer interaction. Although the traditional models reveal detailed information of the ways that people interact with systems, they are incapable of capturing many factors and contexts critical to design for social interaction, adaptive systems, and serious games. Personas, on the other hand, represent rich information about users’ lifestyles and attitudes and bring various benefits to design practices. However, there is little consensus regarding suitable methodology for how to develop reliable and representative personas. Both user modeling and personas narrow down their audience groups by overlooking activity contexts. User modeling only focuses on direct users of systems, and personas mainly represent target consumers, rather than describing the many participants involved in the activities.

In this research, we compare the strengths and weaknesses of different design and research methods. We, first, present additional types of user information that were gained through application of multiple research methods. Second, we discuss the benefits of having this information in a design process, including being able to identify invisible users, increase awareness of different user types, and develop more informative and representative user portraits. We also suggest suitable research methods in each project phase and list the important information needed to develop user portraits. In addition, we provide many examples through our cases to illustrate how the knowledge of user types can be applied in a design process, in particular in the development of large-scale services, adaptive systems, and serious games. We are aware that the cost of applying long-term ethnographic approaches and multiple modeling strategies to gain these insights is high. However, following scientific approaches can contribute high-quality and reliable user portraits with significant dynamic and static user information, which can later be reused and reapplied in different future projects.

References


Usability Testing for Serious Games: Making Informed Design Decisions with User Data

Pablo Moreno-Ger,1 Javier Torrente,1 Yichuan Grace Hsieh,2 and William T. Lester2

1 Facultad de Informática, Universidad Complutense de Madrid, 28040 Madrid, Spain
2 Laboratory of Computer Science, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02114, USA

Correspondence should be addressed to Pablo Moreno-Ger, pablom@fdi.ucm.es

Received 19 May 2012; Revised 9 October 2012; Accepted 11 October 2012

Academic Editor: Kiju Lee

Copyright © 2012 Pablo Moreno-Ger et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Usability testing is a key step in the successful design of new technologies and tools, ensuring that heterogeneous populations will be able to interact easily with innovative applications. While usability testing methods of productivity tools (e.g., text editors, spreadsheets, or management tools) are varied, widely available, and valuable, analyzing the usability of games, especially educational “serious” games, presents unique usability challenges. Because games are fundamentally different than general productivity tools, “traditional” usability instruments valid for productivity applications may fall short when used for serious games. In this work we present a methodology especially designed to facilitate usability testing for serious games, taking into account the specific needs of such applications and resulting in a systematically produced list of suggested improvements from large amounts of recorded gameplay data. This methodology was applied to a case study for a medical educational game, MasterMed, intended to improve patients’ medication knowledge. We present the results from this methodology applied to MasterMed and a summary of the central lessons learned that are likely useful for researchers who aim to tune and improve their own serious games before releasing them for the general public.

1. Introduction

As the complexity of new technologies increases, affecting wider portions of the population, usability testing is gaining even more relevance in the fields of human-computer interaction (HCI) and user interface (UI) design. Brilliant products run this risk of failing completely if end users cannot fully engage because of user interface failures. Consequently, product designers are increasingly focusing on usability testing during the prototype phase to identify design or implementation issues that might prevent users from successfully interacting with a final product.

Prototype usability testing is especially important when the system is to be used by a heterogeneous population or if this population includes individuals who are not accustomed to interacting with new technologies. In this sense, the field of serious games provides a good example where there should be special attention paid to usability issues.

Because educational serious games aim to engage players across meaningful learning activities, it is important to evaluate the dimensions of learning effectiveness, engagement, and the appropriateness of the design for a specific context and target audience [1]. Yet because serious games target broad audiences who may not play games regularly, usability issues alone can hinder the gameplay process negatively affecting the learning experience.

However, measuring the usability of such an interactive system is not always a straightforward process. Even though there are different heuristic instruments to measure usability with the help of experts [2]; these methods do not always identify all the pitfalls in a design [3]. Furthermore, usability is not an absolute concept per se but is instead relative in nature, dependent on both the task and the user. Consider the issue of complexity or usability across decades in age or across a spectrum of user educational backgrounds—what is usable for a young adult may not be usable for an octogenarian. It is situations like these where deep insight
into how the users will interact with the system is required. A common approach is to allow users to interact with a prototype while developers and designers observe how the user tries to figure out how to use the system, taking notes of the stumbling points and design errors [4].

However, prototype evaluation for usability testing can be cumbersome and may fail to identify comprehensively all of the stumbling points in a design. When usability testing sessions are recorded with audio and/or video, it can be difficult to simultaneously process both recorded user feedback and onscreen activity in a systematic way that will assure that all pitfalls are identified. Thus usability testing using prototype evaluation can be a time-consuming and error-prone task that is dependent on subjective individual variability.

In addition, many of the principles used to evaluate the usability of general software may not be necessarily applicable to (serious) games [5]. Games are expected to challenge users, making them explore, try, fail, and reflect. This cycle, along with explicit mechanisms for immediate feedback and perception of progress, is a key ingredient in game design, necessary for fun and engagement [6]. So the very context that makes a game engaging and powerful as a learning tool may adversely affect the applicability of traditional usability guidelines for serious games.

For example, typical usability guidelines for productivity software indicate that it should be trivial for the user to acquire a high level of competency using the tool, and that hesitation or finding a user uncertain about how to perform a task is always considered as unfortunate events. A serious game connects the pathways of exploration and trial and error loops to help the player acquire new knowledge and skills in the process [7]. This makes it imperative to differentiate hesitations and errors due to a bad UI design from actual trial and errors derived from the exploratory nature of discovering gameplay elements, a nuance typically overlooked using traditional usability testing tools.

In this paper we present a methodology for usability testing for serious games, building on previous instruments and extending them to address the specific traits of educational serious games. The methodology contemplates a process in which the interactions are recorded and then processed by multiple reviewers to produce a set of annotations that can be used to identify required changes and separate UI issues, game design issues, and gameplay exploration as different types of events.

Most importantly, a main objective of this methodology is to provide a structured approach to the identification of design issues early in the process, rather than to provide an instrument to validate a product achieving a “usability score”.

As a case study, this methodology was developed and employed to evaluate the usability of a serious game developed at the Massachusetts General Hospital’s Laboratory of Computer Science. “MasterMed” is a game designed to help the patients understand more about their prescribed medications and the conditions for which they are intended to treat. The application of this methodology using an actual game has helped us to understand better the strengths and limitations of usability studies in general and of this methodology in particular. From this experience, we have been able to synthesize the lessons learned about the assessment methodology that can be useful for serious games creators to improve their own serious games before releasing them.

2. Usability Testing and Serious Games

Usability is defined in the ISO 9241-11 as “the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use” [8]. This broad definition focuses on having products that allow the users to achieve goals and provides a base for measuring usability for different software products. However, digital games are a very specific type of software with unique requirements while serious games have the additional objective of knowledge discovery through exploratory learning. This presents unique usability challenges that are specific to serious games.

In this section we provide an overview of the main techniques for usability testing in general, and then we focus on the specific challenges posed by serious games.

2.1. Usability Testing Methods and Instruments. Usability represents an important yet often overlooked factor that impacts the use of every software product. While usability is often the intended goal when developing a software package, engineers tend to design following engineering criteria, often resulting in products that seem obvious in their functioning for the developers, but not for general users, with correspondingly negative results [9].

There are a variety of methods typically used to assess for usability. As described by Macleod and Rengger [4], these methods can be broadly catalogued as (i) expert methods, in which experienced evaluators identify potential pitfalls and usability issues, (ii) theoretical methods, in which theoretical models of tools and user behaviors are compared to predict usability issues, and (iii) user methods, in which software prototypes are given to end users to interact.

Among user methods, two main approaches exist: observational analysis, in which a user interacts with the system while the developers observe, and survey-based methods, in which the user fills in evaluation questionnaires after interacting with the system. Such questionnaires may also be used when applying expert methods, and they are typically based on heuristic rules that can help identify potential issues [10].

There are a number of survey-based metrics and evaluation methodologies for usability testing. A method most commonly cited is the System Usability Scale (SUS) because it is simple and relatively straightforward to apply [11]. SUS focuses on administering a very quick Likert-type questionnaire to users right after their interaction with the system, producing a “usability score” for the system. Another popular and well-supported tool, the Software Usability Measurement Inventory (SUMI), provides detailed evaluations [12] by measuring usability across five different dimensions (efficiency, affect, helpfulness, control,
and learnability). In turn, the Questionnaire for User Interaction Satisfaction (QUIS) [13] deals in terms more closely related with the technology (such as system capabilities, screen factors, and learning factors) with attention to demographics for selecting appropriate audiences. Finally, the ISO/IEC 9126 standard is probably the most comprehensive instrument, as described in detail in Jung and colleagues’ work [14].

However, many of these metrics suffer from the same weakness in that they can yield disparate results when reapplied to the same software package [15]. In addition, it is very common for such questionnaires and methods to focus on producing a usability score for the system, rather than the identification and remediation of the specific usability issues. This focus on identifying remediation actions as well as the prioritization of the issues and the actions surprisingly is often missing in studies and applications [16].

When the objective is to identify specific issues that may prevent end users from interacting successfully with the system, the most accurate approaches are observational user methods [4], as they provide direct examples of how the end users will use (or struggle to use) the applications. However, observational analysis requires the availability of fully functioning prototypes and can involve large amounts of observational data that requires processing and analysis. The experts may analyze the interaction directly during the session or, more commonly, rely on video recordings of the sessions to study the interaction. This has also led to considerations on the importance of having more than one expert review each interaction session. As discussed by Boring et al. [16], a single reviewer watching an interaction session has a small likelihood of identifying the majority of usability issues. The likelihood of discovering usability issues may be increased by having more than one expert review each session [17]; but this increased detection comes at the expense of time and human resources during the reviewing process.

In summary, usability testing is a mature field, with multiple approaches and instruments that have been used in a variety of contexts. All the approaches are valid and useful, although they provide different types of outcomes. In particular, observational user methods seem to be the most relevant when the objective is to identify design issues that may interfere with the user’s experience, which is the focus of this work. However, these methods present issues in terms of costs and the subjectivity of the data collected.

2.2. Measuring Usability in Serious Games. In the last ten years, digital game-based learning has grown from a small niche into a respected branch of technology-enhanced learning [18]. In addition, the next generation of educational technologies considers educational games (or serious games) as an instrument to be integrated in different formal and informal learning scenarios [19].

Different authors have discussed the great potential of serious games as learning tools. Games attract and maintain young students’ limited attention spans and provide meaningful learning experiences for both children and adults [20], while offering engaging activities for deeper learning experiences [21].

However, as games gain acceptance as a valid educational resources, game design, UI development, and rigorous usability testing are increasingly necessary. And while there are diverse research initiatives looking at how to evaluate the learning effectiveness of these games (e.g., [1, 22, 23]), the usability of serious games has received less attention in the literature. Designing games for “regular” gamers is reasonably straightforward, because games have their own language, UI conventions and control schemes. However, serious games are increasingly accessed by broad audiences that include nongamers, resulting occasionally in bad experiences because the target audience “does not get games” [24].

Designing for broad audiences and ensuring that a thorough usability analysis is performed can alleviate these bad experiences. In this context, Eladhari and Ollila conducted a recent survey on prototype evaluation techniques for games [25], acknowledging that the use of off-the-shelf HCI instruments would be possible, but that the instruments should be adapted to the specific characteristics of games as reported in [26]. In this context, there are some existing research efforts in adapting Heuristic Evaluations (with experts looking for specific issues) to the specific elements of commercial videogames [27, 28]. However, usability metrics and instruments for observational methods are not always appropriate or reliable for games. Most usability metrics were designed for general productivity tools, and thus they focus on aspects such as productivity, efficacy, and number of errors. But games (both serious or purely entertainment) are completely different, focusing more on the process than on the results, on enjoyment than on productivity, and on providing variety than on providing consistency [5].

Games engage users by presenting actual challenges, which demand exploratory thinking, experimentation, and observing outcomes. Ideally, this engagement cycle intends to keep the users just one step beyond their level of skill for compelling gameplay whereas a game that can be easily mastered and played through without making mistakes results in a boring game [6]. Therefore, usability metrics that reflect perfect performance and no “mistakes” (appropriate for productivity applications) would not be appropriate for (fun) games [29].

A similar effect can be observed with metrics that evaluate frustration. Games should be designed to be “pleasantly frustrating experiences”, challenging users beyond their skill, forcing users to fail, and therefore providing more satisfaction with victory [6]. In fact, the games that provide this pleasantly frustrating feeling are the games that are the most addicting and compelling. On the other hand, there are games that frustrate players because of poor UI design. In these cases, while the user is still unable to accomplish the game’s objectives, failure is the result of bad UI or flawed game concepts. Usability metrics for serious games should distinguish in-game frustration from at-game frustration [30], as well as contemplating that “obstacles for accomplishment” may be desirable, while “obstacles for fun” are not [5].
Unfortunately, as game designers can acknowledge, there is no specific recipe for fun, and as teachers and educators can acknowledge, eliciting active learning is an elusive target. The usability and effectiveness of productivity tools can be measured in terms of production, throughput, efficacy, and efficiency. But other aspects such as learning impact, engagement, or fun are much more subjective and difficult to measure [31].

This subjectivity and elusiveness impacts formal usability testing protocols when applied to games. As White and colleagues found [32], when different experts evaluated the same game experiences (with the same test subjects), the results were greatly disparate, a problem that they attributed to the subjective perception of what made things “work” in a game.

In summary, evaluating the usability of games presents unique challenges and requires metrics and methodologies that aim to contemplate their variability and subjectivity of interacting with games, as well as their uniqueness as exploratory experiences that should be pleasantly frustrating.

3. General Methodology

As discussed in the previous section, gathering data to evaluate the usability of a serious game is an open-ended task with different possible approaches and several potential pitfalls. Therefore, there is a need for straightforward and reliable methods that help developers identify usability issues for their serious games before releasing them. In our specific case, we focus on facilitating an iterative analysis process based on observational methods, in which users play with early prototypes and researchers gather data with the objective of identifying and resolving design and UI issues that affect the usability of the games.

3.1. Requirements. From the discussion above it is possible to identify some initial requirements to perform usability testing of serious games.

(1) Test Users. First, it is necessary to have a set of test users to evaluate the prototype. These test users should ideally reflect the serious game’s target audience in terms of age, gender, education, and any other demographic characteristics that might be unique or pertinent to the educational objective of the serious game. In terms of number of test users, according to Virzi [33], five users should be enough to detect 80% of the usability problems, with additional testers discovering a few additional problems. In turn, Nielsen and Landauer [34] suggested that, for a “medium” sized project, up to 16 test users would be worth some extra cost, but any additional test users would yield no new information. They also suggested that the maximum benefit/cost ratio would be achieved with four testers. We suggest selecting at least as many users that would span the range of your target audience, but not so many users that hinder the team performing the usability data analysis.

(2) Prototype Session Evaluators. Another important requirement is the consideration of the numbers of evaluators or raters to analyze the play session of each test user. Having multiple evaluators significantly increases the cost, making it tempting to use a single evaluator. However, while some analyses are performed with a single evaluator observing and reviewing a test user’s play data, Kessner and colleagues suggested that it is necessary to have more than one evaluator to increase the reliability of the analysis, because different evaluators identified different issues [3]. This effect is even stronger when evaluating a game, because their high complexity results in evaluators interpreting different causes (and therefore possible solutions) for the problems [32]. Therefore, we suggest having more than one evaluator to analyze each play session and a process of conciliation to aggregate the results.

(3) Instrument for Serious Game Usability Evaluation. For an evaluator who is analyzing a play session and trying to identify issues and stumbling points, a structured method for annotating events with appropriate categories is a necessity [17]. Because serious games differ from traditional software packages in many ways, we suggest using an instrument that is dedicated to the evaluation of serious game usability. Section 3.2 below is dedicated to the development of a Serious Game Usability Evaluator (SeGUE).

(4) Data Recording Setup. Nuanced user interactions can often be subtle, nonverbal, fast paced, and unpredictable. A real-time annotation process can be burdensome, or perhaps even physically impossible if the user is interacting with the system rapidly. In addition, any simultaneous annotation process could be distracting to the user’s game interactions and detract from the evaluative process. For these reasons, we recommend screen casting of the test play sessions along with audio and video recordings of the user with minimal, if any coaching, from the evaluation staff. These recordings can be viewed and annotated later at an appropriate pace.

(5) “Ready-to-Play” Prototype. “Ready-to-Play” Prototype should be as close to the final product as possible for the test users to evaluate. The prototype should allow the test users to experience the interface as well as all intended functionalities so that the interactions could mimic the real play session, therefore, maximizing the benefits of conducting a usability test. When it is not feasible or cost effective to provide a full prototype, using an early incomplete prototype may fail to reflect the usability of the final product once it has been polished. White and colleagues [32] conducted their usability studies using a “vertical slice quality” approach, in which a specific portion of the game (a level) was developed to a level of quality and polish equivalent to the final version.

(6) Goal-Oriented Play-Session Script. Lastly, prior to the initiation of the study, a play-session script should be determined. The script for the evaluation session should be relatively brief and have clear objectives. The designers should prepare a script indicating which tasks the tester
is expected to perform. In the case of a serious game, this script should be driven by specific learning goals, as well as cover all the relevant gameplay elements within the design. There may be a need for more than one play session to be exposed to each user so that all the key game objectives could be included.

3.2. Development of the Serious Game Usability Evaluator (SeGUE). Evaluators who analyze a prototype play session will need a structured method to annotate events as they try to identify issues and stumbling points. This predefined set of event types is necessary to facilitate the annotation process as well as to provide structure for the posterior data analysis. This evaluation method should reflect the fact that the objective is to evaluate a serious game, rather than a productivity tool. As described in Section 2.2, serious games are distinct from other types of software in many ways. Importantly, serious games are useful educational resources because they engage the players on a path of knowledge discovery. This implies that the evaluation should focus on identifying not only those features representing a usability issue, but also the ones that really engage the user.

Since the objectives of evaluating a serious game not only focus on the prototype itself but also the process of interacting with the game and the user’s experience, our research team developed a tool, the Serious Game Usability Evaluator (SeGUE), for the evaluation of serious game usability. The SeGUE was derived and refined using two randomly selected serious game evaluation sessions, in which a team comprising game programmers, educational game designers, and interaction experts watched and discussed videos of users interacting with an educational serious game. Two dimensions (system related and user related) of categories were created for annotation purposes. Within each dimension, several categories and terms were defined to annotate events.

Within the system-related dimension, there are six different event categories. Two event categories are related to the game design, including gameflow and functionality. Events of these categories are expected to require deep changes in the game, perhaps even the core gameplay design. Three event categories are related to the game interface and implementation, including content, layout/UI, and technical errors, where solutions are expected to be rather superficial and have less impact on the game. A nonapplicable category is also considered for events not directly related to the system, but still deemed relevant for improving the user experience.

In the user-related dimension, there are ten event categories across a spectrum of emotions: negative (frustrated, confused, annoyed, unable to continue), positive (learning, reflecting, satisfied/excited, pleasantly frustrated), or neutral (nonapplicable and suggestion/comment). For researchers’ convenience an additional category named “other” was included in both dimensions for those events that were hard to categorize. Such events may be an indication that a new category is required due to specific traits of a specific game. More details about the categories and their meanings are presented in Tables 1 and 2.

3.3. Evaluation Process. We present here a step-by-step methodology to assess for usability events in serious games. Additionally we will show as a case study how we employed this methodology to assess for usability while accounting for the MasterMed game’s specific learning objectives. According to the requirements described above, the methodology is organized in discrete stages, from the performance of the tests to the final preparation of a list of required changes. The stages of the methodology are as follows.

(1) Design of the Play Session. The evaluation session should be brief and have clear objectives. The designers should prepare a detailed script indicating which tasks the tester is expected to perform. This script should be driven by specific learning goals, as well as include all the relevant gameplay and UI elements within the design. There may be a need for more than one scripted play session to cover all the key objectives.

(2) Selection of the Testers. As noted above, invited testers’ characteristics should closely represent the intended users and mimic the context for which the serious game is designed.

(3) Performance and Recording of the Play Sessions. The testers are given brief instructions about the context of the game and the learning objectives and prompted to play the game on their own, without any further directions or instructions. The testers are instructed to speak out loud while they play, voicing out their thoughts. During the play session, the evaluator does not provide any instructions unless the user is fatally stuck or unable to continue. Ideally, the session is recorded on video, simultaneously capturing both the screen and the user’s verbal and nonverbal reactions.

(4) Application of the Instrument and Annotation of the Results. In this stage, the evaluators review the play sessions identifying and annotating all significant events. An event is a significant moment in the game where the user found an issue or reacted visibly to the game. Events are most commonly negative events, reflecting a usability problem, although remarkably positive user reactions should also be tagged, as they indicate game design aspects that are engaging the user and should be enforced. Each event is tagged according to the two dimensions proposed in the SeGUE annotation instrument (Section 3.2). Ideally each play session should be annotated by at least two evaluators separately.

(5) Reconciliation of the Results. Since multiple reviewers should annotate the videos independently, the annotations and classifications likely will end up being different. Therefore, it is necessary for all of the reviewers to confer for reconciliation of the results. There are several possibilities that result from initial discrepant event assessments: (1) an observed event may be equally recognized by multiple reviewers with identical tagging; (2) a single event might be interpreted and tagged differently by at least one reviewer;
Table 1: Event categories for the system dimension.

<table>
<thead>
<tr>
<th>System-related event</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Functionality</strong></td>
</tr>
<tr>
<td>An event is related to prototype’s functionality when it is the result of the user activating a control item and it is related to one specific action.</td>
</tr>
<tr>
<td><strong>Layout/UI</strong></td>
</tr>
<tr>
<td>An event is related to layout/UI when the user makes a wrong assumption about what a control does, or when the user does not know how to do something (negative events). It is also a layout/UI positive event when a user appreciates the design (figures, attempts, colors, etc.) or having specific information displayed.</td>
</tr>
<tr>
<td><strong>Gameflow</strong></td>
</tr>
<tr>
<td>An event that is caused not by a single specific interaction, but as a consequence of the game sequences interactions and outputs and the specific gameplay design of the game.</td>
</tr>
<tr>
<td><strong>Content</strong></td>
</tr>
<tr>
<td>A content event is related to text blurbs and other forms of textual information provided by the game.</td>
</tr>
<tr>
<td><strong>Technical error</strong></td>
</tr>
<tr>
<td>A technical error event is related to a nonintentional glitch in the system that must be corrected.</td>
</tr>
<tr>
<td><strong>Nonapplicable</strong></td>
</tr>
<tr>
<td>When the event is not related to the system and/or not prompted by a system behavior.</td>
</tr>
<tr>
<td><strong>Other</strong></td>
</tr>
<tr>
<td>An event that is related to the system, but does not match any of the above (this suggests that a new category is needed).</td>
</tr>
</tbody>
</table>

Table 2: Event categories for the user dimension.

<table>
<thead>
<tr>
<th>User-related event</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Learning</strong></td>
</tr>
<tr>
<td>The user figures out how to perform an action that was unclear before (learn to play), or when the user is actively engaging in consuming content (learn content).</td>
</tr>
<tr>
<td><strong>Reflecting</strong></td>
</tr>
<tr>
<td>The user pauses or wonders what to do next. Unlike when the user is confused and does not know what to do, reflecting events indicate pause to create action plans within the game space.</td>
</tr>
<tr>
<td><strong>Satisfied/excited</strong></td>
</tr>
<tr>
<td>The user displays a remarkably positive reaction.</td>
</tr>
<tr>
<td><strong>Pleasantly frustrated</strong></td>
</tr>
<tr>
<td>The user expresses frustration in a positive manner. A pleasantly frustrating moment urges the user to try to overcome the obstacle again.</td>
</tr>
<tr>
<td><strong>Frustrated</strong></td>
</tr>
<tr>
<td>The user voices or displays negative feelings at not being able to complete the game or not knowing how to do something. A frustrating moment urges the player to stop playing.</td>
</tr>
<tr>
<td><strong>Confused</strong></td>
</tr>
<tr>
<td>The user does not know how to perform an action, misinterprets instructions, and/or does not know what he/she is supposed to do.</td>
</tr>
<tr>
<td><strong>Annoyed</strong></td>
</tr>
<tr>
<td>The user performs properly a task in the game (knows how to do it), but feels negatively about having to do it.</td>
</tr>
<tr>
<td><strong>Unable to continue (fatal)</strong></td>
</tr>
<tr>
<td>This is usually the consequence of one or more of the above, or of a fatal technical error. An event is related to when the user becomes definitely stuck and/or cannot continue without the help of the researcher. Such events are highlighted because the origin of these events must always be resolved.</td>
</tr>
<tr>
<td><strong>Nonapplicable</strong></td>
</tr>
<tr>
<td>An event is not related to the user (e.g., it is a remark by the researcher, or a glitch appeared but the user did not notice it).</td>
</tr>
<tr>
<td><strong>Suggestion/comment</strong></td>
</tr>
<tr>
<td>The user verbalizes a comment or a suggestion that is not related to a specific interaction or event.</td>
</tr>
<tr>
<td><strong>Other</strong></td>
</tr>
<tr>
<td>An event is related to the user, but does not match any of the above (this suggests that a new category is needed).</td>
</tr>
</tbody>
</table>

or (3) an event could be recognized and tagged by one observer and overlooked by another. In the latter two cases, it is important to have all the reviewers to verify and agree on the significance of the event and have subjective agreement on the proper tag. Most importantly, the objective of this task is not to increase the interrater reliability, but to study collaboratively the event in order to better understand its interpretation, causes, and potential remediation actions.

(6) Preparation of a Task List of Changes. Finally, the eventual product from this evaluation process should be a list of potential improvements for the game, with an indication of their importance in terms of how often the problem appeared and how severely it affected the user or interfered with the game’s educational mission. For each observed negative event, a remediation action is proposed. Changes proposed should avoid interfering with the design and game-play elements that originate positive events to maintain engagement. Users’ comments and suggestions may also be taken into account. Quite possibly, some of the encountered issues will occur across multiple users, and some events might occur multiple times for the same user during the same play session.
(e.g., a user may fail repeatedly to activate the same control). For each action point there will be a frequency value (how many events were recorded that suggest this action point) and a spread value (how many users were affected by this issue).

Finally after reconciliation, the evaluation team should have an exhaustive list of potential changes. For each modification, the frequency, the spread, and a list of descriptions of when the event happened for each user all contribute to the estimating of importance and urgency for each action, as it may not be feasible to implement every single remediation action.

It must be noted that although a predefined set of tagging categories facilitate the annotation and reconciliation process, the work performed in stages 4 and 5 can be labor intensive and time consuming depending on the nature and quantity of the test user’s verbal and non-verbal interactions with the prototype.

Finally, depending on the scope and budget of the project, it may be appropriate to iterate this process. This is especially important if the changes in the design were major, as these changes may have introduced further usability issues that had not been previously detected.

4. Case Study: Evaluating MasterMed

This SeGUE methodology, including the specific annotation categories, has been put to the test with a specific serious game (MasterMed) (see Figure 1), currently being developed at Massachusetts General Hospital’s Laboratory or Computer Science. The goal of MasterMed is to educate patients about the medications they are taking by asking patients to match each medication with the condition it is intended to treat. The game will be made available to patients via an online patient portal, iHealthSpace (https://www.ihealthspace.org/portal/login/index.html), for patients who regularly take more than three medications. The target audience for this game is therefore a broad and somewhat older population that will be able to use computers, but not necessarily technically savvy. This makes it very important to conduct extensive usability studies with users similar to the target audience, to ensure that patients will be able to interact adequately with the game.

Performing an indepth evaluation of the MasterMed game helped us refine and improve the evaluation methodology, gaining insight into the importance of multiple reviewers, the effect of different user types in the evaluation, or how many users and reviewers are required. In addition, the experience helped improve the definitions of the categories in the SeGUE instrument.

In this section we describe the case study, including the study setup, the decisions made during the process, and the results gathered. From these results, we have extracted the key lessons learned on serious game usability testing, and those lessons are described in Section 5.

4.1. Case Study Setup

4.1.1. Design of the Play Session. The session followed a script, in which each participant was presented three increasingly difficult scenarios with a selection of medications and problems to be matched. The scenarios covered simple cases, where all the medicines were to be matched, and complex cases in which some medicines did not correspond with any of the displayed problems. In addition, we focused on common medication for chronic problems and included in the list potentially problematic medications and problems, including those with difficult or uncommon names. As a user progressed through the script, new UI elements were introduced sequentially across sessions. The total playing time was estimated to be around 30 minutes.

4.1.2. Selection of the Testers. Human subject approval was obtained from the Institutional Review Board of Partners Human Research Committee, Massachusetts General Hospital’s parent institution. The usability testing used a convenience sampling method to recruit ten patient-like participants from the Laboratory of Computer Science, Massachusetts General Hospital. An invitation email message contained a brief description of the study, eligibility criteria, and contact information was sent out to all potential participants. Eligible participants were at least 18 years old and not working as medical providers (physicians or nurses). Based on a database query, our expected patient-gamer population should be balanced in terms of gender with roughly 54% of participants are female. Patient age ranges from 26 to 103 with a mean of 69.3 years (SD = 12.5) for men and a mean of 70.14 years (SD = 12.75) for women. We recruited five men and five women with their age ranged from mid-30 s to 60 s to evaluate the game.

4.1.3. Performance and Recording of the Play Sessions. Each participant was asked to interact with the game using a think-aloud technique during the session. The screen and participant’s voice and face were recorded using screen/webcam capture software. The duration of the play sessions ranged between 40 and 90 minutes.

4.1.4. Application of the Instrument and Annotation of the Results. After conducting the sessions, a team of evaluators was gathered to annotate the videos identifying all potentially significant events. There were four researchers available, two from the medical team and two from the technical team. Five videos were randomly assigned to each researcher to review; thus two different researchers processed each video independently. In order to avoid any biasing factors due to the backgrounds of each researcher, the assignment was made so that each researcher was matched to each of the other three researchers at least once. The annotations used the matrix described in Section 3.2. Two more fields were added to include a user quote when available and comments describing the event in more detail.

4.1.5. Reconciliation of the Results. The reconciliation was performed in a meeting with all four researchers, where (i) each unique event was identified and agreed upon, (ii) each matched event classified differently was reconciled, and (iii) each matched event with the same tags was reviewed.
for completeness. This process was crucial in determining the nature of overlooked events and facilitated the discussion on the possible causes for those events that had been tagged differently by the reviewers.

### 4.1.6. Preparation of a Task List of Changes.

For each observed negative event, a remediation action was proposed and prioritized.

### 4.2. Case Study Results.

The first artifact of the case study was a set of 10 video files resulting from the screen/webcam capture software. Since the evaluation method was experimental, two randomly selected videos were used for a first collaborative annotation process. This step helped refine and improve the tags described in Section 3.2. Therefore, the final evaluation was performed only on the eight remaining play sessions.

The average play session was around 30 minutes in length, although most users took between 40 and 60 minutes (and only one user as much as 90 minutes). A total of 290 events were logged. We summarize the events identified for each user (see Figure 2). A unique event is defined when the event was only tagged by one of the two researchers reviewing the video (and overlooked by the other). A matched event is defined when the event was tagged by both researchers and classified equally with the same tags and interpretation. Finally, a reconciled event is defined when the event was identified by two researchers, but tagged differently and then agreed upon during the reconciliation process.

In Figure 3, we summarize the number of appearances of each tag and the relative frequencies for each event type. The number of negative events (138) was much higher than positive events (46). Also the number of interface and implementation events (179) is greater than events related to design (91).

Finally, in Table 3 we provide an excerpt of the action points that were derived from the analysis of the results. For each action, we also indicate the frequency (number of events that would be solved by this action) and the spread (number of users that encountered an event that would be solved by this action). Both numbers were used to determine the priority of each action.

### 4.3. Case Study Discussion.

An interesting aspect for discussion is the variability of event statistics across users. Figure 2 is sorted according to the number of unique events, as this category requires special attention. Indeed, while a reconciled event indicates an event that was perceived different by each researcher, a unique event indicates that one of the researchers overlooked the event. In a scenario with only one reviewer per play session, such events may have gone unnoticed. The annotations for some users presented very high numbers of unique events. It is possible that this is related to the total number of events, affecting the subjective thresholds of the reviewers when the frequency of events is high. However, the results do not suggest that a correlation between the total number of events and the proportion of unique, matched, and reconciled events. For example, results from users with small total number of events vary, as user no. 2 presents 77.78% unique events while user no. 1 has only 30.77% unique events.

Regarding the tag statistics, the number of negative events in the user dimension is clearly predominant. This result may be considered normal, as evaluators are actively
Table 3: Excerpt of the prioritized action points list. It shows the type (D: design/I: interface), the frequency (number of occurrences), the spread (number of users affected), and the priority they were given according to these two numbers.

<table>
<thead>
<tr>
<th>Priority</th>
<th>Action</th>
<th>Type</th>
<th>Freq.</th>
<th>Spread</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Rearrange the tutorials (shortening and skipping)</td>
<td>D</td>
<td>28</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>Remove “none of the above” feature</td>
<td>D</td>
<td>23</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>Unify “close dialog” interactions</td>
<td>I</td>
<td>37</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>UI tweaking (color schemes, minor layout changes, etc.)</td>
<td>I</td>
<td>22</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>Review wording</td>
<td>I</td>
<td>13</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>Improve mouse clicking accuracy</td>
<td>I</td>
<td>11</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>Improve handout contents (remove unnecessary sections)</td>
<td>I</td>
<td>11</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>User</th>
<th>Unique</th>
<th>Matched</th>
<th>Reconciled</th>
</tr>
</thead>
<tbody>
<tr>
<td>User 7</td>
<td>16.44%</td>
<td>57.53%</td>
<td>26.03%</td>
</tr>
<tr>
<td>User 8</td>
<td>28.13%</td>
<td>40.63%</td>
<td>31.25%</td>
</tr>
<tr>
<td>User 1</td>
<td>30.77%</td>
<td>53.33%</td>
<td>15.38%</td>
</tr>
<tr>
<td>User 3</td>
<td>33.33%</td>
<td>17.54%</td>
<td>49.12%</td>
</tr>
<tr>
<td>User 10</td>
<td>35.29%</td>
<td>29.41%</td>
<td>35.29%</td>
</tr>
<tr>
<td>User 4</td>
<td>40.63%</td>
<td>27.12%</td>
<td>32.20%</td>
</tr>
<tr>
<td>User 5</td>
<td>45.02%</td>
<td>46.67%</td>
<td>8.31%</td>
</tr>
<tr>
<td>User 2</td>
<td>77.78%</td>
<td>11.11%</td>
<td>11.11%</td>
</tr>
</tbody>
</table>

Figure 2: Event statistics. Each bar shows the percentage of unique, matched and reconciled events for each individual user. The total number of events for each user is shown in parenthesis.

Figure 3: Tag statistics: the events are categorized on two dimensions: the source of the event (interface, design) and the reaction of the user.
looking for issues and pitfalls, while regular play working as intended may not be considered as an event. However, the identification of specific positive events was still helpful to identify specific game moments or interactions that really engaged the users in a visible way.

In the game element dimension, the number of events related to the design of the game was significantly less than the number of events related to the interface and implementation (91 versus 179). This data suggests that users were more satisfied with the flow and mechanics of the MasterMed game than with its look and feel. Nonetheless, this difference seems reasonable, as it is easier for users to identify pitfalls in superficial elements like the UI (e.g., font size is too small) then in the design (e.g., the pacing is not appropriate). The correlation between user and system dimensions is also interesting, as positive events are usually related to aspects of the game design. Since the gameplay design is the key element for engagement, this result may be considered an indication that the design was, in fact, successful.

The process to determine the remediation actions and a heuristic assessment of their importance deserves also some discussion. The prioritization of the list is not fully automatable. While the frequency was an important aspect to consider (an event that happened many times), so was the spread (an event that affected many users). These variables allowed researchers to limit the impact of having multiple occurrences of the same event for a single user. A specific example: the action “remove none of the above feature” was regarded as more important than “unify close dialog interactions” because it affected all users, even though the total number of occurrences was significantly lower (23 versus 37).

Other factors such as the cost of implementing a change or its potential return were not considered, but large projects with limited budget or time constraints may need to consider these aspects when prioritizing the remediation actions.

5. Lessons Learned

The result of the case study not only helped to identify improvement points, but also served as a test to improve and refine the SeGUE instrument for annotation. Some design decisions, taken on the base of the existing literature, were put to the test in a real study, which allowed us to draw important conclusions. And these conclusions are helpful for researchers using this methodology (or other variations) to evaluate and improve their own serious games. The main lessons learned are summarized below.

5.1. Multiple Evaluators. As discussed in Section 3.1, different studies have taken different stances when it comes to how many researchers should review and annotate each play session. The key aspect is to make sure that all usability issues are accounted for (or as many as possible).

The interrater reliability displayed by the results for our case study is, in fact, very low (Figure 2). Both matched and reconciled events were identified by both reviewers, but unique events were only registered by one of the reviewers. For most users, the number of unique events is between 33% and 50%, giving a rough estimate of how many events may have been lost if only one reviewer had been focusing on one play session (user no. 2 has an unusually high number of unique events).

This result is consistent with the concerns expressed by White and colleagues [32] and confirms the importance of having multiple evaluators for each play session in order to maximize the identification of potential issues. While it might be very tempting for small-sized teams to use only one annotator per gameplay session to reduce costs, our experience shows that even after joint training the number of recorded unique events is high. Thus, multiple evaluators should be considered as a priority when planning for usability testing.

5.2. Importance of Think-Aloud Methods. Most observational methods do not explicitly require users to verbalize their thoughts as they navigate the software, as it is considered that the careful analysis of the recordings will suffice to identify usability issues, even with only one expert reviewing each recording.

However, the results from the case study indicate the importance of requesting (and reinforcing) users to think aloud while they play. For our case study MasterMed evaluation, there was a direct correlation between the number of unique events tagged and the amount of comments verbalized by users. While all users were instructed to verbalize their thoughts, not all users responded equally. On one extreme, user no. 7 was loquacious, providing a continuous stream of thoughts and comments. On the other extreme, user no. 2 was stoic, apparently uncomfortable expressing hesitations out loud, rarely speaking during the experiment, despite of being reminded by the researcher about the importance of commenting. This had a direct impact in the number of unique events (16.44% unique events registered for user no. 7 and 77.78% unique events for user no. 2), as it made it difficult for the researchers to distinguish between hesitations caused by a usability issue from actual pauses to think about the next move in the game.

5.3. Length of the Play Sessions. The length of the play sessions was estimated to be around 30 minutes, although the range was 40–90 minutes. During the play session, familiarity with the tool and its expected behaviors may improve, and this may mean that most usability issues would be detected in the first minutes of a play session. To get a better insight about this issue, we produced the event timestamp frequency histogram provided below in Figure 4. Most of the events were tagged during the first 13 minutes of the session (44.06%) after which the rate decreases, with only 24.95% of the events tagged in the following 13 minutes. Beyond this point, the rate slowed even further, even though new, more complex gameplay scenarios were being tested.

Users are also encouraged to verbalize their impressions and explain their reasoning when deciding the next move or interaction; but as the play session becomes longer, the users also grow tired. This suggests that play sessions
should be kept short and focused. It should also be noted that researchers observing recorded play sessions thoroughly needed to stop, rewind, and review video footage frequently to tag the issues encountered, thereby requiring lengthy evaluation sessions. When more than 30 minutes are required to explore all the concepts, different sessions with breaks may be desirable.

5.4. Evaluator Profile. Even though the proposed methodology called for multiple experts evaluating each play session, we have found differences between the annotations depending on the researcher’s profile. The foremost difference was between technical experts (developers) and field experts (clinicians).

Technical issues were one of the main sources of events that had to be reconciled (cases in which both researchers tagged the same event, but assigned different categories). Developers would spot subtle technical issues and tag them accordingly, while clinicians often attributed those events to usability problems related to the UI. This does not necessarily mean that an effort should be made to assign field experts and technicians to review each play session (although it may be desirable). However, it does reflect the importance of having experts from all sides participating in the reconciliation stage. In particular, the goal of the reconciliation stage is not necessarily to agree on the specific category of the event, but on its origin, impact on the user experience and significance; so that appropriate remediation actions can be pursued based on the data gathered.

5.5. Limitations. The methodology has a very specific objective: to facilitate the identification of design pitfalls in order to improve the usability of a serious game. As such, it does not deal with other very important dimensions of user assessment in serious games. In particular, it cannot be used to guarantee that the game will be effective in engaging the target audience or to assess the learning effectiveness of the final product. While the methodology takes care of identifying those elements that are especially engaging, this is done in order to help the designers preserve the elements with good value when other design or UI issues are addressed. Before the final version of the game is released for the general public, further assessment of engagement and learning effectiveness should be conducted.

Another limitation that this methodology shares with typical observational methods (and in particular with think-aloud methods) is that the results are subjective and dependant on both the specific users and the subjective interpretations from the evaluators. The subjectivity of the process was highlighted in the case study in the number of events overlooked by at least one reviewer (number of unique events) and the discrepancies when annotating the perceived root cause of each event. While this subjectivity could be reduced by increasing the number of users and evaluators, this increases the cost of the evaluation process. This problem is further aggravated when the process is applied iteratively.

Small and medium sized development projects will need to carefully balance the number of users, evaluators, and iterations depending on their budget, although we consider that having more than one evaluator for each session is essential. Similarly, multiple iterations may be required if the changes performed affect the design or UI significantly, potentially generating new usability issues. In turn, bigger projects with enough budget may want to complement the observational methods by tracking physiological signals (e.g., eye tracking, electrocardiogram, brain activity) to gather additional insight into engagement. However, such advanced measurements fall beyond the scope of this work, which targets smaller game development projects with limited budgets.

6. Conclusions

The design of serious games for education is a complex task in which designers need to create products that engage the audience and provide an engaging learning experience, weaving gameplay features with educational materials. In addition, as with any software product targeting a broad audience, the usability of the resulting games is important. In this work we have discussed the unique challenges that appear when we try to evaluate the usability of a serious game before its distribution to a wide, nongamer audience. The key challenge is that typical usability testing methods focus on measurements that are not necessarily appropriate for games, focusing on aspects such as high productivity, efficacy, and efficiency as well as low variability, number of errors, and pauses. However, games contemplate reflection, exploration, variety and trial, and error activities.

While generic heuristic evaluative methods can be adapted to contemplate the specificities of games, observational instruments that generate metrics and scores are not directly applicable to serious games. In addition, observational data is by definition subjective, making it difficult to translate a handful of recorded play sessions into a prioritized list of required changes.

For these reasons, we have proposed a step-by-step methodology to evaluate the usability of serious games that focuses on obtaining a list of action points, rather than
a single score that can be used to validate a specific game. Observational methods can be useful in determining design pitfalls but, as we have described in the paper, the process is subjective and sometimes cumbersome. The methodology provides a structured workflow to analyze observational data, process it with an instrument designed specifically for serious games, and derive a list of action points with indicators of the priority for each change, thus reducing the subjectivity of the evaluative process.

The Serious Games Usability Evaluator (SeGUE) instrument contemplates tagging events in the recorded play sessions according to two dimensions: the system and the user. Each observed event has an identifiable cause from a certain interaction or UI element and effect on the user (confusion, frustration, excitement, etc.). The categories for each dimension contemplate aspects specifically related to serious games, distinguishing, for example, between in-game frustration (a positive effect within the description of games as “pleasantly frustrating experiences”) and at-game frustration (a negative event when the game interface, rather than the game design, becomes a barrier for achieving objectives).

The inclusion of positive events is relevant when studying the usability of serious games. These games need to engage users by both presenting challenges and variability and achieving a learning objective. The events in which the users are engaging intensively with the game (displaying excitement or pleasant frustration) are important parts of the game-flow, and the action points to improve usability should be designed such that they do not dilute the engagement.

The application of the SeGUE methodology in the MasterMed case study allowed us to draw some conclusions and summarize important lessons learned during the process, as summarized in Section 5. Among them, the experience provided answers to typically open questions regarding observational methods such as (a) the appropriate number of test subjects, (b) number of experts to review each play session, and (c) the importance of the think-aloud technique.

We expect the methodology, the SeGUE tagging instrument, and the summary of lessons learned to be useful for researchers who aim to improve the usability of their own serious games before releasing them. Small- and medium-sized projects can use this methodology to test the usability of their games, record data that is typically subjective and difficult to process, and then follow a structured methodology to process the data. The number of evaluation cycles, the specific designs, and the aspects of the games that need to be evaluated may vary across development projects. Therefore, these steps and the SeGUE instrument might be adapted and/or refined to incorporate any particular elements required by specific serious game developments.

Acknowledgments

This project was funded by the Partners Community Healthcare, Inc. System Improvement Grant program as well as the European Commission, through the 7th Framework Programme (project “GALA-Network of Excellence in Serious Games" -FP7-ICT-2009-5-258169) and the Life-long Learning Programme (projects SEGAN-519332-LLP-1-2011-1-PT-KA3-KA3NW and CHERMUG 519023-LLP-1-2011-1-UK-KA3-KA3MP).

References


J. Kirriemuir and A. McFarlane, Literature Review in Games and Learning, NESTA Futurelab, Bristol, UK, 2004.


Research Article

Assessment of Learners’ Motivation during Interactions with Serious Games: A Study of Some Motivational Strategies in Food-Force

Lotfi Derbali and Claude Frasson

Département d’Informatique et de Recherche Opérationnelle, Université de Montréal, C.P. 6128, Succursale Centre-Ville, Montréal, QC, Canada H3C 3J7

Correspondence should be addressed to Lotfi Derbali, derbalil@iro.umontreal.ca

Received 9 April 2012; Revised 25 September 2012; Accepted 8 October 2012

Academic Editor: Francesco Bellotti

Copyright © 2012 L. Derbali and C. Frasson. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This study investigated motivational strategies and the assessment of learners’ motivation during serious gameplay. Identifying and intelligently assessing the effects that these strategies may have on learners are particularly relevant for educational computer-based systems. We proposed, therefore, the use of physiological sensors, namely, heart rate, skin conductance, and electroencephalogram (EEG), as well as a theoretical model of motivation (Keller’s ARCS model) to evaluate six motivational strategies selected from a serious game called Food-Force. Results from nonparametric tests and logistic regressions supported the hypothesis that physiological patterns and their evolution are suitable tools to directly and reliably assess the effects of selected strategies on learners’ motivation. They showed that specific EEG “attention ratio” was a significant predictor of learners’ motivation and could relevantly evaluate motivational strategies, especially those associated with the Attention and Confidence categories of the ARCS model of motivation. Serious games and intelligent systems can greatly benefit from using these results to enhance and adapt their interventions.

1. Introduction

It is widely acknowledged that learners’ psychological and cognitive states have an important role in intelligent systems and serious games (SGs). For instance, engagement and motivation or disaffection and boredom obviously affect learners’ wills and skills in acquiring new knowledge [1]. SGs cannot, therefore, ignore these states and should take them into account during learning process. One important learners’ state is motivation which plays a crucial role in both the learners’ performance and the use of intelligent systems over time [2]. Motivation is generally defined as that which explains the direction and magnitude of behaviour, or in other words, it explains what goals people choose to pursue and how they pursue them [3]. It is considered as a natural part of any learning process. Several researches have showed that motivated learners are more likely to be more engaged, to undertake challenging activities, and to exhibit enhanced performance and outcomes [4, 5]. Therefore, it is of particular relevance to study motivation and its role in improving learners’ performance during gameplay.

Learners’ interactions with Intelligent Tutoring Systems (ITSs) and especially SGs have always been considered to be intrinsically motivating. One possible explanation is the fact that ITSs generally use pictures, sounds, videos, and so forth, that are considered, crudely, as motivational factors. Intrinsic motivation is possibly gained through challenge, curiosity, control, sensory stimuli, interaction, and fantasy when using SGs [6, 7]. However, many researchers have argued that learners’ negative emotions or amotivational states such as boredom or disengagement have been known to appear following a certain period of interaction with computer systems. These states can be overwhelming to learners and cause motivational problems and decrease learning benefits [2, 8, 9]. Once learners’ psychological and cognitive states are identified, intelligent systems are in
a much better position to act upon them. In this perspective, several studies have described intelligent systems that can provide adapted emotional or cognitive strategies for coping with, or at least reducing, the negative learners’ states [8–12]. Computer systems can also use motivational strategies which are the actions (or tactics) taken in order to scaffold learners’ motivation toward tasks and goals of learning process and to make learning easier, faster, more enjoyable, more self-directed, and more effective. However, it is surprising to find so little mention of the motivational strategies and relatively little is known about coping with motivational problems, which motivational strategies should be used, and to what extent they are employed. Within the researchers who have tackled this issue, some have found that SGs seemed to show a promising potential from a motivational standpoint. It has been consistently shown that SGs have inherent motivational properties and different strategies, allowing them to be used for improving educational applications [7, 13–15]. Game designers, for example, employ a range of Artificial Intelligence (AI) techniques (e.g., controlling the behaviour of the nonplayer characters, providing performance feedback) to promote long-term user engagement and motivation [16].

Moreover, evaluating systems interventions is obviously related to differences in learners’ performance (successful completion of tasks) or judgement (self-report questionnaires). However, by using only performance or judgement in evaluating motivational strategies, intelligent systems risk obtaining delayed or imperfect evaluations or interrupting learning process by repeatedly using self-report questionnaires. This may offer misleading information regarding the impact of motivational strategies on learners’ motivation. Therefore, it is of particular relevance to investigate new ways of evaluating motivational strategies. One promising way is the use of physiological sensors. This is notably explained by the significant results of recent studies involving physiological sensors to assess motivational learners’ states as well as emotional and cognitive systems strategies [10, 17–19].

The present paper examines the implication of different physiological sensors to evaluate some motivational strategies employed in SGs and to highlight the corresponding learners’ patterns. To this end, we use an existing SG called Food-Force presented by the United Nations World Food Programme (WFP) and intended to learn players about the fight against world hunger. The ultimate objective of this intervention study is to assess learners’ motivation when motivational strategies have been used by SGs. We ask the two following research questions: Can we empirically find physiological patterns to evaluate the effects of motivational strategies on learners’ motivation during interactions with Food-Force? If so, Can these patterns feed AI models to predict the level of learners’ motivation to the motivational strategies? Hence, two hypotheses are postulated: (1) it is possible to model learners’ physiological reactions and trends towards motivational strategies in an SG environment and (2) we can discriminate between effective and ineffective motivational strategies using physiological manifestations as well as self-report questionnaires. We designed an experiment using an existing SG called Food-Force and combined both the theoretical ARCS model of motivation and empirical physiological sensors (heart rate (HR), skin conductance (SC), and electroencephalogram (EEG)) to assess the effects of motivational strategies on learners.

The organization of this paper is as follows. In the next section, we present previous work related to our research. In the third section, we explain our empirical approach in assessing learners’ motivation. In the fourth section, we describe the theoretical ARCS strategies to support motivation and the studied strategies in Food-Force. In the fifth section, we detail our experimental methodology. In the sixth section, we present the obtained results and discuss them. Finally, we give a conclusion in the last section, as well as present future work.

2. Related Research

Csikszentmihalyi [20] observed that people enter in a “flow” state when they are fully absorbed in activity during which they lose their sense of time and have feelings of great satisfaction. Games generally catalyze conditions of flow state by their clear goals, balance between challenges and skills, immediate feedback, progress, and control. Furthermore, Ryan and Deci [21] defined the Self-Determination Theory (SDT) and distinguished between intrinsic motivation (to understand the subject) and extrinsic motivation (for the reward of a certificate or employment). They assumed that the individual is normally inclined to be active, motivated, curious, and eager to succeed. They also recognized that some people mechanically perform their tasks, or even people passive and unmotivated. They reported that environments that facilitate the satisfaction of psychological needs can boost the internal dynamism of people to maximize their motivation and to maximize results in personal development and behavior. Ryan and colleagues [22] have studied the SDT and stated that the motivational pull of computer games is attributed to the combination of optimal challenge and informational feedback. Bartle [23], one of the pioneers of the massively multiplayer online games and known for his work on the first MUD (An MUD (originally Multiuser dungeon, with later variants Multiuser Dimension and Multiuser Domain) is a computer program, usually running over the Internet, that allows multiple users to participate in virtual-reality role-playing games.), distinguished several motivational profiles among players: the killer (competitive-ness), the per-former (success), the explorer (curiosity), and socializer (cooperation). He reported that some players strive to achieve all the challenges offered by the gameplay while others seek the company of other players, or want to discover the whole virtual world.

Nevertheless, the effectiveness of any study regarding the assessment of learners’ motivational states depends on two important factors: (1) the choice of proper assessment tools and (2) the accuracy of the selected tools. For example, Schunk et al. [5] used Keller’s ARCS model (see next section for a description of this model) and proposed several rules to infer motivational states from two sources: the interactions of the students with the tutoring system and their motivational traits. Some researchers have analyzed log files and have established correlations between learners’ actions in log files
and their motivational states (e.g., [24]). Other researchers have used physiological sensors to assess learners’ motivation and correlate physiological learners’ responses to some dimensions of motivation such as attention and confidence (e.g., [17, 25]). They have identified that the combination of various physiological sensors may provide perfect measures of learners’ states and consequently enhance systems intervention strategies. They have involved a variety of sensors to assess physiological learners’ states and responses to stimuli in Computer-Based Education (CBE) environment: mouse, electromyogram (EMG), respiration (RESP), HR, SC, and more recently EEG. For instance, Conati [19] has used biometric sensors (HR, SC, EMG, and RESP) and facial expression analysis to develop a probabilistic model of detecting students’ affective states within an educational game. Arroyo and colleagues [26] have used four different sensors (camera, mouse, chair, and wrist) in a multimedia adaptive tutoring system to recognize students’ affective states and embed emotional support.

Others studies have shown that learners have also been known to experience a lower sense of relatedness to the educational systems [27], thus increasing their feeling of isolation and possibly leading to further motivational issues. For example, learners lack the substantial self-monitoring skills that CBE systems require and possibly start “gaming” the system [9]. In addition, CBE systems generally place fewer restrictions on learners and learners must take greater responsibility for their educational experiences. Hannafin and colleagues [28] recommended that students need more support and must be empowered to acquire the necessary skills to effectively progress in an educational environment. For example, it has been found that when collaborative learning strategy is used, fewer errors are made than in individual learning situations, resulting in better outcomes performance, increased confidence, and decreased frustration levels of the learners [29, 30]. Dörnyei [31] has reported that motivational strategies are used not only to maintain students’ motivation but also to generate and increase it. He has defined that more than one hundred motivational strategies can be used not only to maintain students’ motivation but also to generate and increase it. He has identified that more than one hundred motivational strategies can be used by teachers in the classroom. These strategies integrate the creation of the basic motivational conditions, the generation of initial motivation, the maintaining and the protection of motivation, and the encouragement of positive and retrospective self-evaluation. Furthermore, efforts to overcome learners’ motivational problems have mainly been focused on tutor’s strategies or instructional design aspects of the systems. For example, Hurley [12] developed interventional strategies to increase the learner’s self-efficacy and motivation in an online learning environment. She extracted and then validated rules for interventional strategy selection from expert teachers by using an approach based on Bandura’s Social Cognitive Theory and by observing the resulting learners’ behaviour and progress. Goo and colleagues [32] showed that tactile feedback, sudden view point change, unique appearance and behaviour, and sound stimuli played an important factor in increasing students’ attention in virtual reality experience. Arroyo and colleagues [8] evaluated the impact of a set of noninvasive interventions in an attempt to repair students’ disengagement while solving geometry problems in a tutoring system. They claimed that showing students’ performance after each problem reengages students, enhances their learning, and improves their attitude towards learning as well as towards the tutoring software.

3. Assessment of Learners’ Motivation

There are a few studies that have particularly considered the evaluation of motivational strategies. It is of particular significance for this research work that motivational strategies are identified and their impacts on learners are evaluated in a specific CBE environment, precisely serious games. In addition to ARCS self-reported questionnaires, the present study uses three physical sensors (SC, HR, and EEG) to assess motivational strategies while interacting with a serious game called Food-Force. We first need to present the tools used to measure motivation itself.

3.1. ARCS Model of Motivation. In the present study, the ARCS model of motivation [33] has been chosen to theoretically assess learners’ motivation in SG. Keller used the existing research on psychological motivation to identify four components of motivation: Attention, Relevance, Confidence, and Satisfaction. His model has been used in training and games and has also been validated in numerous studies with all education levels and in many different cultures (e.g., [3, 34, 35]), and therefore, it is of particular interest in our study.

(i) Attention: to attract learners’ attention at the beginning and during the process of learning. Diverse activities should be considered to maintain students’ feelings of novelty thus the attention can be sustained.

(ii) Relevance: to inform learners of the importance of learning and to explain how to make the learning meaningful and beneficial.

(iii) Confidence: to allow learners to know the goal and to believe that the goal can be achieved, if enough effort (physical and/or intellectual) has been made.

(iv) Satisfaction: to provide feedback on performance and to allow learners to know how they are able to perform well and apply what is learned in real life situations.

The ARCS questionnaire asks students to rate ARCS-related statements in relation to the instructional materials they have just used. Examples of items related to each ARCS component are as follows.

(i) “uses questions to pose problems or paradoxes.” (Attention);

(ii) “uses language and terminology appropriate to learners and their context.” (Relevance);

(iii) “provides feedback on performance promptly.” (Confidence);

(iv) “makes statements giving recognition and credit to learners as appropriate.” (Satisfaction).
3.2. Physiological Sensors. Considering the motivation as a state of both cognitive and emotional arousal, we have decided to combine several noninvasive physiological sensors in order to empirically evaluate the motivational strategies in serious games context. Besides the SC and HR sensors which are typically used to study human affective states [36], we have considered relevant to use the EEG sensor in our proposed approach. Indeed, brainwave patterns have long been known to give valuable insight into the human cognitive process and mental state [37]. More precisely, our EEG analysis relies on differences between slow and fast wave ratios (i.e., “attention ratio” or Theta/low-Beta) which are correlated with responses to motivational stimuli and emotional traits [38, 39]. For instance, low-level attention is characterized by “a deviant pattern of baseline cortical activity, specifically increased slow-wave activity, primarily in the Theta band, and decreased fast-wave activity, primarily in the Beta band, often coupled” [40]. The power of the EEG “attention ratio” can be explained by Putman and colleagues [39]. According to the authors, a negative correlation exists between the attention ratio and learners’ Attention level. A high Theta/low-Beta ratio is usually correlated with excessive Theta and consequently inattentive state. Conversely, a low Theta/low-Beta ratio is normally correlated with excessive Low-Beta brainwave activity reflecting normal state in adults.

4. Motivational Strategies

The key issue in this paper is related to the identification and assessment of motivational strategies in SGs that support and enhance learners’ motivation. We define a motivational strategy as the use of a game element (or factor) [41] susceptible of providing motivational support for players. Motivational strategies in SG are the key to finding and harnessing players’ motivation to learn and achieve their goals. For example, a virtual companion in SG can offer encouragement to players as well as offering additional aid in their current task. This is can be considered as a motivational strategy related to the Confidence category of the ARCS model only if it increases learners’ belief in competence and consequently their effectiveness. Otherwise, it is simply an SG element and not a motivational strategy. Each of the four categories has also subcategories that are useful in identifying learners’ motivational profiles and in creating motivational tactics (or strategies) that are appropriate for specific situations in SG [3].

Attention Getting Strategies (Capture Interest, Stimulate Inquiry, and Maintain Attention). Before any learning can take place, the learner’s attention must be engaged. The challenge with the attention is to find the right balance of consistency, novelty, and understanding how people differ, what tactics to use, and how to adjust the tactics for the learners and how the tutor will be able to keep them focused and interested.

Relevance Producing Strategies (Relate to Goals, Match Interests, and Tie to Experiences). It is very difficult for students to be motivated to learn if they do not perceive there to be any relevance in the instruction. To stimulate the motivation to learn, it is best to build relevance by connecting instruction to the learners’ backgrounds, interests, and goals.

Confidence Building Strategies (Success Expectations, Success Opportunities, and Personal Responsibility). When people believe that they have little or no control over what happens to them, they experience anxiety, depression, and other stress-related emotions. In contrast, when they believe that they can predictably influence their environment by exercising their efforts and abilities in pursuit of their goals, then they are more motivated to be successful.

Satisfaction Generating Strategies (Intrinsic Satisfaction, Rewarding Outcomes, and Fair Treatment). One of the most important elements of satisfaction is intrinsic motivation; that is, if learners believe that they achieved a desirable level of success while studying topics that were personally meaningful, then their intrinsic satisfaction will be high. Another component of satisfaction is based on social comparisons and comparisons to expected outcomes.

The present study invited participants to play the freely downloadable SG called Food-Force [42]. It is an initiative of the World Food Program (WFP) of the United Nations intended to educate players about the problem of world hunger. Food-Force is comprised of multiple arcade-type missions, each intended at raising players’ awareness towards specific problems regarding worldwide food routing and aid. Food-Force also presents players’ objectives in a short Instructional Video before the beginning of each mission. A virtual tutor also accompanies the player throughout each mission by offering various tips and lessons relative to the obstacles and goals at hand. All participants have never played Food-Force before. We have investigated in details six motivational strategies in Food-Force in order to answer our main research question. (Can we empirically find physiological patterns to evaluate motivational strategies during serious game play?) These strategies are related to the four categories of the ARCS model, see Figure 1.

Problem Solving. Keller’s ARCS motivation theory tells us that the learner’s motivation is also aroused by the mean of “solving a problem or resolving an open issue...” called inquiry arousal. Mission 2 (nutritious meal preparation) presents to learners a challenging problem that consists of finding the right combination of different food items (rice, beans, vegetable oil, sugar, and iodized salt) to create a nutritious and balanced diet, all at a target cost of 30 US cents per person per meal. It has been investigated in our experiment to study the Problem Solving strategy used by Food-Force.

Alarm Trigger. According to Brophy [43], situational interest is triggered in response to something in the situation (e.g., unexpected sound) that catches our attention and motivates us to focus on it and explore it further. Keller’s ARCS motivation theory also argues that the learner’s motivation
is possibly gained by a perceptual arousal (novel, surprising, or incongruous events). We have decided to investigate the three Alarm Triggers as a strategy supporting motivation in mission 5 (UN. food delivery) of Food-Force. An example of an alarm trigger is shown in Figure 2.

**Instructional Video.** Motivational strategies rely on some game elements that make a lesson content relevant to the learners. Keller has reported that the instructor has to tie instruction into the learner’s experience by providing examples that relate to the learner’s work. In Food-Force, Instructional Video segments that draw on players’ existing knowledge have been used in order to show them the real application of presented mission in the field and connect each mission to the problem of world hunger.

**Informative Feedback.** It is important to raise learner’s confidence by offering suitable feedback. According to [44],
negative or positive Informative Feedback tells learners what they are doing. It works much better then Controlling Feedback which simply tells them what to do. For Informative Feedback used in Food-Force, comments like “What was a dangerous drop! Try to be more accurate and watch the wind gauge” (mission 3) or “Won’t arrive immediately, but that might be ok for you” (mission 4) indicate the effects (or benefits) of actions taken by the player.

**Explanatory Feedback.** The learner is open to a brief instructional explanation that will help build the right mental model and/or correct misconceptions. Explanatory Feedback resulted in much better learning than Corrective Feedback [45], which can be automated in many authoring tools with only a few key strokes. The virtual companion of Food-Fore makes comments, such as “Rice: we need a lot of rice. It only a few key strokes. The virtual companion of Food-Fore sets a goal for the learner and encourages them to be proud of themselves and celebrate success, and using rewards. Displaying Score strategy is used at the end of each mission in order to show players their current scores and their overall progress.

5. Experimental Methodology

5.1. Procedure. Thirty-three volunteers (11 females) took part in the study in return of a fixed compensation. Participants were recruited from the University of Montreal. The sample’s mean age was 26.7 (SD = 4.1). Following the signature of a written informed consent form, each participant was placed in front of the computer monitor to play the game. Set on a fictitious island called Sheylan riven by drought and war, Food-Force invites participants to complete 6 virtual missions that reflect real-life obstacles faced by WFP in its emergency responses both to the tsunami and other hunger crises around the world. All participants have played only the first five missions of Food-Force. A pretest and posttest were also administered to compare learners’ knowledge before and after playing the game. Four participants (2 females) were excluded from the study due to technical problems at the time of recording. Technical Fz recording problem with some participants leads us to exclude all Fz data from our analysis because of technical problems at the time of recording. Technical Fz recording problem with some participants leads us to exclude all Fz data from our analysis. Technical Fz recording problem with some participants leads us to exclude all Fz data from our analysis. Technical Fz recording problem with some participants leads us to exclude all Fz data from our analysis. Technical Fz recording problem with some participants leads us to exclude all Fz data from our analysis. Technical Fz recording problem with some participants leads us to exclude all Fz data from our analysis. Technical Fz recording problem with some participants leads us to exclude all Fz data from our analysis. Technical Fz recording problem with some participants leads us to exclude all Fz data from our analysis. Technical Fz recording problem with some participants leads us to exclude all Fz data from our analysis.
relationships in the data. We have normalised each HR and SC data using the following modified formula [49]:

\[
\text{normalised signal}(i) = \frac{\text{signal}(i) - \text{baseline}}{\text{signal}_{\text{max}} - \text{signal}_{\text{min}}},
\]

where signal_{max} and signal_{min} refer, respectively, to maximum and minimum values during interaction period and baseline refers to the average value of physiological data before the beginning of the game. These normalized physiological data reflect signal changes from baseline.

EEG data were segmented into one-second epochs and power spectral densities were calculated for each epoch using Fast Fourier Transformation. Power spectral data were averaged within Theta (4–8 Hz) and Low-Beta (12–20 Hz) bands. For each epoch of every participant the attention ratios (Theta/low-Beta) were calculated as described in Section 3.

5.4. Percent of Time (PoT) Index. We have defined an index representing players’ physiological evolution throughout each mission of the serious game with regards to each signal signification. This index, called Percent of Time (PoT), represents the amount of time, in percent, that player’s signal amplitude is lower (or higher) than a specific threshold. The PoT index is a key metric enabling us to sum up players’ entire signal evolution for a mission. A simple method would be to choose the mean players’ signal amplitude of each physiological sensor as the threshold. The PoT index of HR (or SC) for each player was calculated using values above the HR (or SC) threshold, whereas the PoT index for each EEG sites was calculated when player’s attention ratio was below the threshold since we are looking for positive evolutions. Figure 4 illustrates an EEG attention ratio evolution during 20 seconds. The PoT for the selected 5-second window was 80% (4 values below divided by 5 values) and 70% for the entire 20 seconds (14 values below divided by 20 values).

The idea is to analyze, in a joint venture, PoT indexes of HR, SC, and EEG signals to determine, or at least estimate,
relations between the motivational strategies used in the serious game and the physiological learners’ responses. To that end, various AI models have been constructed using gathered data in order to classify learners in two distinct classes: “Below” and “Above”. Indeed, subjects have been separated into two groups based on their self-reported scores of the ARCS model after each mission of game: those with scores below the overall average (group “Below”) and those with scores above the overall average (group “Above”). For instance, the evaluation of an Attention Getting strategy (e.g., Alarm Trigger or Problem Solving) used by Food-Force will consider the Attention scores to determine the “Below” and “Above” groups of subjects and compare their physiological reactions [50]. The same procedure has been applied for all other strategies. Consequently, the members of each group are different from one strategy to another. A detailed description of all these possibilities is given in the following section.

6. Experimental Results

Before presenting our results, we considered it necessary to quickly explain the statistical approach used in this section. Indeed, we could not rely on the usual parametric statistical tools such as ANOVA and t-test because (1) our sample population is small (N = 29 participants), (2) no justifiable assumptions could be made with regards to the normal distribution of the data, and (3) normality tests run on our data confirmed its nonnormal distribution. Hence, nonparametric Friedman’s ANOVA by ranks (counterpart of the parametric one-way ANOVA) and nonparametric Wilcoxon’s signed ranks test (counterpart of paired sample t-test) have been used. However, P value is interpreted in the same manner in both approaches and to that effect, reported significant P values were all computed at the 0.05 significance level (95% confidence).

6.1. Performance and Motivation. In order to determine if the IMMS scale is reliable, a Cronbach’s Alpha was run on IMMS data gathered after the first mission of Food-Force. The simplified IMMS yielded reliability (Cronbach’s Alpha coefficient) of 0.88 for the overall motivation measure and Cronbach’s Alpha for Attention, Relevance, Confidence, and Satisfaction was 0.91, 0.71, 0.79, and 0.87, respectively. These reliability coefficients are analogous to those found in [51] and showed that the motivational measurement instrument used in the present study was highly reliable.

Since we intend to study several motivational strategies in different missions within the Food-Force game, we evaluated the effects of these strategies on learners’ performance as well as their motivation. We have then conducted statistical tests and we have obtained several results regarding knowledge acquisition (pre- and posttests) and learners’ motivation (ARCS scores). The results of Wilcoxon signed ranks test displayed in Table 1 showed a significant difference between the participants’ scores of the pre- and posttests in terms of knowledge acquisition (Z = 4.65, P < 0.001). Number of correct answers after finishing the game is significantly higher than that of correct answers before start playing. The results of Friedman’s ANOVA by ranks between ARCS scores are displayed in Table 2. Significant differences for the general motivational scores as well as each category of the ARCS model were also observed between missions, except for Relevance (motivation overall score: F(1,4) = 10.16, P < 0.05; Attention: F(1,4) = 19.51, P < 0.001; Relevance: F(1,4) = 7.38, P = 0.12; Confidence: F(1,4) = 16.8, P < 0.05; Satisfaction: F(1,4) = 10.85, P < 0.05). Nonsignificant results of the Relevance category can be explained by the fact that the Relevance Producing strategy (Instructional Video) presented between missions was roughly the same: video segments explain the goal of each mission or its real application in order to connect each mission to the problem of world hunger. Conversely, the Attention category which showed the strongest difference and rank has used various game strategies throughout the missions. Indeed, Food-Force maintains learners’ attention by using Alarm Trigger when they are confronted with an unexpected situation such as attacks to the convoy by local rebel forces or flat tires of trucks (mission 5). It also includes mental tasks that require concentration and attention: drop food from the air without risking human lives (mission 3) and guide a convoy of trucks safely to a feeding centre while overcoming challenges from clearing land mines to rebuilding bridges and negotiating with local rebel forces (mission 5). Finally, learners’ attention is possibly gained by using Problem Solving strategy such as finding the right combination of different food items (rice, beans, vegetable oil, sugar, and iodized salt) to create a nutritious and balanced diet, all at a target cost of 30 US cents per person per meal (mission 2).

An example of an Alarm Trigger used in mission 5 is shown in Figure 2. As described in Section 4, Alarm Trigger is a motivational strategy (Attention Getting strategy) associated to Attention category of the ARCS model. We have then considered self-reported Attention scores to separate participants into two groups: a “Below” class (4 females and 7 males) representing participants who reported an Attention score below that of the overall mean and an “Above” class (5 females and 13 males) presenting the opposite (a score above the overall mean). Three alarms in mission 5 have been investigated. They are a sound trigger followed by Food-Force logistics officer’s comments used to help players to overcome challenges—from clearing land mines to rebuilding bridges and negotiating with local rebel forces. To detect physiological changes for each player, we considered two 5-second windows computed before and after each alarm and calculated their means (meanBefore Alarm, meanAfter Alarm). Fifteen Wilcoxon signed ranks tests (3 alarms × 5 physiological sensors) were run between Before Alarm and After Alarm data and significant results were

---

**Table 1: Results of Wilcoxon signed ranks test.**

<table>
<thead>
<tr>
<th>Test</th>
<th>Mean</th>
<th>Median</th>
<th>SD</th>
<th>Z</th>
<th>Sig. P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pretest</td>
<td>6.07</td>
<td>6</td>
<td>1.387</td>
<td>4.65</td>
<td>.000*</td>
</tr>
<tr>
<td>Posttest</td>
<td>8.86</td>
<td>9</td>
<td>.990</td>
<td>5.12</td>
<td>.000*</td>
</tr>
</tbody>
</table>

*Significance at the .05 level.
Table 2: Results of Friedman’s ANOVA by ranks.

(a) Motivation

<table>
<thead>
<tr>
<th>Mission</th>
<th>Mean</th>
<th>Median</th>
<th>SD</th>
<th>Chi-Square</th>
<th>Sig. P</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>55.14</td>
<td>55</td>
<td>10.347</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>54.66</td>
<td>55</td>
<td>11.321</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>52.00</td>
<td>50</td>
<td>11.206</td>
<td>4.657</td>
<td>.000*</td>
</tr>
<tr>
<td>4</td>
<td>58.14</td>
<td>62</td>
<td>10.535</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>56.12</td>
<td>54</td>
<td>10.377</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Significance at the .001 level.

(b) Attention

<table>
<thead>
<tr>
<th>Mission</th>
<th>Mean</th>
<th>Median</th>
<th>SD</th>
<th>Chi-Square</th>
<th>Sig. P</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>14.334</td>
<td>15</td>
<td>3.351</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>16.310</td>
<td>18</td>
<td>3.883</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>16.000</td>
<td>17</td>
<td>3.595</td>
<td>19.512</td>
<td>.001*</td>
</tr>
<tr>
<td>4</td>
<td>16.862</td>
<td>17</td>
<td>3.090</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>17.620</td>
<td>19</td>
<td>3.121</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Significance at the .01 level.

(c) Relevance

<table>
<thead>
<tr>
<th>Mission</th>
<th>Mean</th>
<th>Median</th>
<th>SD</th>
<th>Chi-Square</th>
<th>Sig. P</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12.689</td>
<td>14</td>
<td>5.745</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>11.000</td>
<td>9</td>
<td>5.855</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>9.482</td>
<td>7</td>
<td>5.369</td>
<td>7.379</td>
<td>.117</td>
</tr>
<tr>
<td>4</td>
<td>12.620</td>
<td>12</td>
<td>5.747</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>10.758</td>
<td>10</td>
<td>4.852</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Significance at the .05 level.

(d) Confidence

<table>
<thead>
<tr>
<th>Mission</th>
<th>Mean</th>
<th>Median</th>
<th>SD</th>
<th>Chi-Square</th>
<th>Sig. P</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>14.689</td>
<td>16</td>
<td>4.629</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>12.655</td>
<td>14</td>
<td>4.760</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>11.241</td>
<td>12</td>
<td>4.725</td>
<td>16.833</td>
<td>.002*</td>
</tr>
<tr>
<td>4</td>
<td>14.586</td>
<td>16</td>
<td>4.452</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>12.344</td>
<td>14</td>
<td>4.466</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Significance at the .01 level.

(e) Satisfaction

<table>
<thead>
<tr>
<th>Mission</th>
<th>Mean</th>
<th>Median</th>
<th>SD</th>
<th>Chi-Square</th>
<th>Sig. P</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13.413</td>
<td>14</td>
<td>2.872</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>14.689</td>
<td>15</td>
<td>3.495</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>15.275</td>
<td>16</td>
<td>3.463</td>
<td>10.852</td>
<td>.028*</td>
</tr>
<tr>
<td>4</td>
<td>14.862</td>
<td>15</td>
<td>2.812</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>15.724</td>
<td>15</td>
<td>2.986</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Significance at the .05 level.

obtained for all data. The 3 Alarms Triggers and learners’ physiological trends are presented in Figure 5. Each dot on the graph represents the difference between the two means for each alarm (meanAfter Alarm − meanBefore Alarm). Figure 5 shows almost complete opposite trends for all physiological data between the “Below” and “Above” classes, except for SC. The physiological analysis pointed towards the fact that the “effect” of an Alarm Trigger seems to decrease over time. We can see on Figure 5(a) that the effect of those alarms on SC seems to slowly fade after the second alarm, contrary to the popular belief. Indeed, one may think that intervening with color and sound tends to capture learners attention, but our findings seem to indicate that this is only partially true. There seems to be a certain “adaptation” on the part of the learner with regards to SC at the very least. Nevertheless, any permanent diagnosis regarding learners’ attention level in reaction to an Alarm Trigger based only on SC at this point may be hasty or even wrong for there are numerous other
physiological trends to consider first. Indeed, even if no clear trends were found in HR, the cerebral data provided clarity in distinguishing between the two classes.

In fact, variations in the attention ratio are clearly evident for both classes. We found numerous occasions when two participants from different classes had the same SC and HR trends but have shown very opposite trends in EEG sites, especially C3 area. An example of this situation is illustrated in Figure 6: two participants had the same HR and SC trends but only an opposite trend in C3 helped us identify their respective attention classes. These results seem to show the relevance and importance of adding the EEG in assessing learners’ attention change, even more so when this change cannot be clearly established by the use of HR and SC alone. Thus, the EEG “attention ratio” generally increases for participants who reported a low Attention category score (class “Below”) whereas the same ratio decreases for the learners in the class “Above”.

6.2. Logistic Regression Analysis. Subjects have been separated into two groups according to their ARCS scores after each mission: those with scores below the overall average (group “Below”) and those with scores above the overall average (group “Above”). We have run logistic regressions
Figure 6: Comparison of physiological trends of 2 learners in 2 different classes: the same HR and SC trends (a) and opposite C3 mean difference Before/After trends (b).

Table 3: Omnibus tests of model coefficients (logistic regression).

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Chi-Square</th>
<th>df</th>
<th>Sig.</th>
<th>Nagelkerke $R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Problem Solving</td>
<td>15,893</td>
<td>5</td>
<td>.007*</td>
<td>.574</td>
</tr>
<tr>
<td>Alarm Trigger</td>
<td>18,706</td>
<td>5</td>
<td>.002*</td>
<td>.647</td>
</tr>
<tr>
<td>Instructional Video</td>
<td>7,563</td>
<td>5</td>
<td>.182</td>
<td>.312</td>
</tr>
<tr>
<td>Informative Feedback</td>
<td>15,468</td>
<td>5</td>
<td>.009*</td>
<td>.563</td>
</tr>
<tr>
<td>Explanatory Feedback</td>
<td>12,103</td>
<td>5</td>
<td>.033*</td>
<td>.464</td>
</tr>
<tr>
<td>Displaying Score</td>
<td>11,974</td>
<td>5</td>
<td>.035*</td>
<td>.460</td>
</tr>
</tbody>
</table>

* Significance at the .05 level.

to predict learners’ group (“Above” or “Below”) for each studied strategy. The dependent variable in logistic regression is usually dichotomous, that is, the “Above” group coded as “1” whereas the “Below” group coded as “0”. Furthermore, logistic regression makes no assumption about the distribution of the independent variables. These variables do not have to be normally distributed, linearly related or of equal variance within each group. Our prediction models used all computed PoT indexes as predictor variables (PoT-SC, PoT-HR, PoT-F3, PoT-C3, and PoT-Pz) and the Enter method for variable selection. Table 3 reports the results of adding five predictors (df = 5) to the regression model. Results indicated that adding predictors to the model has significantly increased our ability to distinguish between “Above” and “Below” groups for all studied motivational strategies, except for Instructional Video (see Table 3: Chi-Square and Sig. values with conventional significance level of 0.05). In addition, Nagelkerke’s $R^2$ values of Table 3 ranged from 46% to 65% and indicated a moderately high relationship between the predictors and the dependent variable. Table 5 showed the classification tables which tell us how many of the cases where the observed values of the dependent variable were 1 or 0, respectively have been correctly predicted. In each classification table, the columns are the two predicted values of the dependent, while the rows are the two observed values of the dependent. Prediction success overall was between 65.5% and 79.3% (see Table 5). The Wald criterion demonstrated that PoT-C3 especially made a significant contribution to prediction (see Table 4). Other variables were not significant predictors. Results of regression models clearly showed that physiological data, especially EEG “attention ratio”, were relevant to evaluate motivational strategies. The most significant differences between groups were shown for Attention Getting Strategies though. One reason may be the limitation of the “attention ratio” (Theta/low-Beta) which seems to be inappropriate to identify EEG patterns other than those correlated with the Attention category. Regarding the physiological analysis, it is preferable to explore alternative EEG frequency ratios based on additional brainwaves such as Alpha (8–12 Hz) and High-Beta (20–32 Hz) in order to highlight other patterns correlated with learner’s motivation. Furthermore, F3 and C3 areas showed more significant differences of PoT trends than Pz area which showed roughly similar trends between groups. This is can be explained by specific functions associated with the middle parietal (Pz) area. These functions...
incorporate appreciation of form, sensory combination and comprehension (pain, pressure, heat, cold, and touch) which are quite sparse or even absent in all missions. Learners tended to rely mostly on the frontal cortex (F3) because it is known to be strongly implicated in taking quick decisions under pressure. The central region of the brain (C3) seems to be the most solicited when a more “generalized” problem solving approach is used. Not only our results show that physiological data can provide an objective evaluation of motivational strategies for clearly distinguishing between learners’ reactions, but also the relevance and importance of adding the EEG in our empirical study. The obtained
Table 5: Classification tables (logistic regression).

<table>
<thead>
<tr>
<th>Problem Solving</th>
<th>Observed</th>
<th>Predicted</th>
<th>Percentage correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Below</td>
<td>7</td>
<td>4</td>
<td>63.6</td>
</tr>
<tr>
<td>Above</td>
<td>2</td>
<td>16</td>
<td>88.8</td>
</tr>
<tr>
<td>Overall percentage</td>
<td></td>
<td></td>
<td>79.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Alarm Trigger</th>
<th>Observed</th>
<th>Predicted</th>
<th>Percentage correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Below</td>
<td>8</td>
<td>3</td>
<td>72.7</td>
</tr>
<tr>
<td>Above</td>
<td>4</td>
<td>14</td>
<td>77.7</td>
</tr>
<tr>
<td>Overall percentage</td>
<td></td>
<td></td>
<td>75.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Instructional Video</th>
<th>Observed</th>
<th>Predicted</th>
<th>Percentage correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Below</td>
<td>10</td>
<td>6</td>
<td>62.5</td>
</tr>
<tr>
<td>Above</td>
<td>4</td>
<td>9</td>
<td>69.2</td>
</tr>
<tr>
<td>Overall percentage</td>
<td></td>
<td></td>
<td>65.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Explanatory Feedback</th>
<th>Observed</th>
<th>Predicted</th>
<th>Percentage correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Below</td>
<td>8</td>
<td>2</td>
<td>80</td>
</tr>
<tr>
<td>Above</td>
<td>4</td>
<td>15</td>
<td>78.9</td>
</tr>
<tr>
<td>Overall percentage</td>
<td></td>
<td></td>
<td>79.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Informative Feedback</th>
<th>Observed</th>
<th>Predicted</th>
<th>Percentage correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Below</td>
<td>7</td>
<td>3</td>
<td>70</td>
</tr>
<tr>
<td>Above</td>
<td>4</td>
<td>15</td>
<td>78.9</td>
</tr>
<tr>
<td>Overall percentage</td>
<td></td>
<td></td>
<td>75.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Displaying Score</th>
<th>Observed</th>
<th>Predicted</th>
<th>Percentage correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Below</td>
<td>6</td>
<td>2</td>
<td>75</td>
</tr>
<tr>
<td>Above</td>
<td>4</td>
<td>17</td>
<td>80.9</td>
</tr>
<tr>
<td>Overall percentage</td>
<td></td>
<td></td>
<td>79.3</td>
</tr>
</tbody>
</table>

results also open the door to the possibility to evaluate other motivational strategies used in different intelligent systems.

7. Conclusion and Future Work

In this paper, we have assessed the effects of some motivational strategies in Food-Force on learners' motivation using the ARCS theoretical model as well as three physiological sensors: HR, SC, and EEG. We have successfully answered our first research question by identifying physiological patterns, especially EEG Theta/low-Beta ratio, to evaluate motivational strategies. We then successfully answered our second research question by using these physiological trends to build prediction models of learners' motivation. These models were able to moderately distinguish motivating strategies from those with low impacts on learners' motivation. Our findings showed that SC and HR may reach their limits in some cases for evaluating the impacts of motivational strategies on learners. In fact, no clear trends were found in SC and HR for evaluating some studied strategies. However, C3 Theta/low-Beta ratio has showed different trends between groups for almost all studied strategies. It can give valuable evaluation of motivational strategies.

Statistical and physiological study of our data has given some insights into the assessment of learners’ motivation during playing a serious game. It has shown that physiological parameters are suitable to assess the effects of motivational strategies on learners’ motivation. The obtained results are very encouraging to an ITS because (1) it is possible to assess the effects of tutor’s interventions on learners’ motivation, (2) we can rely on this assessment as a substitute for self-reports that can disrupt a learning session, and (3) it is possible to enrich the Learner Model (which describes learners' behaviors and evaluates their knowledge) with a motivational component based on our results, thus enabling the Tutor Model (which uses the Learner Model and customizes learning environments by adapting learning strategies in order to respond intelligently to learners’ needs, objectives, and interests) to properly adapt its interventions.

However, one limitation in this work is the assumption that the ARCS categories are independent from each other. Simultaneous strategies in SG can be related to different categories of the ARCS model. One possible extension of the present work would be to consider dependencies between ARCS categories. In addition, we can extend the present work to study more than two classes of motivation. Multinomial logistic regression will be used in this case in counterpart of binary logistic regression. It is also possible to add other variables that can improve the prediction quality of our models. Indeed, some personals characteristics (age, gender, player style, hours spent playing video games, etc.) can be additional predictors for players’ motivation. Furthermore, brain activity can also be better analysed in the future and other EEG analysis methods, such as the event-related potential (ERP) technique, can be used to test whether different events in serious game evoke differential EEG responses. We plan, therefore, to address all these possibilities in a further complementary study.

Conflict of Interests

The authors declare that they have no conflict of interests in the research.

Acknowledgments

The authors would like to thank all the reviewers for their insightful and helpful comments. They acknowledge the support of the Tunisian government and the National Science and Engineering Research Council (NSERC) of Canada for this work. They also address their thanks to Pierre Chalfoun for his participation in the experiment setup and his useful comments.
References


Research Article

Psychophysiology to Assess Impact of Varying Levels of Simulation Fidelity in a Threat Environment

Thomas D. Parsons,1 Albert A. Rizzo,2 Christopher G. Courtney,2,3 and Michael E. Dawson3

1 Clinical Neuropsychology and Simulation (CNS) Lab, Department of Psychology, University of North Texas, Denton, TX 76203, USA
2 Institute for Creative Technologies, University of Southern California, Playa Vista, Los Angeles, CA 90094, USA
3 Department of Psychology, University of Southern California, Los Angeles, CA 90089, USA

Correspondence should be addressed to Thomas D. Parsons, thomas.parsons@unt.edu

Received 25 March 2012; Accepted 13 September 2012

Academic Editor: Pablo Moreno-Ger

Copyright © 2012 Thomas D. Parsons et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

There are many virtual environments found in the serious game community that simulate real world scenarios. There is a broad range of fidelity and experimental controls among these serious games. An important component to most evaluations is the extent to which level of fidelity impacts the persons immersed in the serious game. While a great deal of virtual environment and serious game research has assessed the subjective state or feeling of the participant (e.g., the participant’s sense of presence) through the use of questionnaires, the current study examines participant experience by examining psychophysiological responses of participants to their surroundings. The primary goal in this study was evaluative: will a virtual environment with arousing contents result in increased sensory arousal if it is presented in a highly immersive configuration? A secondary goal of this study was to investigate the utility of our environment to offer varying levels of stimulus threat to impact the user’s experience of the virtual environment. Increased simulation fidelity in an arousing environment resulted in faster heart rates and increased startle eyeblink amplitudes, suggesting that higher fidelity scenarios had great efficacy related to sensory arousal.

1. Psychophysiology to Assess Impact of Varying Levels of Simulation Fidelity in a Threat Environment

Virtual environments (VEs) and serious games offer the potential to stimulate and measure changes in the users’ emotion, neurocognition, and motivation processes. The value in using simulation technology to produce serious games targeting such processes has been acknowledged by an encouraging body of research. Some of the work in this area has addressed affective processes: anxiety disorders, pain distraction, and posttraumatic stress disorder [1–3]. Other work has assessed neuropsychological processes [4, 5]. Further, psychophysiology is increasingly being incorporated into research using virtual reality environments [6–8]. The use of psychophysiological measures in affective and neurocognitive studies of persons immersed in VE scenarios offers the potential to develop current physiological computing approaches [9] into affective computing [10] scenarios.

The incorporation of simulation technology into neuroergonomic and psychophysiological research is advancing at a steady rate [11]. New discoveries and techniques are demanding a more rapid and advanced paradigm. In response to the demands, a wide variety of simulations have been developed. The range and depth of these simulations cover a large domain, from simple low fidelity task environments to complex high fidelity full immersion simulators. All of these simulators rely on some type of representation of the real world. An important issue for research into simulation for social and behavioral sciences is the determination of how advanced the simulator needs to be to adequately assess and/or train a particular individual or team. While high-end simulations can train a variety of user types, the cost associated with these devices can be difficult to justify [12].

In this paper, we attempt to build on earlier work that used psychophysiology to assess the propensity of users to respond to virtually generated sensory data as if they were real [13]. We aim to assess the propensity of users to respond
to virtually generated sensory data as if they were real [14]. In the same way people experience physiological responses to stimuli in the real world, researchers seek to quantify participant experience by measuring responses evoked by stimuli in a VE. A low fidelity VE may be preferable in studies where a maximal amount of experimental control is desired because such environments may increase psychometric rigor through limiting the number of sensory variables available to the user [15]. Contrariwise, high fidelity environments are preferable for studies desiring increased ecological validity because they recreate more of the real world environment—better capture the participant’s performance as it would occur in a real world setting [16].

Discussions of the level of fidelity and experimental control needed for a VE often go beyond simple discussions of the “immersive” qualities of the environment to discussions of the impact upon the perceived feeling of “presence” of the individual while immersed in the environment [17]. A number of discussions of the distinction between the terms “immersion” and “presence” can be found in the literature [18–22]. This distinction is important for the current study because issues of fidelity tend to reflect levels of immersion, while levels of presence reflect the user’s experience relative to the level of fidelity/immersion. For the current study, we focus on fidelity and levels of immersion. We also feel that it is important to differentiate between immersion and presence. By immersion, we follow Slater and Wilburs [21] delineation, in which immersion is seen as an objective description of aspects of the system—that which the overall VE can deliver (e.g., the level of fidelity in representing the real world; the field of view, the number of sensory systems it simulates, the frame-rate, and latency [14]). Hence, the level of immersion is an objective property of a VE that in principle can be measured independently of the human experience that it engenders. We view presence as a subjective phenomenon (e.g., sensation of being in a VE).

Knowledge of the user-state during exposure to the VE is imperative for development and assessment of VE design. A number of presence studies have researched such issues using questionnaires [23–25]. Subjective measures tend to rely on post-test assessments of the user’s feelings during the exposure to the VE, which is dependent on memory of the event [26]. Self-report data, when used in isolation, are highly susceptible to influences outside the subject’s own targeted attitudes [21]. The item’s wording, context, and format are all factors that may affect self-report responses. A limitation of questionnaire measures is that they can only be administered following a participant’s immersion in a VE, but in order to assess participant experience during the actual immersion in a VE, researchers have sought a more objective measure. Online assessment of participant experience is difficult when using subjective measures, in that the very existence of subjective questions during immersion serves to break the continuity of the participant’s conscious awareness currently being experienced. As a result, a quite different view seems to be emerging, in which presence is treated as something rooted in physiological and behavioral activity [14, 22] and there is a growing emphasis upon physiological and behavioral assessment [27], as well as the relation between immersion and emotion [28, 29]. Further, there has been increased use of neuroscience techniques for presence measurement, such as EEG [30], transcranial Doppler [31], and fMRI [32].

Up until this point, VE studies have typically relied on self-report and behavioral measures to assess levels of fear and arousal [33, 34]. Some studies however, have moved toward using more objective psychophysiological measures. Jang et al. [35] measured psychophysiological responses including skin resistance and heart rate variability to assess arousal levels in normal subjects exposed to fear of driving and fear of flying VEs. Subjects showed lowered levels of skin resistance compared to baseline, indicating higher levels of arousal, especially during the first 7 to 8 minutes of exposure to the VE. Other studies have also found VEs to be physiologically arousing [36–38].

Meehan et al. [39] sought to uncover a more objective, valid, and reliable measure of presence through psychophysiological metrics. Because psychophysiological responses can be made without conscious knowledge of the response, the experience of the virtual environment, and feelings of presence, need not be interrupted. Meehan and colleagues found that heart rate and skin conductance increased along with increased feelings of presence. They conclude that psychophysiological measures may be utilized as an objective and reliable measure of presence, though they note that additional research using different environments and stressors is necessary to further elucidate these findings.

Psychophysiological metrics proffer the advantage of an objective measure of response that can be recorded in real-time as the environment is experienced, providing a continuous measure of presence. Indeed, highly immersive presentations are thought not only to produce subjective ratings, but also result in increased physiological responses [18]. As such, researchers may study the impact of VEs on participants by looking at the psychophysiological responses of participants to their surroundings [6, 39, 40]. The recording of psychophysiological variables while participants operate within VEs has produced useful results in studies examining presence and immersion [41–44]. As such, the VE assets that allow for precise stimulus delivery within ecologically enhanced scenarios appears well matched for this research.

1.1. Current Study. In the current study, we aimed to look at the psychophysiological responses of participants experiencing “high” versus “low” levels of immersion into a virtual Iraqi scenario that had varying levels of stimulus intensity. Further, these psychophysiological responses may aid researchers in their development of VEs that balance fidelity and experimental control.

The primary goal in this study was evaluative: will a virtual environment with arousing contents result in increased sensory arousal if it is presented in a highly immersive configuration? To assess this, we looked at subjects immersed in a VE on two separate experimental runs consisting of both a “high” immersion condition and a “low” immersion condition. A secondary goal of this study was to investigate the utility of our environment to offer varying levels of stimulus threat to impact the user’s experience of the VE.
Within each of the immersion conditions (high and low), arousal was manipulated by presenting participants with differing “safe” and “ambush” zones. Safe (low threat) zones consisted of little activity aside from driving down a desert road, while the more stressful ambush (high threat) zones included gunfire, explosions, and shouting amongst other stressors.

In the current study, startle eyeblink and heart rate were measured to assess psychophysiological differences in response to varying degrees of immersion and levels of arousal in a virtual Iraqi environment. Participants encountered a highly immersive VE while wearing a head mounted display (HMD) that used two OLED microdisplays with onboard 3D frame sequential video processors to deliver flicker-free motion. Together with the integrated X, Y, and Z-axis head-tracker, the participant was able to look around the VE by turning his or her head left or right, up or down, and leaning forward or back. Also in the high immersion condition, to increase the potential for sensory immersion, the participant was seated on a tactile transducer. Although participants experienced the same VE content in the low immersion condition, the VE was presented on a computer screen. It was expected that the highly immersive condition would lead to an increased sensory arousal, thus resulting in augmented levels of psychophysiological responding.

1.2. Hypotheses. We hypothesized that the highly immersive condition would elicit a more intense physiological response to the stressful high threat zones due to the higher-fidelity environment. It is important to note that the volume levels in both the high and low immersion conditions were held constant in order to increase experimental control of the arousal manipulation in this study and to limit changes in arousal to changes brought on by increased levels of fidelity, rather than changes in volume level. It is our hope that this research will (1) proffer a greater understanding of the psychophysiological correlates of immersion in an arousing VE and (2) act as an initial validation (ecological validation) of the intended impact of varying degrees of stimulus intensity programmed into our virtual Iraqi city.

1.3. Serious Games and Clinical Psychology. Serious game researchers are increasingly interested in working with clinicians to better understand a military service member’s ability to return to active duty. Recent conflicts have increased the prevalence of blast injuries to the head. Many of these brain injuries may have no external marker of injury. As a result, there is need for the serious games community to research innovative assessment methods. Currently, clinicians make “Return-to-Duty” assessments that are based upon the “Return-to-Play” guidelines found in Sports Medicine. Both have incorporated two dimensional cognitive assessments to aid in decisions related to resuming activities following a concussion. Unfortunately, these two dimensional computerized assessments were not developed with the intention of tapping into everyday behaviors like driving through a Middle Eastern city.

Serious gaming environments can increase the ecological validity of neurocognitive batteries through the use of simulation technologies for assessment and treatment planning. The success of such serious games may lead to a psychophysiological computing approach, in which such data gleaned from persons interacting within a military relevant simulation may be used to develop adaptive virtual environments for training and rehabilitation. A beginning step is the identification of the level of immersion needed for a serious game to proffer the appropriate level of arousal. This is the overarching goal of this study.

2. Methods

2.1. Participants. A total of 50 healthy college aged students (males: N = 23, mean age = 20.4, mean years of education = 14.6; females: N = 27, mean age = 19.8, mean years of education = 14.1) participated in this experiment. An interview with a psychologist and a mental health history form were completed with each participant in order to allow the following of strict exclusion criteria to minimize the possible confounding effects of additional factors known to adversely impact a person’s ability to process information, including psychiatric (e.g., mental retardation, psychotic disorders, diagnosed learning disabilities, attention-deficit/hyperactivity disorder, and bipolar disorders, as well as substance-related disorders within two years of evaluation) and neurologic (e.g., seizure disorders, closed head injuries with loss of consciousness greater than 15 minutes, and neoplastic diseases) conditions.

2.2. Apparatus

2.2.1. Hardware. The apparatus used for the virtual humvee (i.e., a high mobility multipurpose wheeled vehicle) included a Pentium 4 desktop computer with a 3 GHz Processor; 3 GB of RAM; and an nVidia GeForce 6800. Two monitors were used: (1) one for displaying the Launcher application which is used by the Examiner and (2) another for displaying the participant’s view of the VE in the HMD. Participants wore an eMagin Z800 head mounted display, and an InterSense InteriaCube 2+ attached for enhanced tracking. A Logitech Driving Force steering wheel was clamped on to the edge of a table in front of the monitors. A separate module consisting of the gas and brake pedals was positioned under the table. To increase the potential for sensory immersion, we built a tactile transducer using a three foot square platform with six Aura bass shaker speakers (AST-2B-04, 4Ω 50 W Bass Shaker) attached. The tactile transducer was powered by a Sherwood RX-4105 amplifier with 100 Watts per Channel ×2 in Stereo Mode.

2.2.2. Virtual Environment. The software was designed using Virtual Battle Space 2 (VBS2). The VBS2 engine was used due to its robust fidelity simulation, ease of modification, and the fact that many military forces have adopted it. The VBS2 engine offers enhanced capability for interoperability and compatibility with existing standards for simulation. We designed the scenarios using a visual scenario editor and VBS2’s own scripting language. To implement the scenario
we used VBS2-engine specific script language and the built-in Finite State Machine (FSM) functionality.

2.2.3. Neuroscience and Simulation Interface. The application uses the Neuroscience and Simulation Interface (NSI) developed in the Neuroscience and Simulation Laboratory (NeuroSim) at the University of Southern California [45]. The NSI was used for data acquisition, stimulus presentation, psychophysiological monitoring, and communication between the psychophysiological system and the VE. Parameters were saved to files using the NSI and automatically loaded through its control module. The NSI allowed our system to switch between parameter files, or executables modules, in order to perform specific experimental sequences. The NSI enabled the sending of event markers from the stimulus presentation computer to a recording device. Matlab scripts were executed in real-time from within NSI and filters were compiled to execute as stand-alone programs. The software runs on Windows XP 32-bit, and requires 5 Gb of free hard drive space for installation and storage of user data.

2.3. Stimuli and Design. The University of Southern California’s Institutional Review Board approved the study. After informed consent was obtained, basic demographic information was obtained. Next, participants were immersed in a VE on two separate experimental runs consisting of both a “high” immersion condition and a “low” immersion condition. In the high immersion condition, participants wore a head mounted display (HMD) with full tracking capabilities and were free to explore their environment visually. The high immersion condition also made use of headphones and a tactile transducer floor to simulate the experience of a large vehicle. The low immersion condition consisted of the same virtual Iraqi scenario presented on a 17 inch laptop screen while wearing headphones. Stimuli within the virtual environment experienced in both immersion conditions were identical. The only differences between conditions were due to the inclusion of the enhanced presentation quality of the high immersion condition. The presentation order of high and low immersion conditions was counterbalanced across subjects.

The VE used in both immersion conditions was comprised of a series of low threat and high threat zones in a virtual Iraqi city. In both the high immersion and low immersion conditions, participants experienced the VE from the perspective of the driver of a Humvee. The speed of the vehicle was kept constant as it followed a predefined trajectory to control for time spent in each zone of the VE and to keep that time consistent across participants. Participants were given a basic 10◦ steering wheel to limit the trajectory, though they were instructed to stay on the road. This allowed for some level of control of the environment without sacrificing experimental control of the stimuli experienced. Low threat zones consisted mainly of a road surrounded by a desert landscape and were free of gunfire and other loud noises (see Figure 1). The high threat zones included improvised explosive devices (IEDs), gunfire, insurgents, and screaming voices (see Figure 2). The auditory background levels associated with the low threat and high threat zones were identical in both the high and low immersion conditions. Participants passed through three low threat and three high threat zones in an alternating sequence in both immersion conditions. Low threat zones were always experienced first and were used to allow the participant to habituate to the novelty of the virtual environment. High and low threat zones also varied in length, with low threat zones consistently lasting longer than high threat zones. High threat zones averaged 20 seconds in duration, while low threat zones averaged 50 seconds. This was to ensure that zone lengths were not predictable, and so that participants had ample time to return to low levels of responding after experiencing the highly arousing high threat zones. The total length of each run was 210 seconds.

An acoustic startle stimulus was used to elicit startle eyeblink responses. Following accepted guidelines for human startle eyeblink electromyographic studies [46], the startle stimulus was a 110 dB white noise burst 50 ms in duration with a near instantaneous rise/fall time presented binaurally through Telephonics TDH-50P headphones. Decibel levels were measured with a Realistic sound level meter using a Quest Electronics earphone coupler. Startle stimuli were experienced intermittently throughout the experimental runs. A total of four startle stimuli were experienced in both the low threat and high threat zones in each run.

2.4. Dependent Variables. Psychophysiological assessment included: startle eyeblink amplitude and heart rate, which were recorded simultaneously throughout the experiment using Contact Precision Instruments equipment and a computer running SAM1 software.
2.4.1. Startle Eyeblink Response. One psychophysiological measure employed in the current study, and that is widely used as an index of valence (e.g., emotional positive or negative reactions), is electromyographic (EMG) recording of the startle eyeblink reflex. This reflex is often elicited by a burst of loud white noise with a nearly immediate rise and fall time presented at very high decibel levels (e.g., 110 dB) for a brief duration (e.g., 50 ms). Vrana et al. [47] found that startle responses are facilitated when startle stimuli are presented in conjunction with a negative stimulus, and inhibited when presented with a positive stimulus relative to startle presentations with neutral stimuli. It is important to note that the positive and negative stimuli used in the Vrana et al. study were matched on subjective ratings of arousal, meaning the startle reflex can be a sensitive measure of valence. Relevant to the measurement of eyeblink responses during exposure to VEs, these findings have been replicated with moving film clips [48, 49].

Startle eyeblink responses were recorded as electromyographic activity using two small (4 mm in diameter) silver-silver chloride electrodes placed over the orbicularis oculi muscle of the left eye and an 8 mm silver-silver chloride electrode placed behind the left ear to serve as a ground. One 4 mm electrode was placed directly below the pupil in forward gaze while the other was placed about 1 cm lateral to the first. The electrodes were placed as close to the eye as possible while still allowing the participant to open and close his or her eyes comfortably. Impedance between the two electrodes was measured and deemed acceptable if below 10 kΩ.

2.4.2. Heart Rate. A second psychophysiological measure employed in the current study was the electrocardiographic (ECG) recording. Heart rate is a psychophysiological measure that is useful in differentiating between orienting and defensive responses. A person’s heart rate will accelerate during a defensive response and decelerate when orienting was e

2.5. Data Analytics

2.5.1. Startle Eyeblink Response. The raw EMG signal was recorded at a rate of 1000 Hz throughout the experiment using a 10 Hz high pass and 200 Hz low pass filter. Raw signals were stored and exported for analysis in microvolt (µV) values. The raw EMG signal was rectified and integrated for analysis. In order to qualify for scoring, the eyeblink response had to begin within a window of 20 to 100 ms following the offset of the startle stimulus and reach peak activity within a window of 20 to 150 ms following the startle stimulus [46]. Blinks occurring at longer latencies were not considered to be the result of the startle stimulus. Amplitudes were recorded as the difference between the peak activity value and the baseline level present immediately preceding onset of the blink response. If the participant was blinking during the onset of the startle stimulus, that blink response was removed from further analysis due to artifact. Participants who failed to reach 1 µV amplitudes on greater than 50% of startled trials were considered nonresponders and were dropped from further EMG analyses. One participant reached this criterion, leaving 49 participants to be included in EMG analyses.

Due to the high levels of variability between participants in EMG responses, all blink amplitude values were standardized by taking the difference between each participant’s raw EMG amplitude value on each trial and that participant’s mean value across all trials and dividing by the standard deviation of all values. Scores were then subjected to a linear transformation resulting in a mean of 50 and a standard deviation of 10 for display purposes. This helped to ensure that all participants contributed to group means equally, minimizing the influence that one participant could have on the outcome of the subsequent analyses.

2.5.2. Heart Rate. Interbeat intervals (IBIs) were scored as the time difference in milliseconds between successive R waves in the ECG signal. IBIs across a period of 5 seconds during each high threat and low threat zone were analyzed. The 5 second period occurred at least 10 seconds following any startle stimulus or large explosion, and no startle stimuli or explosions occurred during the period. A mean IBI score was recorded for each 5 second period and analyzed.

For each dependent variable, a 2 (immersion level) by 2 (zone type) repeated measures analysis of variance (ANOVA) was utilized to determine whether the high immersion setting was effective in increasing psychophysiological responding in general and whether it affected participants differently in low threat versus high threat zones. All significant main effects and interactions were followed with paired samples t-tests in order to identify the precise nature of these effects. All reported significant t-test results are corrected using a sequentially rejective test procedure based on a modified Bonferroni inequality to prevent inflation of type 1 error rates [52].

3. Results

3.1. Startle Eyeblink Results. A significant immersion level main effect was uncovered, and was the result of increased blink amplitudes when participants were in the high immersion setting, \( F(1,48) = 16.34, \) MSEimmersion = 0.49, MSEzone type = 0.38, MSEimmersion × zone type = 0.33, \( p < 0.001. \) Zone type did not yield a significant main effect. The interaction between zone type and immersion level also...
Table 1: Distribution statistics for EMG eyeblink results.

<table>
<thead>
<tr>
<th>Immersion level</th>
<th>Median safe zone</th>
<th>Median ambush zone</th>
<th>25% quartile</th>
<th>75% quartile</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>50.6</td>
<td>52.3</td>
<td>40.7</td>
<td>58.5</td>
<td>36.3</td>
<td>65.4</td>
</tr>
<tr>
<td>Low</td>
<td>47.8</td>
<td>47.1</td>
<td>38.2</td>
<td>55.7</td>
<td>32.4</td>
<td>62.4</td>
</tr>
</tbody>
</table>

Quartile and range data are given for the entire sample, while separate median values are given for both the safe and ambush zones.

Table 2: Distribution statistics for heart rate results.

<table>
<thead>
<tr>
<th>Immersion level</th>
<th>Median safe zone</th>
<th>Median ambush zone</th>
<th>25% quartile</th>
<th>75% quartile</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>67.5</td>
<td>68.2</td>
<td>60.2</td>
<td>83.5</td>
<td>52.3</td>
<td>95.4</td>
</tr>
<tr>
<td>Low</td>
<td>66.3</td>
<td>66.4</td>
<td>59.2</td>
<td>80.8</td>
<td>49.9</td>
<td>90.2</td>
</tr>
</tbody>
</table>

Quartile and range data are given for the entire sample, while separate median values are given for both the safe and ambush zones.

Figure 3: EMG eyeblink response amplitudes for high and low immersion conditions. All amplitudes are reported as T-scores.

Figure 4: Heart rate responses (in beats per minute) for high and low immersion conditions.

failed to reach significance. However, a post hoc paired samples t-test revealed that the difference between the high and low immersion was only significant while participants experienced the high threat zones, \( t(48) = 3.82, P < 0.001 \). As can be seen in Figure 3 and Table 1, eyeblink amplitudes increased in the high threat zone only in the high immersion setting. Responses in the high threat zone of the high immersion setting were large enough to account for most of the immersion level main effect as differences between high and low immersion in the low threat zones were not significant after Rom correction.

3.2. Heart Rate Results. In general, ECG results were in agreement with EMG results. Again, a significant main effect of immersion level was found, \( F(1, 49) = 10.78, \text{MSE}_{\text{immersion}} = 0.06, \text{MSE}_{\text{zone type}} = 0.03, \text{MSE}_{\text{immersion * zone type}} = 0.03, P < 0.01 \). This immersion level effect was the result of faster heart rates when participants were in the high immersion setting. As can be seen in Figure 4 and Table 2, participants were again evidencing increased responding during the high threat zones only when in the high immersion setting, although this increase in heart rate was not significant. No significant zone type main effect or interaction between immersion level and zone type existed. There was again a significant difference between the high and low immersion presentations only during the high threat zones, \( t(49) = 3.42, P < 0.001 \), as was the case in regards to the eyeblink results.

4. Discussion

4.1. Primary Analysis: Effects of Immersion Level. For our primary analysis in this study we sought to evaluate whether a highly immersive environment results in increased sensory arousal as measured by psychophysiological responses. Immersion effects were consistent with each measure. Participants consistently had faster heart rates when in the high immersion setting, suggesting that highly immersive VEs are more arousing than experiencing the same presentation on a computer screen. Participants also had larger startle eyeblinks when highly immersed, especially during the high threat zones, which suggests that the high immersion format facilitated startle eyeblinks.
Although on first reading these results appear to reflect the possibility that highly immersive VEs are more effective for eliciting increased arousal and producing fear responses than are low immersion VEs, this conclusion cannot be generalized given that there are restorative virtual environments that decrease arousal [53, 54]. The fact is that this VE was a warzone simulation with varying levels of threat stimuli. Both of the dependent measures were shown to mainly vary with immersion, not threat.

Another area that may put our results at odds with those reported by others is the issue that our study was for neuroscientific assessment of varying levels of fidelity and threat in a nonclinical sample of healthy college age students. Clinical populations tend to have significantly greater responding to threat stimuli presented in VEs when compared to nonclinical populations. For example, virtual stimuli that are relevant to a given phobia (e.g., phobics respond with more anxiety to phobogenic stimuli) will have more robust reactions to threatening stimuli. Further, it also seems intuitively clear that participants in the current study would react less to the threatening zones than would persons sensitive to the content of the virtual Iraq (e.g., soldiers returning from a rotation in Iraq, suffering from PTSD, or having been in a war zone) [55].

### 4.2. Secondary Analysis: Effects of Zone Type

A secondary goal of this study was to investigate the utility of our environment to offer varying levels of stimulus threat to impact the user’s experience of the VE. Our analysis revealed that high threat zones were ineffective in creating statistically significant increases in arousal levels compared to the low threat zones, according to eyelink and heart rate responses. However, participants appeared to show the appropriate directional trend toward increased heart rate and eyelink responding in the high immersion setting, lending credence to the notion that the high immersion setting may be more effective in creating differential responding between the two zone types. However, these trends in response did not lead to significant interactions between immersion level and zone type.

### 4.3. Enhancing the Virtual Environment

The lack of differential responding in the high threat and low threat zones may have been due to the fixed order of presentation. While the presentation of the low and high immersion settings was counterbalanced across participants, the order of the zones was not. This meant that in each pair of low threat and high threat zones, the low threat zone was experienced first. While it is impossible to know what the exact effects of a counterbalanced presentation order would have on psychophysiological response, one possible explanation for the lack of differential responding may have been caused by habituation that led to a general decrease in responding during the high threat zones in comparison to the low threat zones that always preceded them. Had the high threat zones occurred prior to the low threat zones, a greater difference between the different types of zones may have been revealed, especially in the high immersion setting.

Additionally, the low threat zones were generally longer in duration than the high threat zones. This may have led to greater habituation taking place during the low threat zones, and created an additional confound that is difficult to account for in participant responses. Moreover, the low threat zones would transition into the high threat zones unpredictably and without warning, making the low threat zones potentially threatening.

The presentation of startle stimuli may also have added to the lack of differential responding in the low threat and high threat zones. In order to make the startle stimuli stand out from the background noise in the environment enough to elicit a startle response, the maximum capacity of the environmental noises were reduced to ten percent of the startle stimulus volume, greatly lowering the potentially arousing effects of gunshots and explosions experienced in the high threat zones.

It is important to note that there is parallel research on the restorative effects of nature that has explored the relationship between presence/immersion, psychophysiological measurements, and virtual reality. Previous research examining whether immersion in a VE simulated nature setting could produce restorative effects found that immersion in virtual nature settings has similar beneficial effects as exposure to surrogate nature. These results also suggest that VR can be used as a tool to study and understand restorative effects [53, 54].

### 4.4. Future Directions

Future studies using this VE may be enhanced through counterbalancing of the order of zones experienced in the VE. Counterbalancing across participants to allow for half to experience low threat zones first and half to experience high threat zones first should help to alleviate the possible order effects that occurred in the present study. In order to better understand which particular zone is the most effective in increasing arousal, it is important that the high and low immersion conditions can begin with any zone. We can then counterbalance whether a low threat zone or a high threat zone is experienced first, and which particular low threat or high threat zone is experienced first. A uniform amount of time spent in each zone will also help to control the effects of habituation from zone to zone. Furthermore, in order to make the low threat zones more clearly perceived as being safe, a cue could be given to warn the user of the impending high threat zone. This way, the low threat zones are clearly separated from the high threat zones.

The removal of startle stimuli to allow background environmental noises to be played at one hundred percent capacity may also be beneficial in creating more arousing high threat zones. Eyelink responses will no longer be an option as a psychophysiological measure of valence in a noisy background environment, but facial corrugator EMG recording can be used as an index of perceived valence in its stead. Other metrics such as electrodermal activity, respiration, and blood pressure may also be useful measures of arousal, and responses would most likely be enhanced by the increased volume levels.

A further enhancement for future studies would be the addition of subjective evaluations. Having both subjective
and objective information would strengthen the validity of the results and allow combining them for the conclusions [56]. The inclusion of the subjective data may have aided this study through greater explanatory power for the nonsignificant but apparent difference between low/high threat in the eyeblink responses found in low immersion.

4.5. Conclusions. One of the main goals of the present research was to assess whether a VE with arousing contents would result in increased sensory arousal if it is presented in a highly immersive configuration. A secondary goal of this study was to investigate the utility of our environment to offer varying levels of stimulus threat to impact the user’s experience of the VE. Increased simulation fidelity in an arousing VE resulted in faster heart rates and increased startle eyeblink amplitudes, suggesting that higher fidelity scenarios with threatening contents were related to sensory arousal. Hence, highly immersive VEs appear to be more effective for eliciting increased arousal and producing fear responses than are low immersion VEs.

Conflict of Interests

No financial Conflict of interests exist for any of the authors of this paper.

Acknowledgment

This research is partially supported by the US Army Research Laboratory, Human Research & Engineering Directorate, Translational Neuroscience Branch (Aberdeen Proving Ground, MD).

References


