Chapter 18
Study Design and Data Gathering Guide for Serious Games' Evaluation

Jannicke Baalsrud Hauge
Bremer Institut für Produktion und Logistik (BIBA), Germany

Elizabeth A Boyle
University of the West of Scotland, UK

Igor Mayer
Technical University of Delft, The Netherlands

Rob Nadolski
Open University of The Netherlands, The Netherlands

Johann C. K. H. Riedel
Nottingham University, UK

Pablo Moreno-Ger
Universidad Complutense Madrid, Spain

Francesco Bellotti
Università degli Studi di Genova, Italy

Theodore Lim
Heriot-Watt University, UK

James Ritchie
Heriot-Watt University, UK

ABSTRACT

The objective of this chapter is to provide an overview of the different methods that can be used to evaluate the learning outcomes of serious games. These include Randomised Control Trials (RCT), quasi-experimental designs, and surveys. Case studies of a selection of serious games developed for use in higher education are then presented along with evaluations of these games. The evaluations illustrate the different evaluation methods, along with an assessment of how well the evaluation method performed. Finally, the chapter discusses the lessons learned and compares the experiences with the evaluation methods and their transferability to other games.

DOI: 10.4018/978-1-4666-4773-2.ch018

INTRODUCTION

In the last decade higher education has taken a digital turn in the use of games and simulations for learning and training. The long and well-established tradition of using teacher-led, no-technology or low-technology simulation games in higher education is 'under the spell' of online simulations, 3-D virtual worlds and digital Serious Games (SGs). So, what have we gained and/or possibly lost with this digital turn to Game-based Learning (GBL)? To answer this question we need to have ways of evaluating the learning impact of games. This chapter sets out to review and provide examples of the different evaluation methods that can be applied to serious games.

Considerable efforts and resources are now being put into the evaluation and assessment of game-based learning. As a result, both the number and the quality of evaluations of games for learning are increasing (see for a recent overview Connolly et al., 2012). However, there are still considerable weaknesses, for example, the absence of tools for unobtrusive, 'stealth' data gathering and assessment, and good research designs other than randomized controlled trials. Here, we wish to make a contribution by looking at how different evaluation methods have been applied to some serious games and to see what has been measured and how.

This chapter will present several case studies of serious games and their evaluation methodologies. It will identify the differences in the evaluation methods, and also discuss what this means for the transferability of the evaluation methods to other types of games.

EVALUATION METHODS FOR SG LEARNING OUTCOMES

The evaluation of games is complex and multidimensional since it involves evaluation not just of whether there is an improvement in performance on the targeted learning outcomes, but also evaluation of the user acceptance of, engagement with, and satisfaction with the game. The introduction of a serious game into the curriculum raises similar issues to any other educational intervention, since the aim of a game is to improve performance on a specific learning outcome. Woolfson (2011) proposes a hierarchy of evidence for evaluating educational interventions:

1. Meta-analyses.
2. Randomised controlled trials (RCT).
3. Quasi-experimental designs.
4. Single case experimental designs–pre & post test.
5. Non experimental designs–surveys, correlational, qualitative.

Meta-Analyses: At the top of the hierarchy of evidence for the effectiveness of interventions are meta-analyses. Meta-analysis combines the results from previous studies to identify patterns in research findings, especially with respect to whether games are effective methods in learning. Meta-analysis requires a reasonable number of empirical studies as input to compare – in serious games we still have a way to go to produce the needed studies, hence it has not been included in this chapter.

Randomised Control Trials (RCT): The Randomised Control Trial (RCT) is considered to be the gold standard for evaluating educational interventions. In a RCT participants are randomly allocated to an experimental (game) group or a control (non-game) group and their performance on the target skill/behaviour before and after the game intervention is tested. Ideally pre-testing should confirm no existing difference between the groups, while post-testing should show whether the experimental group performs better than the control group. Improvements in the target skill/behaviour for the experimental compared with the control group in a follow-up study would allow further confirmation that the intervention was successful.

Papastergiou (2009) developed a game to teach computer memory concepts and carried out a classic RCT, comparing the performance of a games group with a control group on tests of knowledge of computer memory concepts before and after the serious game intervention. She found that students in the gaming group performed better and also liked the game based approach better than students in the control condition. This provides evidence to support the view that educational computer games can be exploited as effective and motivating learning environments. This study raised an interesting methodological point which is true for many educational studies. In a true RCT each participant is randomly assigned to a gaming or non-gaming condition, but in this case participants were randomly assigned by intact classes to gaming or non-gaming groups.

Beale, Kato, Marin-Bowling, Guthrie and Cole (2007) carried out a RCT to investigate whether a video game, Re-Mission, could actively involve young people with cancer in their own treatment and increase self-care and cancer illness knowledge. A test on cancer-related knowledge was given prior to game play (baseline) and again after 1 and 3 months. Knowledge test scores for both control and experimental groups improved significantly over the follow-up periods, but the significant group by time interaction showed that the scores of the experimental Re-Mission game group improved significantly more than the control group($F(1,302)= 4.07$, $p= .04$, $f=.013$).

Quasi-Experimental Designs: While a RCT requires the random assignment of participants to experimental or control groups, in educational interventions this is not always possible. In that case a quasi-experimental design would have to be used (Field and Hole, 2003). This kind of design is also used to refer to a one group post-test design where participants' behaviours are measured following an intervention and to a one group pre-test/post-test design where participants' performance is measured before and after the intervention. In group comparison designs, the performance of two (or more) groups is measured after the intervention. These designs are all of lower quality than a RCT, but for pragmatic reasons may have to be used in real world research. An example of a study that compared four different groups but only after the intervention was Cameron and Dwyer (2005) who compared the impact of four different instructional conditions on knowledge acquisition in learning about the operation of the human heart: the digitised instructional unit with (a) no game plus questions, (b) game plus questions, (c) game plus questions plus knowledge of accuracy of response to questions, and (d) game plus questions plus elaborative feedback which provided the answer to the question and reasons why that was correct. The results showed that there was no difference in performance in the no-game condition (a) and the game condition (b), suggesting that the competitive structure of the game was not sufficient to increase knowledge retention. However, there were significant advantages on two outcome measures when response feedback was introduced and on all the performance measures when elaborative feedback was included, indicating that feedback to players about the accuracy of their responses was more important than the competitive structure of the game. While not an RCT, this kind of study can clearly provide detailed information about how different kinds of game mechanics provide support for learning in a game.

Surveys: Survey research typically uses a questionnaire methodology to ask many respondents about their attitudes to, perceptions of, or use of games generally, or of a specific game. The results are typically reported in terms of descriptive statistics reporting for example what percentage of people play games, intend to play games, enjoyed a game or felt that the game had helped them achieve the intended skills. Some studies, such as Connolly et al (2007) and Karakus et al (2008) examined game playing generally, while others, such as Lindh et al (2008), studied students' use of a specific game. Surveys can also be used as part of a formative evaluation or user

requirements analysis to assess whether potential players of a game would perceive a particular kind of game as useful.

Connolly et al (2007) surveyed Scottish students about their game playing habits, their motives for playing both entertainment and educational games and their acceptance of educational games in Higher Education. Findings confirmed the popularity of playing entertainment games as a leisure time activity for students, especially male students. There was also a high level of acceptance amongst students that games could be used for learning in Higher Education. Fewer female students played games and those who did play played less and played a less varied selection of games than males, suggesting that there may still be some way to go in persuading female students of the value of computer games in learning.

Rather than just reporting descriptive data, it is possible to carry out more sophisticated analysis with survey data, looking at links between variables and this would typically be done where a theoretical model is being tested. Weibel et al (2008) for example used regression analysis to examine the relationship between engagement variables, presence, flow and enjoyment, in an online game. They found that flow mediated the relationship between presence and enjoyment.

Structural equation modelling has also been used and again this kind of analysis would typically test a theoretical model. The Technology Acceptance Model (TAM) proposes that the perceived ease of use and perceived usefulness of a software application determines how much it will be used. Hsu & Lu (2004) tested an extended version of the TAM model and found that social norms (i. e. players' perceptions of other people's views of the technology), critical mass (the number of people using the technology) and flow were more important in predicting time spent playing entertainment games than the traditional TAM variables.

Qualitative Research: In terms of the hierarchy of evidence, qualitative research is regarded as lower quality than quantitative research. Qualitative research is more subjective than quantitative since it is more interpretative, but it can provide a much broader brush approach to examining the skills that playing games can support.

Steinkuehler and Duncan (2008) reported a high quality qualitative analysis of the scientific reasoning skills displayed by players in their contributions to the online discussion boards while they played the popular online game, World of Warcraft (WoW). Steinkuehler and Duncan developed a rigorous coding system for players' contributions based on the benchmarks of the American Association for the Advancement of Sciences (AAAS, 1993) for scientific reasoning, Chinn and Malhotra's (2002) theoretical framework for evaluating enquiry tasks and Kuhn's (1962) framework for categorising epistemological stances in argumentation. They found that WoW players demonstrated an impressive variety of higher order scientific reasoning skills in these fora, such as using data and argument, building on others' ideas and using system based reasoning. Players' contributions to discussion boards provided evidence of the higher level evaluative thinking demonstrated in discussion, knowledge sharing and debate and 86% of players' contributions to the fora were examples of this kind.

The following table summarises the types of evaluation methods that can be used and when they can be used for evaluating serious games. It is followed by examples of studies using some of the methods.

Evaluation data can be gathered through mixed methods, mostly combining pre-game and post-game questionnaires of the players, live or video observations, transcripts of after-action reviews and game results. In a few cases, methods are applied more rigorously with in-game knowledge tests or network and communication analyses from logging tools or video observations. Table 1 gives an overview of how to mix the various methods in pre-game, in-game and post-game stages.

Table 1. What to measure, how and when

How		What?	Pre-Game	In Game	Post-Game
Self-reported	Qual.	Personality, player experiences, context, etc.	Interviews, focus group, logbook.	Logbook, interviews or small assignments as part of the game.	Interviews focus group, after-action review.
	Quant.	Social/ demographic, opinions, motivations, attitudes, engagement, game-quality learning, power, influence, reputation, network centrality, learning satisfaction, etc.	Survey, questionnaire, individual or expert panel.	In-game questionnaires	Survey, questionnaire, individual or expert panel
Tested	Qual.	Behaviour, skills, etc.	E.g. actor role-play, case-analysis, assessment, mental models.	Game-based behavioural assessment.	Game-based behavioural assessment.
	Quant.	Values, knowledge, attitudes, skills, personality, power.	Psychometric, socio-metric tests: e.g. personality, leadership, team roles, IQ.	Game-based behavioural performance analysis.	Game-based behavioural performance analysis.
Observed	Qual.	Behavioural performance of student, professionals, player and/or facilitator, others; decisions, strategies, policies, emotions, conflicts, etc.	Participatory observation, ethnographic methods.	Video, audio personal observation, ethnography, Maps, text, figures, drawings, pictures, etc.	Participatory observation, ethnographic methods.
	Quant.	Biophysical–psychological responses, like stress (heart rate, perspiration).	Participant observation, network analysis, Biophysical–psychological observation.	In-game tracking and logging, network analysis, data mining, biometric observation.	In-game log file analysis, network analysis.

In the previous chapter, Mayer et al. discussed the need for proper methods, tools and principles for the evaluation of serious games and game based learning was discussed. Mayer also stated that there is a "lack of comprehensive, multipurpose frameworks for comparative and longitudinal evaluation". While RCT is the gold standard for evaluating educational interventions, very often it cannot be applied in practice due to the difficulties in having randomly selected control groups, and the arising ethical issues and practical concerns. So there is a need for other kinds of evaluation. Furthermore, an upcoming issue is the need for seamless, or "stealth" data-gathering and assessment in SGs (Bellotti et al, 2013a) as well as for performance based evaluation (Bellotti et.al 2013b). These are all activities under development, and thus not yet deployed on a large scale. However, every teacher being interested in using serious games in his /her classes, has, at the end, to deliver a proof of effectiveness and to show how the game supported the learning objectives of the course as well as the individual learning outcomes.

This section has reviewed the different study designs that can and have been applied to evaluating the learning effectiveness of computer games. The next section presents case studies of several of these methods.

CASE STUDIES

The objective in this section is to show different approaches for the evaluation of the learning outcomes of serious games and to discuss the advantages and disadvantages of the methods used. This discussion is based on seven case studies reporting the authors' own experiences in using games in their own courses. In this chapter we present the evaluation of these serious games. (see Table 2

)The serious games we have looked at here are used in different settings in higher education and vocational training. Most of them are facilitated and used in a blended learning approach, only one case study reports on a game which is not facilitated. There is a mixture of individual and team-based games. The topics addressed by the games are varied, ranging from aquaculture to supply chain management.

Table 2. Overview of the case studies and the evaluation methods

Game	Authors	Application Domain	Evaluation Method	Outcomes Measured	Individual/ Team Game
Supply Net Game	Baalsrud-Hauge et al (2007) Delhoum (2009)	Supply Chain and Inventory management	RCT	Marginal inventory costs	Team
Hemocrit (HCT)	Moreno-Ger et al. (2010)	Health	Quasi-experimental; comparison of game group and control group	Rating of difficulty in understanding and performing procedure and in using equipment ; variance in performance	Individual
Beware	Baalsrud Hauge et al. (2008)	Supply Chain Management Risks	Formative; Quasi-experimental: pre, during & post questionnaire	Assessment of knowledge risk management procedures and methods, PKI on users' performance in the game (time, quality, costs, collaboration (no. of interaction with the other players)), scores on final report	Team
SimVenture	Bellotti et al. (2012); Bellotti et al. (2013c)	Enterpreneur-ship Management	Quasi-experimental: pre & post tests	Assessment of knowledge of entrepreneurship-related topics; user acceptance of the serious games and of the overall course based on them	Individual game played by teams
Emergo	Hummel et al. (2011)	Aquaculture	Quasi-experimental: pre & post-tests	Scores on preliminary and final feasibility reports	Individual
Cosiga	Riedel, Pawar, & Barson (2001)	New product development	Survey. In process/ during game tests	Questionnaire on subjective situational awareness administered at regular intervals during game play.	Team
Shortfall	Corriere (2003)	Inventory Management	Surveys: usability survey and player perceptions survey	System Usability Scale (SUS) questionnaire; 10 question post-test survey on player perceptions of game	Team

Supply Net Game-Case Study Using RCT

Description

This case study describes the use of a serious game for system analysis – the Supply Net Game. The game is simulation based and uses the system dynamics methodology (Coyle, 1977). The simulation of a production network was produced using the VensimDSS software (Scholz-Reiter and Delhoum, 2007). Vensim is simulation software usable for modelling dynamic systems (http://vensim.com/vensim-software/) in a realistic way. It is a collaborative game with four participants, each of them being responsible for the inventory and the replenishment in one of four factories, thus the players have to place orders in each simulation period. They also have to control the cash-flow as well as make sure that they do not run out of stock. Each player has an overview of their costs. The aim of the players is the minimisation of the inventory costs. The GUI delivers enough information for taking decisions and comprises: work in progress, back logs, etc. The interface of the game offers the participants feedback so that they can decide on the level of their orders.

Learning Objective

The aim of the game is to support systems thinking in a dynamic environment. The participants are required to learn about inventory management, back logs and the bullwhip effect (Arnold et al, 2002), as well as experience how important communication is. The target for the participants during the game is to minimize their costs, while still being able to deliver. Marginal inventory costs are the key performance measure of the game (Baalsrud Hauge et al., 2007).

Evaluation Method

The game was evaluated using a Randomised Control Trial (RCT) with 106 students, 78 in the experimental group and 28 in the control group at the University of Bremen. There were two groups, one group (the experimental group) only playing the game, and one group (the control group) first getting an introduction to the left-hand elicitation (Delhoum, 2009) method before playing the game. The game included a systems-thinking intervention with a method for mental model elicitation. For the pre and post-tests, we used questionnaires. Ten of the questions were objective, while two of them were judgmental. The same questionnaire was used twice, before and after the main phase of testing to the participants to identify learning effects after running the simulation game for the control group, or after experiencing the left-hand column elicitation method and playing the serious game for the experimental group. Learning was measured by (i) the responses of the students and decision makers to a questionnaire that tests systems-thinking skills and (ii) total inventory costs achieved by a team during the serious game.

Experimental Setup

The game was embedded in a five-step workshop based on Kolb's learning cycle. The participants were divided into two groups. The first group was the control group with 28 participants and met twice. Due to organisational constraints the experimental group had 78 participants. The experimental group was also introduced to elicitation and mental models before they played the game. On an organisational level, two principal characteristics were retained. First, the distribution of the students' pool to the teams was random in the first round. Second, the same teams were built and maintained in both rounds whether this was for the control or experimental group.

Results

The students had lower costs in the second part of the lab, so it was expected that the level of detail and the complexity of the answers given in the questionnaire should have improved. However this could not be verified since the students in the control group scored equally in the pre-test, while the experimental group answered marginally better in the pre-test than in the post-test.

Evaluation of the Evaluation Method

While this study used a RCT, there were pragmatic difficulties in actually implementing a RCT in a regular course at a university. The curriculum specifies how teaching should be delivered, and there was little room for change or innovation. For example for practical reasons it was necessary to include 78 students in the experimental condition but only 28 in the control condition when ideally there would be equal numbers in each. Secondly, if we could produce the evidence that a specific method (in this case the elicitation method) would bring the student a specific advantage, it would not be ethical to randomly exclude students from the same opportunity. In addition, including a control group increases the workload, and thus it is not always possible when running courses.

Validation of the Learning Goals

The learning goals for the supply net game were to understand how inventory control works in a dynamic environment as well as to get a better understanding of system dynamics. Even though the results showed a decrease in costs in the second round of the game, the results do not show significantly higher achievements on the learning objective when comparing the experimental and control groups. The absence of a significant effect is disappointing but has to be viewed in the context of students appearing to enjoy the game and learning how complex any decision in a dynamic environment is.

EMERGO-A Game on Aquaculture Management Game

Aquaculture deals with the development of flora (plants) and fauna (animals) in water. To assess the influence of the new use on the system and other purposes, professionals working in the domain of water management have to both possess natural science knowledge and have a keen eye for the context of policy-making that is involved. Aquaculture is a relatively new sector. Governmental and licensing institutions still struggle to find their way in dealing with entrepreneurs that want to start new businesses in this sector.

Learning Objective

The serious game on aquaculture is the practical part of the aquaculture course that most students follow during their third year of the Bachelor of Water Management programme at OUNL. The main learning objective is to deal with conflicts and dilemmas and to negotiate. The student is assigned the role of an externally hired project leader and is asked to investigate and draw up a feasibility report on what would be the most suitable location to start a new shellfish production site.

Evaluation Method

We compared the quality of advisory reports that students in the domain of water management had to draw up for an authentic case problem, both before and after collaborating on the problem with (virtual) peer students in the game. Peers studied the case from either an ecological or governance perspective, and during collaboration both perspectives had to be confronted and reflected upon. Twelve water management students of the HZ University of Applied Science in the Netherlands participated in this case study. The average age of the participants was 22 years, with a range from 19-26. Seven were male and five were female.

Experimental Setup

For research purposes, the course tutor allocated one of the two perspectives to each student and they had one month to deliver the final report. Virtual collaboration on average took place after about 75% of the period. The same (real life) tutor collected, scored and compared both the preliminary (before virtual collaboration) and final (after virtual collaboration) reports, in close cooperation with another tutor, using a learning effect correction model. Although we did not explicitly measure the inter-subjective reliability of the correction model, both tutors assessed the reports and agreed upon the scores to be given on the various items of the model. Partial elaborations (preliminary reports) before collaboration were assessed as pre-test results, and integrative elaborations (final reports) after collaboration were assessed as post-test results. Appreciation of the serious game was measured by online questionnaires that students had to fill in at the start and at the end (i. e. after sending in their final reports).

Results

A paired t-test (two-tailed) confirmed that the mean scores following the collaborative intervention ($M = 54.00$, $SD = 6.28$) were significantly higher than the scores before the intervention ($M = 19.92$; $SD = 8.47$), ($t = -14.53$; $p < 0.001$). The most important hypothesis therefore can be confirmed: virtual collaboration indeed improves learning effectiveness. We controlled for the influence of perspective on this learning effect (i. e. on the increase of scores), which appears to be missing ($F(1, 11) = 0.72$, $MSE = 46.67$, $p = 0.42$, $\eta p2 = 0.07$).

While assessing the quality of the reports, tutors observed a number of more qualitative results that also provide evidence for the contribution of collaboration. Increases between preliminary and final reports were to the largest degree attributable by gains in scores on the integration items of the correction model. For instances, an integrated map was distilled from information from both perspectives, information about known cultivation methods (ecological perspectives) was linked to existing legislation (governance perspective), and confrontation of perspectives led to better rethinking the selection of most suitable shellfish species. Overall, it is the opinion of both tutors, that the conclusions could not be reached based on one perspective, or learning trajectory alone.

Evaluation of the Evaluation Method

The evaluation method used in this study was a pre and post-test. There was good agreement between the tutors' assessments, showing the reliability of the scoring method. It was planned to compare these results from a brand new course with the results from the previous ones that might have been working as a control group. The issue with a control group is that this is a brand new course on Aquaculture, so there was an existing course which could have been used for a control group. It was decided that there was no real control group possible, mainly as the only alternative for the game might have been face-to-face (f2f) or virtual working groups with high tutor load. Such working groups were practically not feasible because of tutors' limited availability due to other working obligations. Students were dispersed through the region (Province of Zeeland), which made it practically infeasible for them to work together in f2f working groups, so virtual working groups might have been the best alternative. However, the issue with limited tutor availability would still have been the case and considerable costs for setting up a virtual working groups course environment was beyond project budget for game development and testing.

Validation of the Learning Goals

Results from this case study using the educational (serious) game 'Aquaculture' have shown that scripted collaboration significantly improved the quality of learning output. Furthermore, students indicated that the game helped them gain more insight into the various perspectives that play a part in their professional development. According to the questionnaire results, participants preferred real life collaboration over virtual collaboration, although they see that online education does increase the flexibility of study. It therefore could not be concluded that students prefer these kinds of virtual learning environments over more traditional face-to-face settings of collaboration.

Beware-A Game on Supply Chain Risk Management

This game was developed for use in a blended learning environment as part of a course for masters students at the University of Bremen. It is a multi-user, role based game. It has been in use since 2006, and is continuously improved. It is process driven and comprises two levels. The game is facilitated and played in a distributed environment. The facilitator has a monitoring tool, which allows him/her to monitor the game without taking an active part in the game. It also offers the possibility of actively controlling the game by setting events. The facilitator can also communicate with the players via the chat function; she/he can set events and reset processes.

Learning Objective

The objectives of the Beware game are to increase the understanding and awareness of risks in enterprise networks and to improve the players' skills in risk management in a supply network as well as to apply common risk management methods to gain some experience in a risk free environment. Thus, the knowledge on methods and procedure on risk management was measured. In addition it was assessed well the students were able to apply the methods and to apply the methods. In addition, during the game we measure the interaction among the players, the costs, net –margin, logistics cost, performance, delivery on time etc is measured and compared in each round.

Evaluation Method

In this game two forms of evaluation were used. The first is formative - the facilitator monitors the gaming process, collects information on how the different players are playing and on the communication and collaboration between them. Also a set of indicators is continuously collected. These can also be used by the players to evaluate how they played during the game play. This information is used in the debriefing stage in order to analyse and evaluate what happened in the game and thus to construct new knowledge.

The second part of the evaluation is the use of pre, mid-term and post-game questionnaires completed by the players to find if the players have gained knowledge from playing the game. It is only on reconstructable knowledge, so it does not deliver enough information concerning if the player has improved his/her skills on resilience. The outcome of the evaluations is used for improving the game.

Experimental Setup

The Beware game concept foresees that the teacher can introduce the theory to students in advance. Even though the game is process driven, the levels are scenario based. Normally, the students complete two levels. The playing time is 3.5-4 hours, followed by a debriefing and reflection phase. In order to internalise the knowledge acquired during the class, students meet one week after to explain the tasks and the analysis they need to carry out during the two gaming sessions. The observation of how the other participants solved their tasks

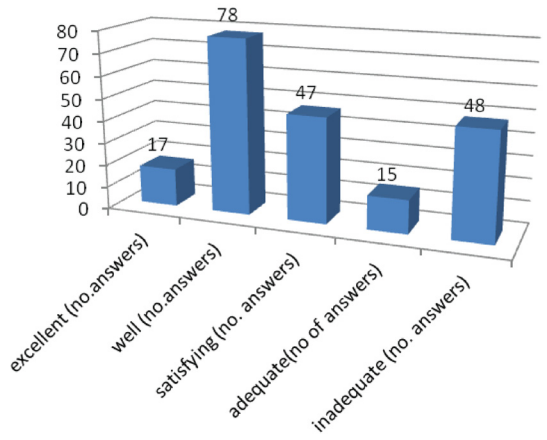
and applied the methods leads to a reflection on the method and thereby to improving the understanding among the participants. Finally, the last step for the participants is to prepare a report in which they reflect on the problems experienced and to assess the strategies they developed at the beginning to reduce the occurring risks (Baalsrud Hauge et al. 2008).

Results

The evaluation of the learning outcome on risk awareness and management showed that the students were able to identify risks, apply risk assessment and management methods, as well as reporting that the game helped them to apply their theoretical knowledge and develop strategies. Applying risk management successfully requires that the participants know the steps of the process. The tests show that the players are able to apply the theory and to use different methods. It also shows that the longer they play, the better they get at identifying and assessing the different types of risks at an early stage. However, if we compare the mid test with the final test, the results show that the level increases more after the game than after the introduction. The participants mentioned two main challenges (provoked by the game); first, they lost the overview and did not manage to deal with the user interface and what was happening. Secondly, they found it difficult to identify hidden risks. We see an example of the outcome of the post test in Figure 1.

The performance in each game is dependent on the players and on the communication level. At the beginning, before the facilitator's tool was in use, it was sometimes the case that the game hardly worked well. The facilitator's tool offers the possibility to track the communication flow against the performance in the game. The communication carried out by using the chat function is stored in the database, and the facilitator can monitor the communication throughout the game

Figure 1. Results of the question: Please list as many main steps in a risk management process as you know and put them in the order you will carry them out.



play. Debriefing is a central part of the two stage game, and time is set aside to analyse the communication and collaboration problems identified during the game in this phase. The trend in these discussions supports the impression of the author/facilitator, that the communication level has an important impact both on the key performance indicators (KPI) as well as the risks the participants need to deal with.

Evaluation of the Evaluation Method

Tracking the communication as well as all the actions taken by the participants is very helpful, but requires a lot of experience of the facilitator. This information also helps in the debriefing sessions.

Using pre, mid and post questionnaires as well as collecting communication data and using in-built performance measures is time consuming. The experiments shows, that the students are motivated and reach the learning goals. However, the evaluation process is complex (especially the part based on interaction and communication), gives good results, but is time consuming and does not support immediate feedback.

Validation of the Learning Goals

The results show that for students without any, or with a little knowledge of risk management, it is important to make their task more visible in the first game level. Furthermore, it was seen that the process of playing one game, debriefing it, and then playing another game level helps to increase the performance on the second game because of the transfer of knowledge from one game to another through debriefing. The participants identified the risks, as well as developed strategies for reducing the collaboration risks to a much higher degree. The continuous evaluation of the learning effect demonstrates that the time required to transfer information into knowledge not only depends on the essential debriefing phase, but also relies on the experience that the participant already has, and that this needs to be taken into consideration at an early stage of the experimental set up, so that the students can be supplied with the necessary information on methods and approaches in advance.

Cosiga–Evaluating a Team-Based Multiplayer Serious Game

Cosiga is a multimedia computer based simulation game of new product development for the education of European engineers, designers, managers and students.

Learning Goals

Cosiga was designed as a complement to engineering and manufacturing courses to give an experience of what the new product development process is like. The game aims to realistically simulate the collaborative and co-operative process of the new product development process inherent in a concurrent engineering approach (Riedel et al, 2001). It is a role playing game with five participants which requires participants to work collaboratively together, using communication tools to specify, design, and produce the final product which is a type of truck. The final product's conformance to specification, development time and costs are used to calculate the team's score.

Evaluation Method

In this study situational awareness (SA) was used to measure the performance of participants during the Cosiga simulation game. The aim was compare the performance of collocated teams and virtual teams. SA is conceptualised as the current knowledge about what is actually happening in a given situation, what it means and what to do about it. It is a mental model of the dynamic context in which one is operating, including its status and dynamics, with which one evaluates current and possible future situations in terms of one's goals, thereby optimising decision-making and performance. "*SA is the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future*" (Endsley, 1995). In our study questionnaires which measure subjective situational awareness were used (while subjective SA has its limitations compared to objective SA, nevertheless if participants do not subjectively feel they have good awareness – whether or not it matches the reality of the situation – they are likely to make mistakes (Endsley 1998).

Experimental Setup

A controlled empirical study was conducted using engineering personnel from the UK aerospace and defence company BAE Systems. Two conditions were set up – a collocated game with all five participants located in the same room and a virtual condition with the five participants located in different rooms, see Table 3.

Typically a game run takes one working day, starting with participants' briefing, game practice session, gaming session, and debriefing. Once the

Table 3. Overview of the Cosiga experimental design

Condition 1 – Virtual Game	**Condition 2 – Collocated Game**
All players in separate rooms with only telephone and text messaging for communication. No face-to-face contact during gaming session.	All players in one room with text messaging for communication. Face-to-face contact at participants' discretion.

five participants had gathered in the room, they were allocated with their role in the game. For the situation awareness evaluation, in-game questionnaires were issued at four ~~regular~~ intervals – starting after one hour of gaming and roughly one hour apart. The data was analysed using analysis of variance – ANOVA. The main aim is to look at the effect of the virtual/collocated condition on the situational awareness of the players over the duration of the game. The data were pooled for all roles to give a sample size of 10 (5 roles and 2 conditions).

Results

For brevity only the results for two items of the situation awareness questionnaire are presented here – as the results for most of the SA items were similar. The first item is Question 1: Would you say that you have observed all events and information that are relevant to your role in the production of the truck? The second item is Question 2: Would you say that you have a good sense of the future course of events and likely outcomes with regard to the production of the truck? (See Figure 2.)

The interaction between condition and time was significant (Two-way interaction condition by time $F(3,15)=3.32$, $p=0.0488$). Analysis suggests that there is a significant difference between collocated and dispersed teams at time intervals 3 and 4. It would appear that the difference between groups became apparent only after a prolonged period of time (2 hours) was spent working on the game. The data suggests that this finding is a result of the virtual team's understanding becoming worse with time and the collocated team's understanding improving with time (see Figure 3.)

Figure 2. SA question 1 Awareness of Events, means over the gaming period

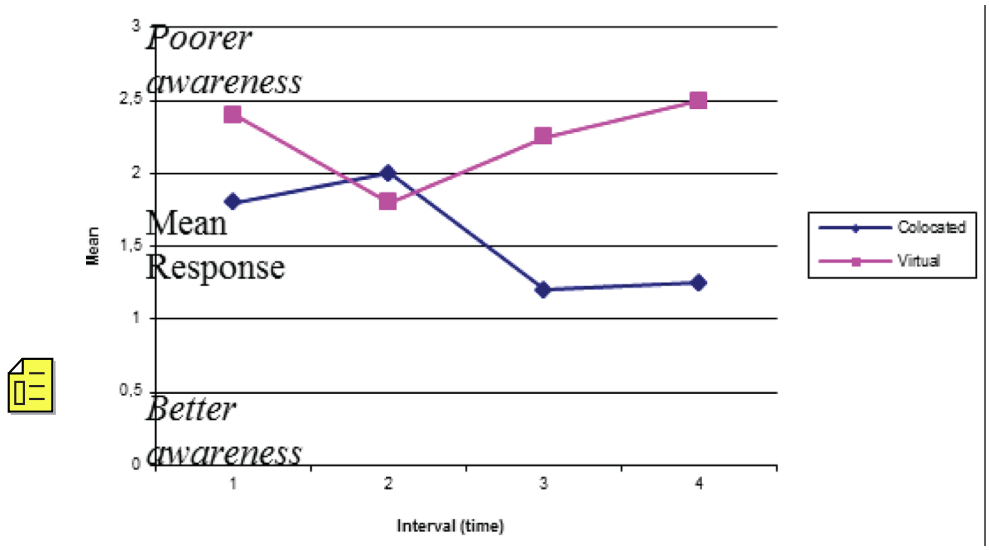
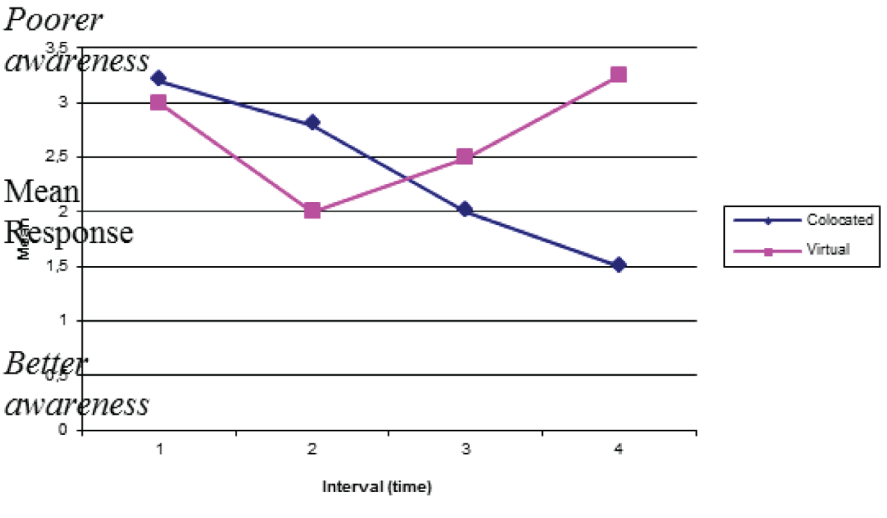


Figure 3. SA question 3 Sense of the future course of events, means over the gaming period



There is again a significant interaction between condition and time (two way interaction $F(3,15)=4.08$, $p=0.0265$). Post hoc tests showing that measures of situation awareness are different between collocated and virtual teams at interval 4 and that situation awareness is better (smaller values) at intervals 3 and 4 from interval 1 for the collocated team. The data suggest that the collocated team was building up an understanding of the future course of events as time elapsed during the game, whereas the virtual team was unable to develop an understanding of the future course of events. In addition the situation awareness in the collocated team was better at the end of the game (interval 4) than in the virtual team.

The results have highlighted statistically significant differences between the two conditions with virtual teams having less situation awareness. There were significant differences between the collocated and virtual teams at the third and fourth measurement points. The collocated team was much more aware of what was going on than the virtual team. In terms of understanding – that is, how easy it was to make sense of the information being provided - the results were not significant, but there was a clear trend indicating that the collocated team found it easier and increasingly easier to make sense of the task as the game progressed. This was not so evident for the virtual team.

Evaluation of the Evaluation Method

The use of situation awareness for evaluating the process and performance of team role play games was successful and produced interesting results. The situation awareness questionnaire used covered only subjective (participant evaluations) SA. Nevertheless, it was short – 8 questions, took only two to three minutes to complete and was easily completed at regular intervals during the game. This enables its use to monitor the progress of players during the game – which can even be shown to them graphically afterwards during the debriefing to get them to reflect upon the result. This evaluation method is suitable to see how players' awareness developed during the game. However, it is not so suitable for identifying what the players had learnt, which is something that needs to be done with objective pre and post-tests.

Validation of the Learning Goals

The evaluation of Cosiga indicated that measuring situation awareness was a valuable method to understand what was happening during the gameplay. It showed that as the game progressed the collocated players improved their understanding of what was going on, whereas the virtual team did not. Analysis of the data from the experiment shows that progression requires the involvement of different disciplines and continuous information sharing, which is the crux of team work.

The Hematocrit (HCT) Game

The HCT game was designed to facilitate the activities of medical students in laboratory sessions (Moreno-Ger et al., 2010). The game simulates the steps of performing an experiment to determine the Hematocrit (HCT) of a blood sample through centrifugation, allowing the students to fail in different steps, providing feedback on what went wrong and allowing the students to repeat the exercise until the highest possible score is obtained. The game simulates the actual workstations at the laboratory, allowing the students to interact with the different objects (Figure 4). The game also exaggerates the negative outcomes, often humorously, balancing interest and providing feedback in a way that students can relate to (Figure 4).

The evaluation process for the game did not focus on knowledge gain, but on other (often overlooked) aspects including student confidence in handling the equipment, anxiety during the laboratory sessions and reliability of the results.

Learning Objective

The exercise of determining Hematocrit in a blood sample is performed in a laboratory, using blood samples from laboratory animals. As such, the students are only allowed to rehearse it once per year during the course and (sometimes) as part of their practical exam. However, the practical exercise can hardly be replaced, even with the most realistic game: handling the actual blood, touching the instruments and even sometimes receiving a spoiled blood sample are parts of the experience that cannot be substituted with current technology. Therefore, the main objective in this game was not to learn how to process the blood sample, but to allow the students to rehearse the procedure before going into the lab, thus taking full advantage of their limited time in the laboratory.

Evaluation Method

Unlike many studies, the evaluation did not focus on knowledge gain, but on whether the game was successful in improving the learning experience

Figure 4. Screenshots from the game showing the work station and a (badly) spoiled centrifuge

in the posterior practical session. The evaluation assessed the following aspects:

- Perceived difficulty of the experiment.
- Performance in handling the equipment.
- Reliability of the results.

After the laboratory session, all students were given access to the game so that they could rehearse before the practical exam.

Experimental Setup

The students were assigned to an Experimental Group (n=66) or a Control Group (n=77) according to their laboratory groups, which are designated alphabetically (and therefore practically random). Students in the Experimental Group were invited to play the game during a class which took place one week before attending the laboratory session, while the control group did not receive any intervention, and went to the laboratory after receiving the usual lecture on dealing with blood samples.

After the laboratory session, the students answered a very brief questionnaire with one 5-point Likert item for each research question:

Q1: Please rate the difficulty you experienced to understand and perform this procedure.
Q2: Please rate the difficulty you experienced to use the required equipment for this procedure.
Q3: Please indicate the HCT value you have obtained.

The answers to questions Q1 & Q2 were compared through an unpaired Students' t-Test as well as Mann-Whitney U tests, to identify differences in perceived difficulty (it was assumed that the results were parametric, but the U test was used for verification). In turn, Q3 was analyzed by comparing the Standard Deviation for each group. Since each group worked with blood from different animals, the results were normalized and an F test was used to statistically compare the variances.

Results

The results for Q1 indicated a mean perceived difficulty of 3.52 (SD 0.28) for the experimental group and 4.39 (SD 0.16) for the control group, a 0.86 difference considered significant after a Mann-Whitney U test ($P = .016$). Regarding the perceived difficulty in using the equipment (Q2) the result for the experimental group was 3.41 (SD 0.40) and 4.02 (SD 0.20) for the control group. The difference was lower (0.31) and not significant ($P = .47$).

As described above, the obtained values for Q3 were normalized and studied in terms of variance, as this would give an estimate of how reliably the students were obtaining the required values. Higher variance would mean less precise results. Comparing the normalized responses of the students, a much lower variability was observed in the results of the experimental group (3.10 vs 26.94 SD for experimental group and control group groups, respectively). An F test showed that variances between the two groups were significantly different ($P(68.19) < .001$; F: 75.25).

Evaluation of the Evaluation Method

The evaluation method compared the perceived difficulty for the experimental group which played the game with that of a control group who did not. The design paid attention to balancing the students and keeping the groups random. The study was limited to a single class, so that all students would have been taking classes from the same instructors, and the assignment to the groups was almost random: for logistic reasons the laboratory groups were used, but the groups were assigned alphabetically and can be considered random in practice.

The evaluation was successful in measuring its initial objectives, even though no gains were

evidenced in one of the research questions. The most difficult item to assess was the reliability of the group, which required normalization processes before the statistical analysis.

Validation of the Learning Goals

The learning goals for the HCT game were not the standard "knowledge gain", but more abstract measures such as confidence, perceived difficulty and skill-based achievements (in the form of lower error rates). The game was successful in allowing the students to focus more on the exercise and less on the steps and proper handling of the materials, resulting in a lower perceived difficulty and, most significantly, in a higher reliability in the results. In turn, the game did not make it easier for the students to use the actual equipment, a conclusion that reinforces the importance of the actual physical practice session.

SimVenture- a Game on Entrepreneurship and Managerial Skills

SimVenture is a single player business simulation game which aims to teach the basis of company management (www.simventure.com). The player is an entrepreneur who manages a small computer assembly and selling company.

Learning Objective

The player's managerial/ entrepreneurial skills are solicited and put to the test in this simulated environment. The simulation is quite detailed and the options and parameters to be managed are numerous. The player is exposed to a number of factors concerning business development in the four functional areas of: sales and marketing, organisation, operations (design and production) and finance.

This complexity allows for the division of responsibilities within a single company team (e.g., director of marketing, director of purchases, financial director, etc.) and also allows experimenting with several different strategies and situations. This is very positive since a player can explore and play for a long time without repeating or getting bored. However, the simulation algorithms are completely opaque, and thus the outcomes of the simulation are not easy to understand and interpret by the players, who have difficulty in learning from their own experience and mistakes. After every simulated month, SimVenture supplies a detailed report, with graphs providing a very detailed break-down of the player's activity and company status (profit and loss, cash-flow, production, employee satisfaction, etc.).

The simulation game provides pre-defined scenarios (e.g., start-up, company growth management, cash-flow crisis), that put the player in different problematic situations, that need to be addressed in different ways. The scenarios vary also in terms of complexity (number of available modules and of parameters modifiable by the player) and difficulty (e. g. initial availability of money).

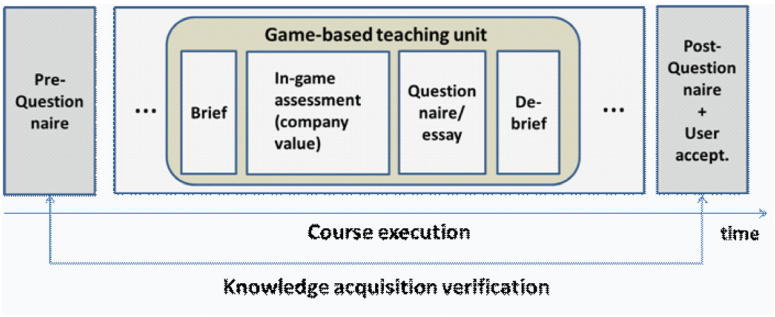
Experimental Set Up

SimVenture has been used in a short course (20 hours) on entrepreneurship at the University of Genoa within the Stimulating Entrepreneurship in Higher Education through Serious Games (eSG) project (www.esg-project.eu). The course was attended by around 40 volunteer BSc (2nd year), MSc (2nd year) and PhD students. The process of collection of requirements for the course, the game selection process and the structure of the course are described in detail in (Bellotti et al., 2012 and in Bellotti et al., 2013c).

Evaluation Method

Here we focus on the user assessment, which involved several aspects, as reported in the following (and synthesised in Figure 5):

Figure 5. Assessment steps in the Entrepreneurship course



- A questionnaire about knowledge on entrepreneurship-related topics was administered before and after the course, in order to assess the students' improvement. Quantitative results are still being processed, but they are positive both for the open and closed questions. The post-questionnaire also involved questions about user acceptance of the serious games as a tool to support learning.
- The students - divided in teams of three participants each - played 6 matches, of increasing levels of simulation complexity and difficulty, also exploiting SimVenture's pre-defined scenarios. Before every game the teacher gave a briefing to introduce the simulation match. After the match the score was assigned by the teacher on the basis of the SimVenture simulation's reports. In particular, the main assessment parameters considered by the teacher in defining the teams' scores were the company's profit and cash-flow levels.
- A questionnaire was also administered after every game session, where the teams could provide free comments and were asked questions about the game's usability and, overall, economic and management topics covered in the simulation. The questionnaire was evaluated with a score which was added (with similar weights) to those coming from the match (previous bullet) and from the debriefing session (next bullet).
- During the de-briefing session held in class after each game competition (which was done at home) each team was asked to discuss their performance and the strategies they employed in the game. Also the de-briefing sessions were assessed with a score assigned to each team by the facilitator.

Evaluation of the Evaluation Method-

This assessment is complex and involves several steps. However, we believe that it was necessary in order to consider all the different aspects. On the one hand, games are very interesting for competitive people. But it is important to combine extrinsic motivation, which should be aroused through games, and intrinsic motivation, which is the fundamental element, in the long-term. On the other hand, the complexity of reality (in particular the world of entrepreneurship and innovation) cannot be fully captured in a game/simulation.

Validation of the Learning Goals

Some teams tended to over-fit the game mechanics (once they had discovered them) and developed some misconceptions, also because of some possible software bugs. Overall, these motivations strengthen the need for the presence of the teacher

as a competent expert that introduces the young generations to the reality. In conclusion games are a new, powerful tool that needs to be carefully studied and employed by teachers to improve the students' understanding and practice.

Shortfall – A Supplement to Teaching Manufacturing Principles

The Shortfall game (Corriere, 2003) was designed to facilitate learning the principles for creating and managing a sustainable manufacturing enterprise. The game centres on raising awareness of the environmental impact in the supply chain as a result of decisions made. The goal of game play is to minimise environmental impacts while maximising profits. The context of Shortfall used in this case study pertains to supplementing taught modules on engineering manufacturing at Heriot-Watt University, across two campuses; UK and Dubai. The aim was to provide students with an alternative approach to classical teaching about the principles of manufacturing enterprises and the technologies employed therein. From the academic perspective, it was about how Shortfall could be used as an abstraction layer to broaden their knowledge and implementation strategy of manufacturing concepts and as a platform for team work.

Learning Objectives and Outcomes

Successful manufacturing requires the integration of the latest manufacturing methodologies, techniques, and innovative technologies. Energy conservation and environmental impact are part and parcel of 21st century manufacturing. Students on completion of the course should have acquired a detailed understanding of the product development process as well as appreciate the impact of sustainability at each stage of the process on the business and organisation with respect to information dependence and manufacturing processes employed.

Evaluation Method

The evaluation was not focused purely on knowledge gain, but on whether the game was successful in improving the learning experience, and if students were able to apply the abstractness provided by the game with that of the taught material and its implementation.

The evaluation follows a usability study and compares the following aspects:

- Perceived usefulness of the game.
- Knowledge of manufacturing practices.
- Tandem use of game with class lectures.
- Reliability of the results.

The focus of this course is the application of knowledge and the development of decision making skills. Teaching is through a combination of core lectures and supported by coursework, in this case a 12-week case study on sustainable manufacturing for which Shortfall was used as a supplement. The assessment is a combination of exam and coursework. After the introductory session to the course and case study, all students were given access to the game.

Experimental Setup

The students formed teams of three and are tasked to play three rounds of the game. A total of 17 groups of three were formed in the UK while a total of 8 groups of three were formed in Dubai. The UK students were largely un-facilitated while the Dubai students were facilitated. This difference was due mainly to the cultural difference and also because the student cohort was smaller. It should be noted here, that unlike standard practice of control groups, each location could be considered its own control group.

A total of 3 rounds were played by each team during the 12-week case study. After each round was played, the team spends two weeks reflecting on how they performed, their strategy, how the

lecture material compares to real world scenarios, and how it contextualises with the game. Each team was expected to keep a log book in the form of a wiki. Reflective questions such as "How can manufacturers use existing technologies in new ways?" and "What innovations are in store for users (in materials, equipment, software, systems, and/or design)?" were asked during the reflection surgery classes to encourage students to probe the philosophical conundrums. The lectures were scheduled such that the knowledge accrued would increase the relevance of game play.

After the 12-week case-study, each team gives a short presentation on how they conducted their game rounds; the time spent in making the choices, the reasoning of the choices made, the consequences of those choices and any other pertinent information with regards the strategy the team chose, etc. It is important that each team member demonstrates their understanding of manufacturing methods and the use of technologies thereof.

Students are then asked to complete a System Usability Scale (SUS) questionnaire, which is a well established tool used in usability engineering (Brooke, 1996). An open question in the exam was used to gauge if the game had influenced their understanding to create and manage a benign manufacturing system/enterprise. The question raised issues on technology causality both on the game side and that of the real world, and how it might influence supply chain strategies and decision making processes. Rather than issuing a questionnaire, the exam probes how elements in the game enable students to identify and link the knowledge and information accrued from the taught material to implement these concepts within the manufacturing domain.

Results

It is important to note that the SUS scores for individual items are not meaningful on their own. The scoring of SUS requires a composite measure of the overall usability. The SUS average score for the UK was 81.6 and for Dubai it was 78.3. This suggests that as a supplement to core material playing Shortfall was deemed by the students to be beneficial.

An additional 10 question post-test survey (Gennett, 2010) was also adopted for this course and used to evaluate student perceptions regarding the effect that the game play had on their knowledge and ability for a variety of areas such as supply chain management and strategies, manufacturing practices, working as a team and individually. The average score for the UK was 65.2 and for Dubai it was 65.8.

Results Analysis

From the perspective of conducting an RCT, the study was limited due to the fact that different staff were involved due to the different campus locations. While every effort was made to ensure consistent teaching and grading of coursework, there will inevitably be some discrepancies. This is due to having different instructors, students' academic level and engagement, and the fact that culturally there is a vastly different approach required for the two student populations. It was perceived that facilitation would be required for the Dubai campus. On the other hand, this would be an advantage in the sense that it would enable the usability of Shortfall to be assessed - the hypothesis being that if the game was designed well, it could be operated with or without facilitation.

The score averages indicate that Shortfall was indeed useful as a supplement to the learning of sustainable manufacturing. The post-test SUS reveals that Shortfall can be conducted either fully facilitated or not at all; i.e. 65.2 and 65.8 respectively. To validate the post test, the exam averages were compared across both campuses. The UK average was 68% (A/B grades) and Dubai average was 66% (A/B grades) for the course. To further establish if Shortfall as a supplement had an effect, results over two preceding years were compared indicating an improvement in top level grades of 12.7%.

Mean scores for the post test questions on enjoyment of the game were 3.21 and 3.34 respectively, suggesting that the students enjoy the game and found it a useful supplement to the taught material. Students have however indicated that they would prefer a game with more realism.

Validation of the Learning Goals

The learning goals for Shortfall were not simply focused on knowledge acquisition. The fact that the game abstracts the values and paradigms of a manufacturing enterprise allowed the students to contextualise with the real world which classical methods could not. Playing the game allowed students to learn and understand new technologies, linking that to new knowledge while expanding on how to best adapt and integrate existing technologies. The team play also enriches communication and collaboration. Shortfall was successful in allowing the students to focus on the application of knowledge and the causality of strategy selection and decision making.

DISCUSSION AND CONCLUSION

In the previous chapter Mayer et al. provided a comprehensive methodology for evaluating Serious Games and Games Based Learning. The current chapter complements Mayer's by providing an overview of the different study designs that can be used to evaluate the learning outcomes of serious games, along with specific examples of these from the literature on serious games. Seven case studies of games were then presented with details of evaluations that have been carried out on these games. The games described in the case studies were all developed for use in higher education to support teaching across a range of disciplines including business, health and engineering manufacturing. The games had varied, and usually complex, learning objectives. The case studies illustrate some of the pragmatic and ethical issues which arise for researchers and teachers in developing and using games in education.

With respect to research design, only one of the case studies, Supply Net, reported using an RCT to evaluate students' performance. Despite optimism that Supply Net would improve systems thinking, the evaluation showed that students in the experimental group performed no better than those in the control group and they actually performed worse in the post-test than the pre-test. While the RCT is the design of choice in evaluating games, this study highlights an interesting dilemma for teachers in utilising the results of the evaluation. How much weight should they attach to the RCT results compared to subjective feedback from several years of using the game for teaching which suggests that students like it!

The other case studies used the less rigorous quasi-experimental design in evaluation. The evaluations of the Emergo, Beware and SimVenture studies all used a pre-test/ post-test design where a performance measure was administered before and after participants played the game. Use of the game improved students' performance in writing feasibility reports on water management (Emergo), knowledge of risk assessment (Beware) and knowledge of entrepreneurship (SimVenure) respectively. While we can conclude that the game did help students to achieve better performance on the specified outcome measures, the absence of a control group in these studies constrains our conclusion about the effectiveness of the game compared to traditional teaching methods.

The HCT study also used a quasi-experimental design. In this case a between groups design was used, with around half of the students assigned to the game group and the other half to a control group which received the traditional teaching only. Care was taken to randomly allocate students to the experimental or control conditions, although, as is frequently the case in educational studies, this was at the level of lab groups rather than individuals. The evaluation of Cosiga also compared two groups. This was not exactly an

RCT, as a pre-test/post-test was not used, but it was rigorous in looking at changes in several variables over time.

Ideally researchers will use objective tests of performance on the specified learning outcomes to assess whether the use of a game leads to better performance than traditional methods of learning. This requires a very clear specification of what the required learning outcome for the game are. Most of the games described in the case studies present players with complex problems which require them to integrate several different dimensions of the problem where opposing views might be presented. Objective measures of performance used in the case studies included knowledge of systems thinking (Supply Net), knowledge of risk management (Beware) and entrepreneurial knowledge (SimVenture). Knowledge was typically assessed by multiple choice questions, although in the case of Emergo was assessed by a written report about an authentic problem case, which was part of the students' coursework. This example illustrates how it is possible to weave the use of a game into the achievement of curricular objectives.

Subjective assessments of performance can be used instead of objective assessments, but these have less validity. HCT game focused on preparing medical students for subsequent work in the laboratory and the evaluation compared students' self-assessments of the perceived difficulty of the task rather than their actual performance on the task. The Cosiga study also used a subjective measure, collecting data on players' situational awareness. This is the players' self-evaluation of their abstract understanding of the current state of the complexity of the problem. While these subjective perceptions provided evidence about players' experiences of the impact of the game, ideally they would be supplemented by objective measures.

Evaluation of the Shortfall game used a survey method looking at players' perceptions of the game and its usability. Surveys are typically used early on in the evaluation to establish that players perceive that the game is fit for purpose. This study had a further objective which was to compare whether the game had a similar impact on students on both campuses, UK and Dubai, of the university. This was important not just to identify possible cultural differences but because different tutors delivered the course on the two campuses. The evaluation confirmed that students on both campuses benefited equally from the game.

The case studies discussed in the current chapter help to illustrate some of practical and ethical issues in evaluating games in the classroom. An important practical consideration is the recognition that it is asking a lot of tutors to take up their limited teaching time to pilot the use of a game which has not yet been shown to produce effective learning outcomes. University lecturers are accountable for the delivery and quality of their modules. They want students to learn and they want to include activities which will help them to learn. Teachers typically have a limited amount of class contact time and, although they might be keen to use the game, they might not be happy about the amount of time it takes away from class time with no guaranteed outcome.

Lack of time was the main difficulty mentioned in the Emergo study as a reason for not setting up a suitable control group. Tutors would have had to mark pre and post intervention assessments for this group. This would have placed an unacceptable load on staff both in terms of their time and the expense. In addition the "traditional teaching" experience might have been of limited value to students.

To provide an objective assessment of the value of a game the evaluation should be carried out by an independent evaluator. However, very often the evaluation of the game is carried out by the people who are developing it.

A number of ethical issues are evident in evaluating games. Games are frequently introduced into modules where students' performance is going to be assessed. Students who are assigned to the control group of an RCT may feel disadvantaged

if other students (in the experimental group) are perceived to be given an advantage, even although playing the game has not yet been shown to provide an advantage in learning. In medical RCTS participants typically do not know whether they are in the experimental group or the control group, whereas with RCTS of games it is clear to both staff and students who is in which group. This may be less obvious if the random assignment is done at the level of classes, although there are still possible confounding variables with this solution, such as different groups having different tutors.

Assigning students randomly to treatment groups, as required in an RCT, is also problematic for the reason that many universities specify that students should be treated equally. A possible solution to the problem of equity is to carry out the evaluation at the beginning of a module and then release the game to everyone later in the semester, to ensure that there are not issues with unfairness before the exams.

In the previous chapter Mayer et al. suggested that the participation of students in evaluating a game should be voluntary. Frequently however, students will feel under pressure to participate in the piloting of the game, if the tutor recommends this as part of their learning experience.

Education policy is following medicine in its demand for evidence based practice. Games developers and researchers are well aware of the need to provide evidence that serious games and games based learning supports students in learning. Although RCTs provide the best evidence about the effectiveness of learning in a game, the use of quasi-experimental evaluation methods can also provide useful information. However as Mayer's evaluation model in the previous chapter shows, comprehensive evaluation of a game goes beyond an RCT based on performance to include very many different aspects of games, players and contexts and changes in these over time. Each of the case studies considers only a small subset of these different facets.

The breadth of factors that need to be taken into account in evaluation raises a further difficult question for teachers and researchers: "Which factors are most important?". We have already considered the case of a game which was not shown to be more effective than traditional teaching but which students liked. What if a game did result in better performance on an RCT but the players didn't actually like it? Should we take on board Mayer's suggestion that hard learning is not all about having fun and insist the students play the game?

There are many areas to develop in our understanding of the evaluation of games and the implications of evaluation for classroom teachers. An interesting area for future development, especially with respect to making evaluation less time consuming, is stealth assessment, assessment of in-game behaviours of players as they play games looking for example at decisions they have made which reflect their level of understanding.

ACKNOWLEDGMENT

The research reported in this paper has been partially supported by the European Union, particularly through the projects: GaLA - The European Network of Excellence on Serious Games (FP7-ICT-2009.4.2-258169) www.galanoe.eu; Cosiga (MM1003). The work in the Stimulating Entrepreneurship in Higher Education through Serious Games (eSG) project (www.esg-project.eu) was partially funded by the EACEA under the Lifelong Learning Programme, contract number: 518742-LLP-1-2011-1-IT-ERASMUS-FEXI.
The study on Shortfall was partially funded by EPSRC-IMRC (EP/F02553X/1) under the theme Serious Games for Computer Aided Engineering

REFERENCES

American Association for the Advancement of Science (AAAS). (1993). *Project 2061: Benchmarks for science literacy*. New York: Oxford University Press..

Arnold, D., Isermann, H., Kuhn, A., & Tempelmeier, H. (2002). *Handbuch der logistik*. Berlin: Springer..

Baalsrud Hauge, J., Delhoum, S., Thoben, K.-D., & Scholz-Reiter, B. (2007). The evaluation of learning the task of inventory control with a learning lab. In *Proceeding Learning with Games 2007*. Learning with Games..

Baalsrud Hauge, J., Duin, H., & Thoben, K.-D. (2008). Increasing the resiliency of global supply network by using games. In *Proceedings of ISL 2008, Centre for Concurrent Enterprise*. Nottingham, UK: Nottingham University Business School.

Baalsrud Hauge, J., & Riedel, J. C. K. H. (2012). Evaluation of simulation games for teaching engineering and manufacturing. In *Proceedings of the 4th International Conference on Games and Virtual Worlds for Serious Applications (VS-GAMES'12),* (vol. 15, pp. 210-220). London: Elsevier Procedia Computer Science.

Beale, I. L., Kato, P. M., Marin-Bowling, V. M., Guthrie, N., & Cole, S. W. (2007). Improvement in cancer-related knowledge following use of a psycho-educational video game for adolescents and young adults with cancer. *The Journal of Adolescent Health*, *41*, 263–270. doi:10.1016/j.jadohealth.2007.04.006 PMID:17707296.

Bellotti, F., Berta, R., De Gloria, A., Lavagnino, E., Antonaci, E., Dagnino, F., & Ott, M. (2013c). A gamified short course for promoting entrepreneurship among ICT engineering students. In *Proceedings of IEEE International Conference on Advanced Learning Technologies (ICALT).* Bejing, China: IEEE.

Bellotti, F., Berta, R., De Gloria, A., Lavagnino, E., Dagnino, F., Ott, M., et al. (2012). Designing a course for stimulating entrepreneurship in higher education through serious games. In *Proceedings of the 4th International Conference on Games and Virtual Worlds for Serious Applications (VS-GAMES'12).* Elsevier Procedia Computer Science.

Bellotti, F., Kapralos, B., Lee, K., & Moreno-Ger, P. (2013a). *User assessment in serious games and technology-enhanced learning*. Hindawi Advances in Human-Computer Interaction. doi:10.1155/2013/120791.

Bellotti, F., Kapralos, B., Lee, K., Moreno-Ger, P., & Berta, R. (2013b). *Assessment in and of serious games: An overview*. Hindawi Advances in Human-Computer Interaction. doi:10.1155/2013/136864.

Brooke, J. (1996). *SUS—A quick and dirty usability scale*. Reading, UK: Redhatch Consulting Ltd.

Cameron, B., & Dwyer, F. (2005). The effect of online gaming, cognition and feedback type in facilitating delayed achievement of different learning objectives. *Journal of Interactive Learning Research*, *16*(3), 243–258.

Chinn, C. A., & Malhotra, B. A. (2002). Epistemologically authentic inquiry in schools: A theoretical framework for evaluating inquiry tasks. *Science Education*, *86*(2), 175–218. doi:10.1002/sce.10001.

Connolly, T. C., Boyle, E. A., Hainey, T., McArthur, E., & Boyle, J. M. (2012). A systematic literature review of empirical evidence on computer games and serious games. *Computers & Education*, *59*, 661–686. doi:10.1016/j.compedu.2012.03.004.

Connolly, T. M., Boyle, E., & Hainey, T. (2007). A survey of students' motivations for playing computer games: A comparative analysis. In *Proceedings of the 1st European Conference on Games-Based Learning* (ECGBL). Paisley, UK: ECGBL.

Corriere, J. D. (2003). *Shortfall: An educational game on environmental issues in supply chain management*. (M.S. Thesis). Northeastern University, Boston, MA.

Coyle, R. G. (1977). *Management system dynamics*. London: John Wiley & Sons..

Delhoum, S. (2009). *Evaluation of the impact of learning labs on inventory control: An experimental approach with a collaborative simulation game of a production network*. Berlin: Gito-Verlag..

Endsley, M. R. (1995). Towards a theory of situation awareness in dynamic systems. *Human Factors*, *37*(1), 32–64. doi:10.1518/001872095779049543.

Endsley, M. R. (1998). A comparative analysis of SAGAT and SART for evaluations of situation awareness. In *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting* (pp. 82–86). Santa Monica, CA: The Human Factors and Ergonomics Society.

Field, A., & Hole, G. (2003). *How to design and report experiments*. London: Sage Publications..

Gennett, Z. (2010). *Shortfall online: The development of an educational computer game for teaching sustainable engineering to millennial generation students*. (MS Thesis). Deptartment of Mechanical and Industrial Engineering, Northeastern University, Boston, MA.

Hsu, C.-L., & Lu, H.-P. (2004). Why do people play on-line games? An extended TAM with social influences and flow experience. *Information & Management*, *41*, 853–868. doi:10.1016/j.im.2003.08.014.

Hummel, H. G. K., van Houcke, J., Nadolski, R. J., van der Hiele, T., Kurvers, H., & Lahr, A. (2011). Scripted collaboration in serious gaming for complex learning: Effects of multiple perspectives when acquiring water management skills. *British Journal of Educational Technology*, *42*(6), 1029–1041. doi:10.1111/j.1467-8535.2010.01122.x.

Karakus, T., Inal, Y., & Cagiltay, K. (2008). A descriptive study of Turkish high school students' game-playing characteristics and their considerations concerning the effects of games. *Computers in Human Behavior*, *24*(6), 2520–2529. doi:10.1016/j.chb.2008.03.011.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. Chicago: University of Chicago Press..

Lindh, J., Hrastinski, S., Bruhn, C., & Mozgira, L. (2008). Computer-based business simulation games as tools for learning: A comparative study of student and teacher perceptions. In *Proceedings of the 2nd European Conference on Games-Based Learning (ECGBL)*. Barcelona, Spain: ECGBL.

Mayer, I. S. (2012). Towards a comprehensive methodology for the research and evaluation of serious games. In *Proceedings VS-Games 2012* (Vol. 15, pp. 1–15). Genoa, Italy: Elsevier Procedia Computer Science. doi:10.1016/j.procs.2012.10.075

Mayer, I. S., Bekebrede, G., Harteveld, C., Warmelink, H. J. G., Zhou, Q., & Lo, J. et al. (2013). The research and evaluation of serious games: Towards a comprehensive methodology. *British Journal of Educational Technology*. doi:10.1111/bjet.12067.

Moreno-Ger, P., Torrente, J., Bustamante, J., Fernández-Galaz, C., Fernández-Manjón, B., & Comas-Rengifo, M. D. (2010). Application of a low-cost web-based simulation to improve students' practical skills in medical education. *International Journal of Medical Informatics*, *79*(6), 459–467. doi:10.1016/j.ijmedinf.2010.01.017 PMID:20347383.

Papastergiou, M. (2009). Digital game-based learning in high school computer science education. *Computers & Education*, *52*(1), 1–12. doi:10.1016/j.compedu.2008.06.004.

Riedel, J. C. K. H., Pawar, K. S., & Barson, R. (2001). Academic & industrial user needs of a concurrent engineering computer simulation game. *Concurrent Engineering: Research & Applications*, *9*(3), 223–237. doi:10.1177/1063293X0100900304.

Scholz-Reiter, B., & Delhoum, S. (2007). The effect of enhanced collaboration on the performance of subjects for the task of inventory control. In K. D. Thoben, J. Baalsrud Hauge, R. Smeds, & J. O. Riis (Eds.), *Multidisciplinary research on new methods for learning and innovation in enterprise networks*. Verlagsgruppe Mainz GmbH Aachen..

Steinkuehler, C., & Duncan, S. (2008). Scientific habits of mind in virtual worlds. *Journal of Science Education and Technology*, *17*, 530–543. doi:10.1007/s10956-008-9120-8.

Weibel, D., Wissmath, B., Habegger, S., Steiner, Y., & Groner, R. (2008). Playing online games against computer- vs. human-controlled opponents: Effects on presence, flow, and enjoyment. *Computers in Human Behavior*, *24*(5), 2274–2291. doi:10.1016/j.chb.2007.11.002.

Woolfson, L. M. (2011). *Educational psychology: The impact of psychological research on education*. London: Prentice Hall, Pearson Education..

KEY TERMS AND DEFINITIONS

Case Study: An investigation of a phenomenon (in this case a game) in its real life context.

Evaluation: Judgement of the value or the strengths and weaknesses of an intervention using established procedures.

Game Based Learning: Learning methods which use games to deliver educational outcomes.

Learning Outcomes: The skills, knowledge or understanding that a student should have as a result of completing specified learning tasks or activities.

Quasi-Experiment: An experimental method where participants are not randomly allocated to conditions or where the experimenter does not have control over the manipulation of the independent variable.

Randomised Control Trial (RCT): An experimental method of evaluating an intervention where participants are randomly allocated to an experimental group or a control group and their performance on the target skill/behaviour before and after the intervention is tested. Ideally pre-testing should confirm no existing difference between the groups, while post-testing should show whether the experimental group performs better than the control group.

Serious Games: Include games for learning and behaviour change and have purposes which go beyond entertainment.