The current issue and full text archive of this journal is available at
www.emeraldinsight.com/1741-5659.htm

Received 5 February 2013
Revised 15 March 2013
Accepted 15 March 2013

Serious game evaluation as a meta-game

Lynne Hall and Susan Jane Jones
University of Sunderland, Sunderland, UK
Ruth Aylett
Heriot-Watt University, Edinburgh, UK
Marc Hall
Orchidsoft, Gateshead, UK
Sarah Tazzyman
University of Sunderland, Sunderland, UK
Ana Paiva
INESC-ID, Lisbon, Portugal, and
Lynne Humphries
University of Sunderland, Sunderland, UK

Abstract

Purpose – This paper aims to briefly outline the seamless evaluation approach and its application during an evaluation of ORIENT, a serious game aimed at young adults.

Design/methodology/approach – In this paper, the authors detail a unobtrusive, embedded evaluation approach that occurs within the game context, adding value and entertainment to the player experience whilst accumulating useful data for the development team.

Findings – The key result from this study was that during the "seamless evaluation" approach, users were unaware that they had been participating in an evaluation, with instruments enhancing rather than detracting from the in-role game experience.

Practical implications – This approach, seamless evaluation, was devised in response to player expectations, perspectives and requirements, recognising that in the evaluation of games the whole process of interaction including its evaluation must be enjoyable and fun for the user.

Originality/value – Through using seamless evaluation, the authors created an evaluation completely embedded within the "magic circle" of an in-game experience that added value to the user experience whilst also yielding relevant results for the development team.

Keywords Computer software, Learning methods, Serious games, Evaluation

Paper type Research paper

Introduction

The term "magic circle" was first coined by Huizinga (1950) and describes the boundary of game play, within which the player is immersed and engaged in a self-created and maintained experience, a "temporary world within the ordinary world" creating a novel and engaging reality (Salen and Zimmerman, 2003). In today's technological context, computer games reinforce and facilitate this magic circle, providing sets, scenarios, characters and narrative elements, with certain genres, particularly those involving role and character play, being characterised by the powerful immersive experiences they create for gamers.

Interactive Technology and Smart Education
Vol. 10 No. 2, 2013
pp. 130-146
© Emerald Group Publishing Limited
1741-5659
DOI 10.1108/ITSE-02-2013-0003

Where games are explicitly evaluated, this magic circle is typically ruptured with evaluation conducted as a discrete, separate activity with users transformed from players into subjects (Baranowski *et al.*, 2008), critics (Klesen, 2005) or designers (Stromberg *et al.*, 2002), reflecting upon their game experience outside the game itself. The context is evenmore complex for games-based learning applications and serious games which also need to demonstrate learning, with those essential games metrics of fun and enjoyability subsumed by the need to collect summative evaluation data. Most evaluations focus upon the requirements of the development team, with approaches and instruments geared towards result provision rather than onenhancing, or at the very least maintaining, the quality of the user experience (Sheng *et al.*, 2007). This must call into question the value of the evaluation process itself, while it is situated outside the immersive experience of the game.

In order to reframe evaluation so that it meets the needs of designers and developers, while maintaining the integrity of the game experience for users, evaluation needs to take place within the magic circle itself, so that it is just another, element of the player experience. This does not simply relate to physical and interaction space, but additionally includes evaluating within the player's concept of the experience. This is based not only on consideration of the player role, but additionally the expectations and perceptions of player activities that are inherent in the role itself. This implicitly requires that the player is unaware of the evaluative nature of the activities in which they are involved. Rather, their completion of evaluation tasks, instruments and discussions for example, must appear to be just one more aspect of the player experience. Non-functional evaluation of games for entertainment must seek to establish the level of engagement with the game or the satisfaction of the user. This is true of serious games too, but serious games extend artefact requirements to include a need for pedagogical value as well as entertainment value (De Freitas, 2010). It is not enough that the player has fun but also that they learn something, that there is "transfer" or some other form of measurable pedagogic outcome, placing an even greater burden on the evaluation to detect this impact.

In this article, we present seamless evaluation, an approach to the player evaluation experience that maintains the integrity of the magic circle. It is based upon the intuitively appealing idea of incorporating evaluation as part of an in-role activity, something that has been applied in psycho-drama, with the player continuing to operate in-role throughout the evaluation activities. This paper briefly outlines this seamless evaluation approach and then discusses its application within the context of a serious game. An 18 participant study is briefly detailed, highlighting the added value that evaluation can provide to players.

Evaluation of serious games

The idea that learning can be an immersive process is not new, and the importance of context, application and practice in learning is well established and can be traced back to the earliest days of civilisation (Johnson and Levine, 2008). However, the technologies which support immersive role-playing scenarios and case studies and the simulations that underpin game playing today become increasingly more abstract to the point where the user can experience a, "[. . .] transcendence of the body and the inhabitation of a new one", understanding a virtual world from a completely different viewpoint (Boyer, 2009). The effective evaluation of the outcomes of such immersive

experiences arguably demands approaches which can penetrate and access the inner workings of the game experience, but given the metaphysical and experiential boundaries of game play outlined above, this is not straightforward. Added to this is the nature of immersion itself which Gunter *et al.* (2008) refer to as an "immersion hierarchy" comprising interaction, engagement and immersion and for game-players to be immersed in the game, all three levels require active participation, with both interaction and engagement necessary for immersion but interaction and engagement alone not guaranteeing immersion.

Problems associated with the evaluation of immersive game experiences, are not just restricted to serious games however, but are an industry wide phenomena. Defining an objective "player-in-the-loop" evaluation of quality of experience (QoE) which recognises the interdependence of the quality metrics of game playability, with the involvement of players' subjective factors, presents a serious challenge with some concern that there are aspects of the human experience that are individualistic, and therefore "[...] not measurable" at all (Chen and Zarki, 2011). Certainly, approaches adopted in serious game evaluation tend to either acknowledge this apparent "difficulty" or choose to adopt more traditional approaches, either accepting or not fully aware of the inherent methodological pitfalls. For example, Rojas-Barahona *et al.* (2012) review a method for evaluating two state-of-the-art dialogue systems developed to support conversation with French speaking virtual characters in the context of a serious game called The Mission Plastechnologie designed to promote careers in the plastic industry. The evaluation comprised a quantitative evaluation comparing the accuracy of the interpreter and dialogue manager integrated into each system, a corpus based evaluation where criteria such as dialogue coherence, success, interpretation and generation errors were collected and a user based evaluation involving 22 subjects. In the user based evaluation subjects were asked to play the game twice, once with each system and after playing each, users completed a questionnaire designed to evaluate the interpretative quality, overall system quality, dialogue clarity and timing. So in this case users oscillated in and out of the game experience, traversing the "immersion hierarchy" described by Gunter *et al.* (2008) four times with standard questionnaire completion in between and at completion; effectively subjecting users to an immersion roller coaster and expecting them to report objectively on their experiences.

To match the complexities of evaluating immersive experiences, some approaches use a battery of traditional tools and techniques. The educational game-based learning environment entitled Murder on Grimm Isle (MOGI) for example, is an immersive 2D gaming environment used to foster augmentation and persuasion writing for Grades 9-14 which uses a qualitative, grounded theory method with data collection including observations of student interaction, including chat within the game, observations of student activity within the computer laboratory setting, questionnaires, interviews, peer debriefing, members' check, negative case samples and audit trails (Dickey, 2011). The iterative features of grounded theory method, is arguably a good fit for the iterative aspects of participative design, and data collection is extremely comprehensive but much of the data collection from users falls outside of the game itself, is extensive and logistically cannot have all happened within a short time after playing the game. So not only is the data collection outside of the game, but the time lag between when players experience the game and report on it in all the various qualitative forms, surely compounds the problems further. The tool developed by

Hong *et al.* (2009) sets out to help educators assess the effectiveness of digital game based learning through a set of 74 game evaluation indices sorted into seven categories, namely mentality change, emotional fulfillment, knowledge enhancement, thinking skills development, interpersonal skill development, spatial ability development and bodily coordination. While this is a summative evaluation tool, not specifically designed for a games design context, the approach it adopts, clearly demonstrates the dilemma of trying to understand the impact of game based learning through the decomposition of an immersive experience, in this extreme example where the rubric is totally disconnected with the user.

In seamless evaluation the role of the participating user intensifies as the game and its evaluation are experienced as "one"; an approach which has close parallels with the idea of "in-world" co-creation of game development. In the games industry where computer and video game markets reached levels of motion-picture box office sales in 2006, there is much more intense user participation in the form of games modifications and massively multiplayer online games (MMOGs), although in general there is still an adherence to top-down development approaches. However, Volk describes an "in-world" development approach for a serious game, or meta-design which sets out to "[. . .] close the user-developer gap by intertwining [. . .] the concepts of development and play", and their corresponding platforms. This player-driven kind of bottom-up participation is viewed as a specialised version of paradigmatic change which bypasses the design conflict between the role of domain designer and player. Volk (2008) goes on to discuss the need for stronger emancipation of the user as co-developer at its core and the appearance of, "[. . .] immaterial good changes from versioned and desktop-centric products to ever-changing continuums". Volk calls this the "prod-user" continuum where game content becomes a kind of perpetual beta and the corresponding product-centric game turns into a service platform, "for playing the mods" and "[. . .] the act of playing the game and the process of developing it need to build a smooth transition", which do not force a player into a specific role, neither consumer nor designer.

While the scope of these "prod-user" contexts are beyond the scope and budgets of most educational game projects, some designers of serious games are however beginning to tackle the methodological challenge presented by the evaluation of immersion and feelings of "presence" they invoke. De Freitas *et al.* (2010) see learning design through the lens of "immersive learning experiences", rather than sets of knowledge to be transferred between tutor and learner, that require "[. . .] new methodologies for evaluating the efficacy, benefits and challenges of learning in these new ways". De Freitas *et al.* propose a methodology based upon inductive methods, augmented by a four-dimensional framework constructed around the learner, the pedagogic perspective, the representation itself and the concurrent position of the learning in physical and virtual space. This interdisciplinary approach does set the scene for closer consideration of the immersive experience and more critically game boundaries and suggests that the learning experiences need to be designed, used and tested in a multidimensional way due to the multimodal nature of the interface. However, the nature of research instruments and how these impinge on the immersive experience are not discussed and the focus is more on unpicking the complexities of the broader learning evaluation context than close consideration of how to actually implement any game evaluation experience, which reflects the significant gap between the variety of games based learning systems and the number of reliable ways to evaluate them.

Approaches are beginning to emerge however, which like the seamless approach used here attempt to embed evaluation in some way. For example, the evaluation of an immersive web based training programme for health care professionals called AOC – Anatomy of Care incorporates an embedded assessment algorithm, known as a knowledge assessment module called (NOTE) to capture both user interaction with the educational tool and information about knowledge gained from the training (Libin *et al.*, 2010). This learning application was developed for The Simulation-based Role-Playing Intervention Laboratory, affiliated with the Washington Hospital Centre in partnership with Potomac, Maryland-based WILL International Inc. In AOC learners are presented with stress "slice-of-life" scenarios in which they must make tough decisions and live out the consequences of their actions. AOC was introduced to 1,500 employees in various health care facilities over a three month period and the need for tools tailored to the evaluation of virtual role-playing training clearly emerged. The subsequent knowledge assessment module or NOTE developed for role-playing interventions focused on:

- exploring the interaction between gaming parameters and individual profiles; and
- analysing the interrelations between knowledge gain based upon comparison of preassessment and various individual factors.

AOC incorporates two major elements – five 3-minute video clips representing one of the stressful situations and a multiple choice-scale, representing different – effective and ineffective-decision making strategies. The authors describe the content of the video-clips and multiple-choice scale as reflecting an "[. . .] organisation tailored system of values that is part of the corporate culture" – suggesting careful co-design of learning and instrument and a close matching which helped to maintain game-based learners within their immersive context.

In seamless evaluation, instruments become part of the broader game experience; in fact inclusion of other elements or add-ins to games to achieve some kind of pedagogical impact, is an increasingly commonstrategy that many serious game based developers are adopting in order to achieved desired learning outcomes. Bellotti *et al.* (2008) discusses user exploration and learning in virtual worlds and the challenges of providing more in-depth information without interrupting the flow of the game. The Traveller in Europe (TiU) project is a treasure-hunt game where the player has to accomplish a mission by visiting cities spread across a map of Europe. Small embedded microGames (mGs) or trial games are used to enhance the learning by providing local contextualised heritage knowledge, for example, a game concerning Van Cleve's "Adoration of the Magi" triptych is played in a 3D construction of the San Donato church in Genoa's historical centre, where the picture is conserved. Preliminary informal tests on the impact of these 2D mGs on game play suggest that the approach is a valid one and the authors provide some guidelines about how to properly and smoothly integrate the trial games into a 3D environment so as not to distract the game-player, for example mGs should be short and focused upon a specific item, should have a precise educational/knowledge/skill acquisition target, difficultly should be scaled with the players performance and should not interrupt the player's expected flow of actions. Critically, mGs must not be boring educational add-ons but should, "[. . .] significantly enrich the environment [. . .] and should be well integrated in

the game logic and aesthetic". Similarly Rankin *et al.* (2008) discuss the development of a game plug-in prototype consisting of non-player characters or NPCs capable of real-time multi-modal conversation that emphasises L2 vocabulary to provide conversational support for foreign language students and enhance the game play experience. The authors stress that designers of serious games need a different design framework that "[. . .] interweaves social interactions, learning objectives and elements of play" in order to design serious games that have a positive learning impact on the player. While mGs are clearly explicit and obvious to the user, occasions arise when systems need to be embedded which are not overtly obvious to the user. Liu *et al.* (2011) developed a new real-world music composition application called MusicScore and used it as a running example and experimental testbed to evaluate design choices and implementation of an application called TouchTime – a mobile device using multi-touch gestures that can be streamed on the fly among multiple participating users, making it possible for users to engage in a collaborative or competitive experience. So the MusicScore experience takes full advantage of the TouchTime system with students benefiting from a live educational experience, but without any knowledge of the architectural design choices being made in TouchTime; effectively one application being used to evaluate another, without the knowledge of the user.

In summary, seamless evaluation addresses a fundamental methodological problem of trying to understand the user experience *B* game play. While other designers and developers of serious games are approaching this problem in similar ways, both pedagogically and practically, the contribution of seamless evaluation is that it embeds a summative, multiple instrument evaluation entirely within the game-play, making a serious attempt to capture data during an immersive game experience.

Seamless approaches and validity

Ethical issues need to be considered in any kind of research, but come to the fore particularly where there is any kind of deception of participants (Orb *et al.*, 2000). The key rationale for conducting embedded game evaluation comes from an inherent weakness in trying to establish a causal connection between a treatment and an outcome where boundaries are blurred (Moore *et al.*, 2003), in this case an immersive educational game experience and learning outcomes. As Howe (2004) explains, "[. . .] acquiring a better understanding of causal mechanisms requires substantive knowledge of the contents of the black box", something that cannot be achieved by employing more formal experimental devices. To get inside the "black box" of game-play immersion, a seamless evaluation approach necessitates the design of research instruments which mislead users into thinking that they are actually part of the game, the rationale being to keep users within the magic circle of game-play in order to collect more reliable data about the experience; this inevitably raises ethical issues because of the deception involved. In a research context "deception" occurs when subjects are misled about the nature of research procedures (Chambliss and Schutt, 2009). The word deception itself, even without any precise definition, carries negative connotations. However, Hertwig and Ortmann (2008a, b) discuss the consensus across disciplinary boundaries that it is the, "[. . .] intentional and explicit provision of erroneous information", or lying, that counts as deception, "[. . .] whereas withholding information about research hypotheses, the range of experimental

manipulations, or similar, should not be thought of as deception. Or in other words, there is a "[...] world of difference between not telling subjects things and telling them the wrong things" (Hey, 1998). From an analysis of US research regulatory bodies which define the parameters for the use of deception, Hertwigg and Oattman conclude that deception is viewed as a last-resort measure in social science research, but in an analysis of the frequency of deception in the *Journal of Experimental Social Psychology* (*JESP*) in 2002, they reported that out of 117 studies, 63 or 53 percent used deception in some form in their experimental methods, so the use of deception in a research context has been shown to be widespread, in the context of social science research.

So what are the advantages of deception and how can this be used as a means of enhancing study validity? In the seamless evaluation approach, the goal is to control the research environment by maintaining the user inside the game play experience. Given that the game play itself is an educational intervention designed to be fun, engaging and ultimately to transfer knowledge, the extended game-play is intrinsically beneficial in itself to the user, this is easy to defend. However, this can raise questions as to the validity of the instruments, how this impacts upon the user and their response set, and consequently the reliability of the data collected. Pascual-Leone *et al.* (2010) point out that establishing control is particularly important when the aim is to study behaviour that can only be accessed where participants are uninformed. They allude to a range of factors that impact upon how people respond to research instruments, and the importance of eliciting spontaneous behaviour from research participants. They argue that deception can thus be used as a tool to enhance both the internal and external validity of a study by encouraging responses in a more natural and uninhibited manner and it can be inferred from this that respondents do not need to know the purpose of a research instrument to answer it with the spontaneity of the response being the critical issue.

Goals of the seamless evaluation meta-game

With two primary users of our approach (the players and the developers), the goals of seamless evaluation were twofold:

(1) For the development team: to create an evaluation that gathered quantitative and qualitative data that would enable assessment of the application's ability to achieve the desired goals and outcomes and to inform future design and development.

(2) For users to take part in an evaluation that was invisible and seamless, adding value to the player experience and occurring within the magic circle of game play (and ensuring that evaluation was not a burden).

To retain magic circle integrity, seamless evaluation had its own plot, story and rationale that complemented, and in places expanded upon the game being evaluated, thus creating a broader and richer meta-game. The entire evaluation was undertaken in-role: that is all participants (players and evaluators or members of the development team), the physical space, artifacts, instruments, interactions and measurements were designed around the game play. Thus, everything from the basic wording of the instruments to the overall look and feel of the supporting artifacts was:

- Designed to be congruent with the narrative and context of the game being evaluated.
- Met participant expectations and added value (e.g. additional information, rewards, enjoyment) to the player experience.
- Provided the development team with essential information whilst placing as little burden as possible on the participant.

Seamless evaluation has been specifically designed for role-play games and focuses on ensuring that players are in-role from the moment that they enter the evaluation, which occurs only moments after they have arrive for a session, right until returning to the real world, moments before they depart. In this seamless evaluation study, players received very limited information about the experience that they were about to participate in and were told simply that it involved playing a computer game.

Evaluation context: ORIENT

ORIENT provides users with an intelligent computer assisted, semi-immersive, graphical role-play environment depicting an imaginary culture, including "Spryte" characters. It is aimed at teenagers and young adults who interact in groups of three, taking roles in Space Command (a benevolent United Nations type of organisation with a galactic focus) with the goal of helping the Sprytes to save their planet from imminent destruction. ORIENT's learning focus is cultural understanding and sensitivity.

The Sprytes characters inhabiting this world are autonomous agents, based on an extension of the FAtiMA agent architecture (Dias and Paiva, 2005). Emotional appraisal is based on the OCC cognitive theory of emotions (Ortony *et al.*, 1988) extended by incorporating aspects of a needs driven architecture, PSI (Dörner, 2003). To enable cultural adaptation of the agents, Hofstede's cultural dimension values were added to the agent minds for the culture of the character, cultural specific symbols culturally specific goals and needs, and the rituals of the culture (Hofstede *et al.*, 2010).

Users interact with the Sprytes using a Wii-mote to provide gestures and speech recognition of character names. They interact with the ORIENT world using a scanner phone with an RFID reader. Additionally, the users are provided with the Onboard Resource Agent – Cultural and Liaison Engagement (ORA-CLE), a mobile phone based embodied conversational agent whose role is to support the users in their interaction. Figure 1 shows an overview of ORIENT's main components. At the core of the system is the virtual world model that is presented tothe user as 3D graphics on a large screen, in front of which the users interact with ORIENT as a group.

Developed as part of an interdisciplinary project, the pedagogical and psychological evaluation aimed to investigate the effectiveness of ORIENT in fostering cross-cultural acceptance through the promotion of collaborative practices and the appreciation of similarities and differences between cultures. From the technical perspective, evaluation focused upon the coherence and comprehensibility of the narrative, the believability and credibility of the agents that underpin the characters, and participant engagement with the cultures of ORIENT and the Sprytes themselves. With the interaction approach, we focused upon evaluating the participant's views of the impact of unusual interaction devices and mechanisms, exploring device usability and user satisfaction with unusual interaction mechanisms. This resulted in a wide range of evaluation goals and corresponding instruments, as detailed in Table I.

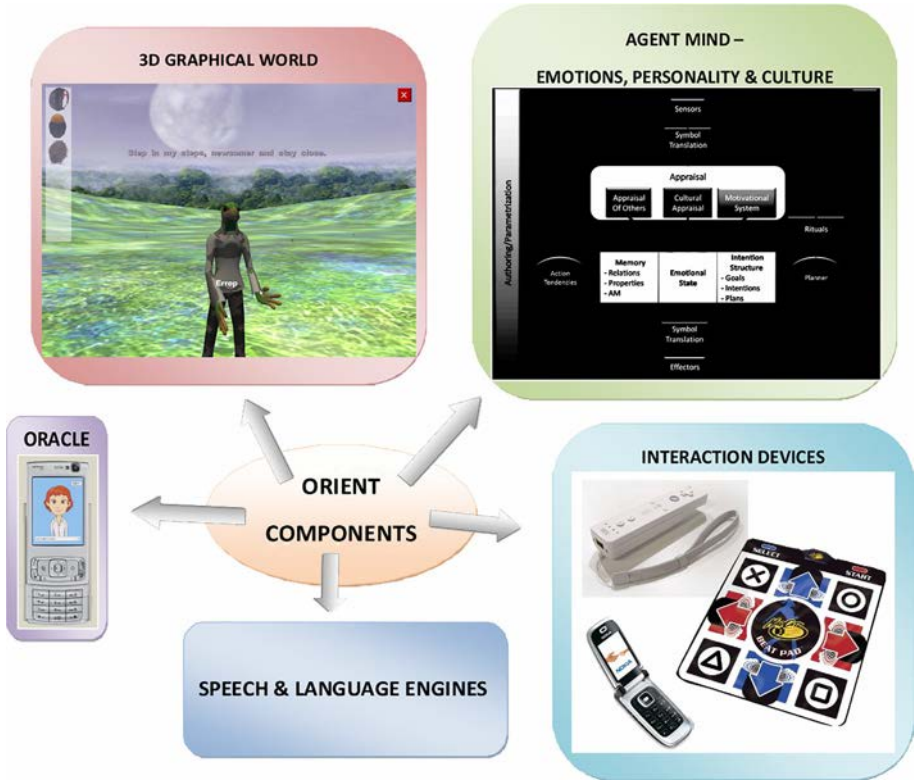

Figure 1.
ORIENT overview

From obtrusive to seamless evaluation

As can be seen from Table I, a significant amount of evaluation was required. If a traditional approach had been taken to ORIENT's evaluation, the user would have engaged in several non-player related roles, as evaluators (e.g. they would have been told that they were interacting with innovative software to evaluate it), subjects (e.g. completing various pre- and post-interaction questionnaires, etc.), as critics (e.g. taking part in a debrief to discuss their experience) and as learners (e.g. being taught how to use the various devices). Instead our goal was for players to have only one experience, that of being a player in a role-play game.

In the stand-alone prototype, the original intention had been for users to take the role of Space Command Staff on a mission to the planet ORIENT. Through slightly changing the player role from staff to intern and placing the ORIENT interaction within a training programme rather than as a mission, we provided a context that could both support the congruity of our evaluation requirements and provide a familiar situation. This resulted in a four-stage session:

- *Stage 1.* Into role: focus on Space Command with user as Intern; mission aims for ORIENT.
- *Stage 2.* Preparing for ORIENT: increasing knowledge about Sprytes and ORIENT, training with devices, planning the mission.

Evaluation goals	Instrument/approach
Demographic characteristics/cultural profile	Participant questionnaire
Cultural intelligence	Cultural intelligence scale (Ang *et al.*, 2007)
Perception and expectations of game play	Qualitative/open instrument to assesspre-interaction views and expectations of game play
INTERACTION with ORIENT	ORAT – logging and assessment of user behaviour based on researcher observation
Outgroup/cultural view (with regards to the outgroup "Sprytes")	Cultural activities questionnaire (amalgamating the intergroup anxiety scale (Stephan, 1985) and the general evaluation scale (Wright *et al.*, 1997)
Cultural understanding	Cultural understanding questionnaire: qualitative and quantitative measures to assess users' understanding of the Spryte culture
Device use	Device questionnaire: qualitative usability evaluation questionnaire with open questions/free text
Quantitative evaluation of ORA-CLE	Usability questionnaire
Response to ORIENT and Sprytes (e.g. graphics, speech, storyline, agent believability, etc.)	Based upon the character evaluation questionnaire (Hall *et al.*, 2006)

Table I.
Constructs and corresponding instruments

- *Stage 3.* ORIENT interaction: carry out and complete fact finding mission.
- *Stage 4.* Debrief and post-test questionnaires: report on mission, response to Sprytes and ORIENT, response to Space Command's technology and gadgets.

The set, scripts, cast and props had to be created to support the evaluation session. The set was provided as the Space Command Training Room, with a large wall for projection. This location was set up (with chairs and desks) to enable users to complete the instruments, to participate in the training and interact with ORIENT. The technology, interfaces and interaction devices were visible and accessible to the users from the moment they entered the room. The cast was composed of Space Command Staff with scripts providing dialogue, key information to be given by cast to the players and stage directions. Props involved hard copy materials (e.g. crib sheets of known Sprytes), the evaluation instruments andinteractive elements, including a 10 minute training video and interactive demo used to support players with their use of the devices.

The evaluation required by the development team was achieved mainly through transforming traditional and/or well established data gathering instruments into their "in role" counterparts. These were then embedded into the role-play and reinforced with supporting artifacts. Each instrument was given archetypal branding (adding value to the role play context) and an age appropriate format and aesthetic (meeting user expectations). Instruments were individually developed, piloted and refined. Although initial instrument design favoured a consistency in terms of look and feel, a final "harlequin" approach was adopted in response to users preferences for a mixture of different designs and layouts. The resulting battery of piloted instruments (Figure 2 and Table II) were then assessed to ensure complementarity, before integration into the role play experience with the aim of adding maximum value to the over-arching role playing game, while collecting key evaluation data to help developers assess the user experience from a number of theoretical perspectives.

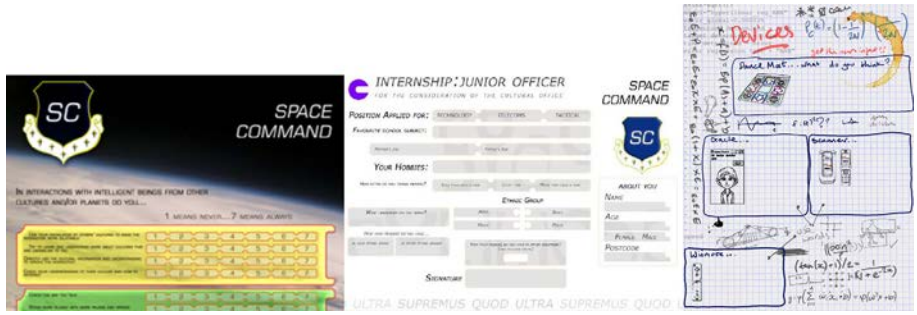

Figure 2.
Questionnaire examples

Instrument/approach	Seamless evaluation instrument
Participant questionnaire	Culture office: internship registration form
Cultural intelligence scale (Ang *et al.*, 2007)	Cultural role allocation form
Qualitative/open instrument to assess pre-interaction views and expectations of game play	Mission plan
ORAT – logging and assessment of user behaviour based on researcher observation	Users taking part in Space Command Intern training, being assessed by mission command using ORAT
Cultural activities questionnaire (amalgamating the intergroup anxiety scale (Stephan, 1985) and the general evaluation scale (Wright *et al.*, 1997)	Intelligence form 1B
Cultural understanding questionnaire: qualitative and quantitative measures to assess users' understanding of the Spryte culture	Cultural form 4C: report on alien intelligence
Device questionnaire: qualitative usability evaluation questionnaire with open questions/free text	Devices, etc.
Usability questionnaire	Neurotek A.I. evaluation form
Based on the character evaluation questionnaire (Hall *et al.*, 2006)	Neural networks assessment form

Table II.
Seamless evaluation instruments

The instruments were not only used in an evaluation sense, but also to increase the coherence and credibility of the game play. For example, it was critical that the users were quickly embedded in role and the use of the participant questionnaire, clearly placed them in their role as Interns on a training programme. With the interaction devices questionnaire, we made specific attempts to reduce the burden of an instrument that requires written input, through creating an instrument that bears no real resemblance to a questionnaire or meets the perceptions of an evaluation instrument. In role, the interaction devices questionnaire had its own unique branding, simply a scrap of paper from the "techies" in the Space Command Gadget Department. With the aim of increasing user willingness to engage with this instrument, the interaction device questionnaire was administered by the technical team, who had assisted the participants in training and with technical problems with device use during the session. The role-play provision of this instrument encouraged participants to feel that they were helping the "lab rats."

Overview of framework for seamless evaluation

Reviewers ask for overview/visual/summary of approach including specific details about novelty (Figure 3).

Did the evaluation add value?

We initially developed this evaluation approach for ORIENT without really considering its impact on improving fun in game play. Instead, our focus had been to remove the burden of assessment. During the initial evaluations of ORIENT, in brief informal discussions after the role-play ended, the players were unanimously positive about their experience. Notably, the players did not seem to realise that they had participated in an evaluation. Rather they saw the completion of the evaluation instruments as part of the game play. All of the players had enjoyed what had been an innovative and engaging few hours and although they actually filled in many questionnaires (nine), not a single user alluded to any evaluation burden. Even when we briefly discussed the instruments, users did not focus on evaluation, rather they were more interested on discussing the branding and content, with almost all comments being positive.

Our approach had been an intuitive response to reducing the evaluation burden, however, although our intention had been to retain the integrity of the magic circle through seamlessly embedding the evaluation, we were extremely surprised at how effective this had been. Seamless evaluation had enabled us to create an evaluation experience congruent with the narrative and context of ORIENT and to gather useful data. It had enabled us to gather results that were useful for the development team, and further, the evaluation instruments and approach had appeared to add value to the user experience. Although this was the "gut" feeling we had from observing players and anecdotal evidence, we decided to further assess the potential of this seamless evaluation approach by explicitly evaluating the entire ORIENT experience (evaluation, interaction and all).

Whilst our goal with the evaluation of ORIENT had been to retain the game-play experience, we also conducted a final assessment of the seamless evaluation approach itself, in which we regressed to traditional approaches and ruptured the magic circle. This involved the creation of an out-of-role questionnaire that we explicitly asked the users to fill in. This questionnaire was a standard, lightly branded (with project information) text-based questionnaire. It focused on all elements of the role-play, including ORIENT and the usability of interaction mechanisms and devices.

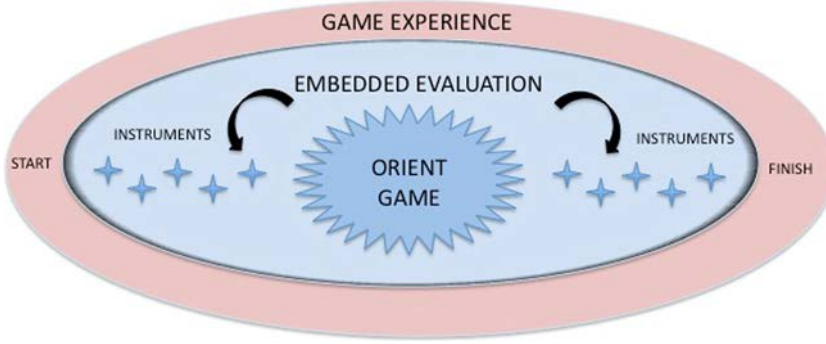

Figure 3.
Model of embedded evaluation

The questionnaire was composed primarily of five-point Likert scales using bi-polar adjectives/statements, along with some open questions. It included sections on:

- taking part: levels of engagement in the role play, contribution to the team, involvement in activities;
- interacting with ORIENT: views of immersion, control, feedback, clarity of goals;
- impact on role play: contribution of different elements of the experience to the role play, role play support material, interaction with ORIENT, ORIENT's story/narrative, interaction partners (users and evaluation team) and the interaction devices; and
- learning potential: relevance of ORIENT role-play experience to learning about different cultures and for the intended user group (teenagers).

Method

The assessment of the seamless evaluation of ORIENT involved 18 adults aged from 18 to 27. They participated in the four stage session with a final additional session, where they were asked to complete the out-of-role questionnaire and to participate in a post-evaluation debrief, where they were directly asked about their role-play experience.

Results

The key result from this out-of-role questionnaire and debrief was that during the seamless evaluation, users were unaware that they had been participating in an evaluation at all, with the instruments enhancing rather than detracting from the in-role experience. During the debriefing sessions, the majority of users were surprised to discover that they had been filling in evaluation and assessment instruments during the role-play experience. The only evaluation instrument that all of the users were aware of was the final out-of-role evaluation questionnaire.

All of the debrief groups were positive about the whole role-play experience with criticism focusing primarily on the interaction devices and the difficulty of engaging in the interaction due to technological barriers. The evaluation instruments were viewed positively by participants and as adding to the experience, with no users feeling that the activity had involved too much paperwork or form filling.

The users were asked to rate the impact of several aspects of the preparation phase of the interaction (the surrounding information given, the support by staff and documentation given during the interaction); "Surrounding Information" was rated highest with a narrow range of results (mean = 4.39, SD = 0.608). The other two categories were scored at 4.06 suggesting the users were generally happy with the impact of these elements of the role-play experience. There was a relatively high correlation between the extent to which the participants felt immersed within the interaction and their rating of the contribution that the support documentation, including evaluation instruments, made to the experience ($df = 9$, $\chi^2 = 11.938$). The participants' rating of how much they lost track of time was also strongly correlated to their rating of the support documentation ($df = 12$, $\chi^2 = 10.933$).

The impact of both the user's collaboration with colleagues and the evaluation support team were scored highly as having a significant impact on the interaction experience (mean = 4.2, SD = 0.77). The level to which the users felt immersed was

also correlated with their rating of the impact of the support evaluation team on the interaction (df = 9, $\chi^2 = 9.870$), though this correlation was not as strong as between user immersion and support documentation. The correlation between immersion and the contribution of training was weaker still (df = 9, $\chi^2 = 8.501$). It seems that in this case the users report documentation as being more strongly correlated to immersion than either the effect of the evaluation support team or the use of training.

These results highlight that the seamless evaluation approach used with ORIENT was clearly within the magic circle and added value to the user experience. The evaluation instruments and activities were seamless and data capture invisible for the user. Rather than the evaluation instruments and supporting artifacts adding a burden to the user, they seemed instead to enhance the game, actually increasing the magic circle and the immersion of the users.

Discussion

Evaluation is predominantly an activity associated with appraisal, assessment, judgement and explanation, which to the player is typically not productive, not fun and yields no tangible return. Seamless evaluation aims to provide an improved experience of the evaluation process. It strives to deliver all data gathering processes within an experiential context that is meaningful and fun to the players. In the case of ORIENT, seamless evaluation effectively created a smoke screen around the appraisal process, so that it did not detract from the experience of the game being evaluated, and actually added to it instead.

Seamless evaluation has considerable applicability for other games, games-based learning, entertainment and social applications and clearly demonstrates the potential of extending the magic circle through incorporating evaluation to enhance the user experience. Although this approach was initially developed to evaluate intelligent computer assisted role-play games, it is applicable to the majority of social and recreational software. It is of particular relevance to those trying to evaluate personal, social and emotional impact of interactions.

Seamless evaluation has a positive impact on the validity of assessment instruments, primarily using established instruments to form the basis of in-role evaluation which maintains the integrity of the research tool but delivers it visually and structurally in a manner which is cohesive with the in-role experience. Further the seamlessness by which instruments are administered enables data to be generated in a way that is more meaningful to users. However, although seamless evaluation offers an improved user experience of evaluation, it does have significant resource implications. In addition to the integrity of instrument design, data capture methods needed to be visually and textually consistent with the application to be evaluated. The holistic in-role experience is challenging to develop and requires considerable time, effort and piloting to ensure that it is effective.

Due to the experimental nature of the protocol and time constraints, all instruments used with ORIENT were hard copy. It could be argued that delivering these tools digitally through the game would have thickened the smoke screen of evaluation even more; this is certainly an approach that needs further investigation. However, the materiality of the research instruments, designed as they were to extend and contribute to the role-play, arguably strengthened the magic circle even more by bringing to specific features of the virtual world a physicality which is immediately perceptible to

users through handling and completion of a game artefact with a pen or pencil. Further these artefacts, like the various interaction devices, blurred the boundaries between virtual and real space for the player.

Conclusions

Evaluation methodologies need to recognise and respond to user expectations as well as generating useful data for development teams. The seamless evaluation approach offers considerable potential for engaging users and has been highly successful in its application to the evaluation of ORIENT. The instruments and approach generated useful pedagogical, psychological, technological and interaction results for the development team. Notably, users were unaware that they were being evaluated, viewing the instruments and activities as part of the game experience, with the evaluation adding value to the participant's enjoyment and engagement.

In evaluating and improving games our aim must be to evaluate the game not the player. Doing this invisibly seems sensible. Doing this in a way that adds value and increases fun is even better. Seamless evaluation achieves these aims, fusing play and evaluation together into a meta-game, embedding the mechanisms for gleaning meaningful evaluation data within the magic circle of the game itself.

References

Ang, S., Van Dyne, L., Koh, C., Ng, K.Y., Templer, K.J., Tay, C. and Chandrasekar, N.A. (2007), "Cultural intelligence: its measurement and effects on cultural judgment and decision making, cultural adaptation and task performance", *Management and Organisation Review*, Vol. 3 No. 3, pp. 335-371.

Baranowski, T., Buday, R., Thompson, D.I. and Baranowski, J. (2008), "Playing for real video games and stories for health-related behavior change", *American Journal of Preventive Medicine*, Vol. 34 No. 1, pp. 74-82.

Bellotti, F., Berta, R., De Gloria, A. and Zappi, V. (2008), "Exploring gaming mechanisms to enhance knowledge acquisition in virtual worlds", *ACM DIMEA'08 Athens, Greece*, pp. 77-84.

Boyer, S. (2009), "A virtual failure: evaluating the success of Nintendo's Virtual Boy", *The Velvet Light Trap*, No. 64, pp. 23-33.

Chambliss, D.F. and Schutt, R.K. (2009), *Making Sense of the Social World: Methods of Investigation*, 3rd ed., Pine Forge Press, Thousand Oaks, CA.

Chen, P. and Zarki, M.E. (2011), "Perceptual view inconsistency: an objective evaluation framework for online game quality of experience (QoE)", *IEEE Proceedings of NETGAMES*, pp. 1-6.

De Freitas, S., Rebolledo-Mendez, G., Liarokapis, F., Magoulas, G. and Poulovassilis, A. (2010), "Learning as immersive experiences: using the four-dimensional framework for designing and evaluating immersive learning experiences in a virtual world", *British Journal of Educational Technology*, Vol. 41 No. 1, pp. 69-85.

Dias, J. and Paiva, A. (2005), "Feeling and reasoning: a computational model", in Seabra Lopes, L., Lau, N., Mariano, P. and Rocha, L.M. (Eds), *12th Portuguese Conference on Artificial Intelligence*, Springer, New York, NY, pp. 127-140.

Dickey, M.D. (2011), "Murder on Grimm Isle: the impact of game narrative design in an educational game-based learning environment", *British Journal of Educational Technology*, Vol. 42 No. 3, pp. 456-469.

Dörner, D. (2003), "The mathematics of emotion", in Detje, F. and Schaub, H. (Eds), *The Logic of Cognitive Systems: Proceedings of the Fifth International Conference on Cognitive Modeling*, Springer, New York, NY, pp. 127-140.

Gunter, G.A., Kenny, R.F. and Vick, E.H. (2008), "Taking educational games seriously: using the RETAIN model to design endogenous fantasy into standalone educational games", *Education Technology Research Development*, Vol. 56, pp. 511-537.

Hall, L., Woods, S. and Aylett, R. (2006), "FearNot! Involving children in the design of a Virtual Learning Environment", *Artificial Intelligence & Education*, Vol. 16 No. 4, pp. 237-251.

Hertwig, R. and Ortmann, A. (2008a), "Deception in experiments: revisiting the arguments in its defense", *Ethics and Behaviour*, Vol. 18 No. 1, pp. 59-92.

Hertwig, R. and Ortmann, A. (2008b), "Deception in social psychological experiments: two misconceptions and a research agenda", *Social Psychology Quarterly*, Vol. 71 No. 3, pp. 222-227.

Hey, J.D. (1998), "Experimental economics and deception: a comment", *Journal of Economic Psychology*, Vol. 19, pp. 397-401.

Hofstede, G., Hofstede, G.J. and Minkov, M. (2010), *Cultures and Organisations, Software of the Mind, Intercultural Cooperation and It's Importance for Survival*, 3rd ed., McGraw-Hill, New York, NY.

Hong, J.C., Cheng, C.L., Hwang, M.Y., Lee, C.K. and Chang, H.Y. (2009), "Assessing the educational value of digital games", *Journal of Computer Assisted Learning*, Vol. 25, pp. 423-437.

Howe, K. (2004), "A critique of experimentalism", *Qualitative Inquiry*, Vol. 10 No. 1, pp. 42-61.

Huizinga, J. (1950), *Homo Ludens: A Study of the Play-Element in Culture*, Roy Publishers, New York, NY.

Johnson, L.F. and Levine, A.H. (2008), "Virtual worlds: inherently immersive, highly social learning spaces", *Theory into Practice*, Vol. 47, pp. 161-170.

Klesen, M. (2005), "Using theatrical concepts for role-plays with educational agents", *Applied Artificial Intelligence*, Vol. 19 Nos 3/4, pp. 413-431.

Libin, A., Laouderdale, M., Millo, Y., Shamloo, C., Spencer, R., Green, B., Donnelan, J., Wellesley, C. and Groah, S. (2010), "Role-playing simulation as an educational tool for health care personnel: developing and embedded assessment framework", *Cyberpsychology, Behaviour and Social Networking*, Vol. 13 No. 2, pp. 217-224.

Liu, Z., Feng, Y. and Li, B. (2011), "When multi-touch meets streaming", *ACM MUM'11 Beijing, China*, pp. 23-32.

Moore, L., Graham, A. and Diamond, I. (2003), "On the feasibility of conducting randomised trials in education: case study of a sex intervention", *British Educational Research Journal*, Vol. 29 No. 5, pp. 673-689.

Orb, A., Eisenhauer, L. and Wynaden, D. (2000), "Ethics in qualitative research", *Profession and Society*, Vol. 33 No. 1, pp. 93-96.

Ortony, A., Clore, G.L. and Collins, A. (1988), *The Cognitive Structure of Emotions*, Cambridge University Press, Cambridge.

Pascual-Leone, A., Singh, T. and Scoboria, A. (2010), "Using deception ethically: practical research guidelines for researchers and reviewers", *Canadian Psychology*, Vol. 51 No. 4, pp. 241-248.

Rankin, Y.A., McNeal, M., Shute, M.W. and Gooch, B. (2008), "User centred games design: evaluating massive multiplayer online role playing games for second language acquisition", *ACM Sandbox Symposium, Los Angeles, CA*, pp. 43-49.

Rojas-Barahona, L.M., Lorenzo, A. and Gardent, C. (2012), "An end-to-end evaluation of two situated dialog systems", *ACL 13th Annual SIGDIAL, Seoul, South Korea*, pp. 10-19.

Salen, K. and Zimmerman, E. (2003), *Rules of Play: Game Design Fundamentals*, MIT Press, Cambridge, MA.

Sheng, S., Magnien, B., Kumaraguru, P., Acquisti, A., Cranor, L.F., Hong, J. and Nunge, E. (2007), "Anti-Phishing Phil: the design and evaluation of a game that teaches people not to fall for phishing", *Privacy and Security*, Vol. 229 No. 1, pp. 88-99.

Stephan, W.G. (1985), "Intergroup relations", *The Handbook of Social Psychology*, Random House, New York, NY, pp. 559-659.

Stromberg, H., Vaatanen, A. and Raty, V. (2002), "A group game played in interactive virtual space: design and evaluation", *4th Conference on Designing Interactive Systems*, ACM, New York, NY, pp. 56-63.

Volk, D. (2008), "Co-creative game development in a participatory Metaverse", CPSR Indiana University, Bloomington, IN, pp. 262-265.

Wright, S.C., Aron, A., McLaughlin-Volpe, T. and Ropp, S.A. (1997), "The extended contact effect: knowledge of cross-group friendships and prejudice", *Journal of Personality and Social Psychology*, Vol. 73 No. 1, pp. 73-90.

Corresponding author
Susan Jane Jones can be contacted at: susan.jones@sunderland.ac.uk

To purchase reprints of this article please e-mail: **reprints@emeraldinsight.com**
Or visit our web site for further details: **www.emeraldinsight.com/reprints**