

Evaluating Educational Game Experiences in a Classroom Context Implications for Qualitative Research ⁱ

Steven Malliet

Media, Arts & Design Faculty – University College of Limburg
Centre for Media, Policy and Culture – University of Antwerp

Niels Quinten

Media, Arts & Design Faculty – University College of Limburg

Veerle Van der Sluys

Media, Arts & Design Faculty – University College of Limburg

Abstract

Although quantitative methods for assessing digital game playability have been extensively documented, research on the use of qualitative evaluation methods remains scarce. In this chapter we explore, based upon practical experiences obtained during the development of a number of educational micro-games, the methodology of qualitative playability research. More specific we focus on the use of qualitative evaluation methods in a classroom context. We formulate practical guidelines in a number of areas: duration of evaluation, role of the observers, number of evaluations and observer/participant ratio.

Introduction

In the past years researchers have developed a strong interest in the issue of digital game playability. In a media literacy context, the term has been used as a counterpart to the readability of novels or textbooks (Kücklich, 2003), and as a means to identify a gap between digital natives and digital illiterates (see

Prensky, 2001a). From a cultural perspective, scholars have developed several models to analyze the gameplay mechanisms that constitute the text of a digital game (Aarseth, 2003; Konzack, 2002; Consalvo & Dutton, 2005; Malliet, 2007). Others have argued that, in order to explain why games can be engaging, moving, persuasive or fun, issues of rule-based gameplay should be considered in addition to issues of representation (Frasca, 2003).

The playability of digital games has been an important concern for game developers and game designers as well. Not only has the game industry become a multi-billion dollar business that generates returns comparable to those of other mainstream media industries, but also the development of a contemporary commercial game title involves enormous production costs. Research regarding the elements that can make the difference between commercial success and failure has become invaluable in the game development cycle - to such an extent that, in most contemporary game productions, 8% to 12% of the total budget is spent investigating the experiences of targeted players (Schafer, 2007). Methods of performing player research range from organizing co-design workshops and performing focus group conversations with prospected players, to the use of formal methods to evaluate the flow and the player-friendliness of a game. These latter methods are commonly classified under the denominator of playability research, a field that has evolved from usability research in Human-Computer Interaction.

In this paper, we will focus on a number of methods for playability evaluation that have recently been developed, and we will complement existing research with insights that were gained through our own experience, during the development of a number of educational micro-games. We will argue that, thus far, existing playability methods have focused, either on strictly formal aspects of game content (Desurvivre et al., 2004; Desurvivre & Wiberg, 2009), or on user-experience related aspects such as presence (de Kort et al., 2007). Researchers have also highlighted the mechanisms of interaction between content and player by means of biometric or psychophysiological measures (Nacke, 2009). While the use of quantitative evaluation methods has been well documented in the past years, the use of qualitative game evaluation methods

has received significantly less attention. Nevertheless, in user-centered and design-oriented research, qualitative methods are generally considered a useful, and often necessary addition to the objective methods derived from psychological or physiological research (Laurel, 2004). The main goal of this paper is to fill this gap, and to formulate, based upon practical experience, a number of guidelines for observing and evaluating player-game interactions. More specific, we will focus on the evaluation of player experiences in a classroom context. Qualitative methods have traditionally been considered useful for the investigation of contextual meanings, experiences in a natural context and collective practices (Denzin & Lincoln, 2005; Guba & Lincoln, 2005). These aspects take a central role in the study of educational gaming, given the relevance of issues such as didactic context (Egenfeldt-Nielsen, 2005), group dynamics (Van Eck, 2006) and situated learning (Gee, 2003).

Playability and related issues

Initial research on the method of digital game playability was largely inspired by the evaluation heuristics that have been developed to assess the learnability, efficiency and memorability of interfaces in Human-Computer Interaction research (Schafer, 2007). This has resulted in a number of adaptations of existing usability and likeability questionnaires, applied to the specific needs of digital game design. Within these initial methods, several elements of game content have been identified as constructive to the overall experience of fun. Federoff (2002) made the distinction between play elements (such as AI, storylines, or audiovisual elements), game mechanics (such as involvement, or intuitiveness) and game interface elements (such as feedback and error prevention). Similarly, Järvinen et al. (2002) identified 4 types of playability: structural playability, functional playability, audiovisual playability, and social playability. More recently researchers have extended this framework, and have constructed a set of standardized, formal research tools. Desurvivre et al. (2004) and Desurvivre and Wiberg (2009) constructed the *Heuristics to*

Evaluate Playability (HEP) tool, consisting of 50 items that correspond to a wide range of commonly used design patterns. Schaffer (2007) identified over 50 design guidelines that should enable the game developer to improve the playability of a game during the initial stages of the development cycle.

The studies mentioned above mainly serve to investigate the content and the programmed code of digital games. Only recently the concern has been raised that, in order to sufficiently evaluate the immersive potential of a game, a central role should be given to the assessment of player experiences (Nacke, 2009). Nacke et al. (2009) argue that not only the relationship between game content and game design should be investigated, but also the relationship between game design and user experience. The authors point out that, in a number of related fields (such as interactive art design or multimedia design), experience design has become a common practice, and experience research has become a conventional phase in the development cycle. Following this perspective, researchers have used psychophysiological methods to evaluate the relationship between game content and play experience. Biometric measures have been applied in the evaluation of aspects such as affective level design (Nacke, 2009), the immersive use of game audio (Grimshaw et al., 2008; Nacke, 2009), or the relationship between game controls and flow (Nacke 2009). At the same time, researchers with a background in media research and psychology have developed scales for performing post-hoc tests of experience-related aspects such as social presence (de Kort et al., 2007), involvement (Calleja, 2007), or perceived realism (Ribbens & Malliet, 2009).

Although playability was initially developed as a qualitative method, allowing experts and designers to make temporal and iterative evaluations of games-in-progress, current research strongly focuses on the use of quantitative metrics. A few authors have addressed the need for complementary methods that address the same issues from a qualitative point of view (Poels et al., in press). Methods such as focus group interviews or collaborative design frameworks have been forwarded as valuable additions to the objective methods derived from psychology or psychophysiology, especially when the contextualized and collaborative experience of the prospected players is considered an important

point of attention. Qualitative inquiries into the relationship between player and game have thus far mainly been based upon general methodological literature on focus group interviewing, or on research-by-design.

Setup of this study

In the project described in this paper, extensive use was made of collaborative design methods and qualitative user experience tests, during different phases in the development of a number of educational micro-games. Because these methods were specifically applied to the educational context of testing in a school environment, and because the practical application of these methods depended strongly on the properties of the games that were developed, a number of practical challenges were met that have not been documented in traditional textbooks on qualitative research or on research-by-design. In the remainder of this paper we will give an overview of these challenges. Based upon a critical analysis of the practical solutions that were developed, we will formulate practical guidelines that can be used by researchers who intend to make use of similar methods in future research.

Development context

The games were developed by master students in the context of a game design module at the MAD-Faculty, University College of Limburg. The goal of the module was to translate educational content into game play mechanisms, guided by literature on serious game design and on the learning models associated with digital game play. In collaboration with a local high school, three courses were selected: Biology, Mathematics and French. Within each course, the content of one lesson was selected, and was translated by the students into a digital micro-game. Each team was given creative freedom with respect to the technologies and game play mechanisms that were used. The central focus of the module was on interaction design, and on the adaptation of

traditional textbook material into a meaningful player experience, as described by Salen & Zimmerman (2003).

As is demonstrated in table 1, the approach of the module resulted in the development of different types of games, as well as in the composition of different types of development/evaluation teams. With respect to the communication of educational content, different approaches were taken by the different teams. In two games (G2 and G5) the decision was made to convey information by using a quiz mechanism similar to the traditional assessment methods in the classroom. The developers of one game (G3) made extensive use of narration and dialogue in order to improve the learners' knowledge of French vocabulary and syntax. Finally, in two games (G1 and G4) the educational context served as a background against which an engaging game play experience was elaborated. In these games knowledge of the course content served as an aid to make progress, and knowledge acquisition was implemented as an iterative process of trial and error. As such, with respect to the balance between play elements and educational elements that is propagated in the field of digital learning (Prensky, 2001b), each team took a specific approach and made well-considered decisions.

In addition, different development teams made use of specific interfacing technologies and play mechanisms. In two games (G1 and G5) a physical interface was implemented, resulting in a game play experience that relies heavily upon direct player-game interaction. In one game (G2) the combination of a complex rule system and a team-based setup resulted in a tactical and collaborative game play mechanism. In one game (G3), narration and character design were used as the main vehicle for message construction. Finally, in one game (G4) character design, narration, puzzle solving and strategic game play were integrated into a hybrid online game.

Evaluation context

Player experience evaluations were performed during three phases in the development cycle. First, in the stage of paper prototyping, a co-design session

was organized with students from the cooperating school. These sessions were carried out according to the principles of participatory design described by Muller (2007). Second, a mixture between collaborative design and qualitative playability evaluation was applied to a first electronic prototype of each game. Third, the final version of each game was evaluated. During these final evaluations a post-hoc playability questionnaire was used, based upon the scales developed by Desurvivre et al. (2004) and Desurvivre & Wiberg (2009). The use of this quantitative method was complemented with the use of qualitative interpretive methods. The evaluators made ethnographic notes, observing player responses, player choices, and game progress, according to the principles of ethnographic observation in a school context outlined by Kantor et al. (1981). Finally, post-hoc in-depth interviews (in the case of single-player games) and focus group interviews (in the case of multiplayer games) were performed. These interviews were conducted in a structured way, complementing the quantitative playability evaluation.

The target group were students in the 4th grade of General Secondary Education. The games were tested in three classes (total N=57) during three class moments of one hour. As is demonstrated in table 2, this resulted, with the exception of one game being tested with only three students (G3), in each game being tested with 10-20 persons. The number of evaluations performed and the number of students involved in one evaluation session, depended upon two aspects: the single-player / multi-player characteristic of a game (with multi-player testing involving more than one participant), and the duration of a playing session. The duration of play testing varied from 15 minutes (G3) to 50 minutes (G4).

Although all evaluations were performed following the same principles of ethnographic research and playability evaluation, the choice was made to have different teams emphasize different aspects of the game content during the evaluations. Two teams (G1 and G2) made the choice to emphasize one specific aspect of the game content (i.e. interface, narrations, dialogues, graphic design) during a playing session. The participants were given specific assignments, which resulted in a procedure similar to the task-oriented

approach taken in traditional Human-Computer Interaction research (Dumas & Redish, 1999). With two teams (G3 and G5) the players were encouraged to adopt a personal playing style, and to make their own choices throughout the game – a procedure that is in accordance with the characteristics of digital game play as an experience involving exploration and emergence (Juul, 2005). Finally, in the context of one game (G4), a hybrid approach was taken, where the participants were given specific assignments in addition to being given the freedom to follow their intuition and make personal choices at certain points – a procedure resulting in a significantly longer duration of each evaluation session. In the results section, the benefits and drawbacks of each procedure will be discussed, in relation to the context of play generated in each game.

Results: challenges and responses

Evaluating collective experiences

A first challenge was met during the evaluation of collaborative games that involve multiple players (G2 and G5). Because the evaluations took place during school hours, many participants perceived the sessions as moments of distraction in between working time. One participant noted: 'Cool, we get to play games instead of attending classes!' Many other participants made similar remarks. While the use of the current procedure involves the benefit of evaluating educational games in an educational setting, researchers should watch to it that the evaluation process remains focused upon educational aspects, and that the participants are not distracted from the main focus of the game. A few user experience tests resulted in very low degrees of information. Based upon other experience tests that yielded more significant results, we observed that, in order for a playability test to proceed properly, at least one evaluator is needed to guide each separate participant. In addition, at least two more evaluators are needed in order to make field notes that cover the wide range of activity taking place in a multiplayer game. Specifically, the use of

audio and video recordings of evaluation sessions generated contextual data that allowed the evaluators to frame the user responses more accurately. Finally, during a few experience tests, the teachers of the course were included in the evaluation process. This resulted in a stronger focus of the participants on the main objectives of the games, and in richer degrees of participant responses.

Similarly, the use of physical interfaces (G1 and G5), and the use of competition elements (G1, G2 and G5) had a similar effect on the participants' attitudes towards the testing procedure. Respondents were easily distracted from the educational content of the game, and directed their attention mainly towards elements of play. Consequently, while this enabled us to generate a good understanding of the 'fun' dimension of the game play, often only moderate information was received on the educational content of the games. We concluded, once again based upon experience tests that proceeded more successfully, that the use of scenarios in order to structure the evaluation procedure, may help significantly in overcoming this difficulty. Moreover we observed that the use of multiple evaluators to simultaneously evaluate the players' behavior may result in a stronger research outcome.

Balancing between education and fun

While the *think out loud* method, wherein participants are demanded to comment on their perceptions of game elements, as well as on the strategic decisions made within the game, helped enhance the quality of the field notes that were taken, a number of limitations were encountered. Especially in the context of games that thrive on the creation of a flow experience and on a rapid succession of game stimuli and player responses (most notably, G1), players found it difficult to continue providing verbal cues. At moments, this made it very difficult to assess the players' relationship to the educational contents of the game. This difficulty was partly accounted for in the in-depth interviews that were taken afterward with the respondents.

A comparable challenge was met with games that are narration-driven (G3) and

games that involve puzzle solving (G4). Higher degrees of participant immersion in the games made it difficult to evaluate the balance between play elements and educational elements. Very often, the coders had obtained a good understanding of whether a game was fun or not, but only a moderate understanding of the participants' liking of the educational aspects. A scenario-based or walkthrough-based approach is advised, where the attention of the players is alternately directed towards play elements and instructive elements.

Providing feedback during game play

In usability research the guideline is formulated that the evaluator should interfere as little as possible in the progress of the user through an interactive text (Dumas & Redish, 1999). The main rationale behind this guideline is that the error rate while performing an assignment is a good indicator of the user-friendliness of a procedural system. While educational games can in essence be considered procedural systems (Bogost, 2007), nevertheless a number of difficulties arose from a strict application of this method. In the context of games where repeated play is considered a necessary condition for a learning effect to take place (G1, G2 and G4), the developers may not dispose of the time to let the participants play for several consecutive sessions, nor can they expect the participants to have acquired the level of expert from the first session on. Evaluating these games often proved a time-consuming assignment. Only after a relatively long period of having the participants explore the game rules, interesting insights could be distilled with respect to the game's replayability. In these cases, it proved a useful strategy to have the evaluator participate more actively in the evaluation process, and to provide detailed feedback at moments when the player encountered difficulties proceeding through the game. In addition, the manipulation of the level of difficulty by the developers proved a very useful technique to attune the evaluation test to the level of experience of the participants.

Balancing between structured play and virtual experience

Although thus far we have mainly emphasized the benefits of performing assignment-based user tests, nevertheless in a few cases a *free play*-based approach instigated rich and significant evaluation results. With respect to every game, the evaluation procedures had enabled the developers to discover playing styles and player responses that had not been accounted for on beforehand. Only one developer (G4) had made the choice to adopt a hybrid method, resulting in a longer duration of the user tests, but also in rich, varied and relevant research results. Based upon the outcome of this study, the conclusion is drawn that a well-considered measurement should be made on beforehand of the balance between different evaluation emphases, as well as between the duration of an evaluation session and the number of evaluations that are made.

In table 3, the most important findings are summarized, providing the researcher with guidelines that should be useful in making this measurement. As became apparent throughout the investigation, most of these aspects are dependent of the types of game that are being tested, as well as on the number of evaluators that a team possesses of. Nevertheless, with respect to the number of participants, the participant-evaluator relationship, and the structure of a user test, a number of general conclusions can be drawn.

Conclusions

While quantitative methods have recently proven useful in order to analyze player responses to digital game content, qualitative methods may serve as an important addition and complement to the heuristic and psychophysiological measures that have become extensively used and documented. Especially in the context of educational games, where aspects such as collaborative work, contextual learning and motivation have been strongly emphasized, researchers may benefit from the development of a qualitative approach on playability research. With this paper we wish to draw attention to the possibilities and

challenges included in performing qualitative playability evaluations. Whereas the literature review exposed a number of benefits of this approach, in the results section we highlighted a number of challenges and proposed solutions, based upon a systematic analysis of the evaluation procedures that were applied in the development of five educational micro-games. Although some of the proposed solutions are related to the properties of specific games that have been tested, the analysis enabled us to formulate a number of general guidelines with respect to the methodology of qualitative playability research. More specific, conclusions are drawn in five areas: observer/participant ratio; role assigned to the evaluators; duration of the evaluation sessions; number of tests; and type of evaluation.

Observer / participant ratio. In the context of all games that were developed and tested, we observed that at least one observer/evaluator is needed for every participant in the game play experience. When multiplayer games are tested, using a observer/participant ratio that is larger than 1:1 is highly recommended, because making field notes may become a highly complex assignment, and because specific characteristics of the classroom context demand an enhanced coordinating effort of the researchers.

Evaluator roles. Based upon this study the guideline is formulated that different researcher roles should be assigned to the members of an evaluation team. Taking field notes may become a challenging assignment, given the limitations of the 'think out loud' method in the context of educational game testing. We advise using audio and/or video recordings in order to assess player responses that are overlooked during the playing session. In addition, evaluating games in a classroom context may result in a reduced focus with the participants, and as such in an enhanced need for structuring a user test.

Use of scenarios / cognitive walkthroughs. Although the educational potential of a digital game resides partly in properties such as freedom of action or virtual exploration, this study unveiled a number of benefits associated with

performing player evaluations that are assignment-based. Most importantly, in a classroom context the use of scenarios and cognitive walkthroughs may help maintain the focus of a session on the educational aspects of the games being tested.

Duration / Number of tests. Finally, the researcher should aim for a balance between well-timed player tests, and sufficient amounts of player tests being performed, since performing player tests in a classroom context is often restricted by time limitations. Our analysis unveiled that the most significant results were obtained with evaluation sessions that had a duration of 30 minutes or longer. Especially in the case of games that aim for a high replayability, this duration should be considered a strict minimum. On the other hand, unlike is the case with traditional usability procedures, testing an instructive game with only 5 participants did not always prove sufficient. As a consequence of the variety in player activity, performing player evaluations should be considered a costly, but highly useful phase in the development cycle of an educational game.

References

Aarseth, E. (2003). 'Playing Research: Methodological approaches to game analysis.' Paper presented at the 5th Digital Arts & Culture Conference, May 19 - 23, in Melbourne, Australia.

Bogost, I. (2007). *Persuasive Games. The Expressive Power of Videogames.* Cambridge, MA: The MIT Press.

Calleja, G. (2007). 'Digital Game Involvement.' *Games and Culture* 2(3): 236-260.

Consalvo, M. & N. Dutton (2006) 'Game analysis: Developing a

methodological toolkit for the qualitative study of games.' *Game Studies* 6(1). Available at http://www.gamestudies.org/0601/articles/consalvo_dutton [17 August 2010]

de Kort, Y.A.W., W.A. IJsselsteijn, & K. Poels (2007). 'Digital Games as Social Presence Technology: Development of the Social Presence in Gaming Questionnaire.' *Proceedings of the PRESENCE 2007 Conference*, Barcelona, Spain, 195-203.

Denzin, N.K. & Y. Lincoln (2005). 'The Discipline and Practice of Qualitative Research.' In: N.K. Denzin, & Y. Lincoln (eds.), *The Sage Handbook of Qualitative Research: 3rd Edition*, 1-32. London: Sage.

Desurvire, H., M. Caplan, & J. A. Toth (2004). 'Using heuristics to evaluate the playability of games.' *Proceedings of the CHI '04 Conference on Human Factors in Computing Systems*, Vienna, Austria, 1509–1512.

Desurvire, H. & C. Wiberg (2009). 'Game Usability Heuristics (PLAY) For Evaluating and Designing Better Games: The Next Iteration Lecture.' *Proceedings of the 13th International Conference on Human-Computer Interaction*, San Diego, USA, 557 – 566.

Dumas, J.S. & J.C. Redish (1999). *A Practical Guide to Usability Testing*. Exeter, UK: Intellect Books.

Egenfeldt-Nielsen, S. (2005). 'Beyond Edutainment: Exploring the Educational of Computer Games.' Ph.D. diss., IT-University Copenhagen.

Frasca, G. (2003). 'Simulation versus Narrative: Introduction to Ludology.' In: M.J.P. Wolf & B. Perron (Eds.), *The Video Game Theory Reader*, 221-236. London: Routledge.

Federoff, M. (2002). 'Heuristics and Usability Guidelines for the Creation and Evaluation of Fun in Video Games.' Master Thesis, Indiana University.

Gee, J. P. (2003). *What video games have to teach us about learning and literacy*. New York: Palgrave/St. Martin's.

Grimshaw, M., C. A. Lindley, & L. Nacke (2008). 'Sound and Immersion in the First-Person Shooter: Mixed Measurement of the Player's Sonic Experience.' *Proceedings of Audio Mostly 2008: A conference on interaction with sound*, Piteå, Sweden, 9-15.

Guba, E.G. & Y. Lincoln (2005). 'Paradigmatic Controversies, Contradictions, and Emerging Confluences'. In: N.K. Denzin & Y. Lincoln (eds.), *The Sage Handbook of Qualitative Research: 3rd Edition*, 1-32. London: Sage.

Järvinen, A., S. Heliö & F. Mäyrä (2002). 'Communication and Community in Digital Entertainment Services.' Prestudy Research Report, Hypermedia Laboratory Net Series 2. Available at <http://tampub.uta.fi/tup/951-44-5432-4.pdf> [17 August 2010].

Juul, J. (2005). *Half-Real. Video games between rules and fictional worlds*. Cambridge, MA: The MIT Press.

Kantor, K.J., D.R. Kirby & J.P. Goetz (1981). 'Research in Context: Ethnographic Studies in English Education.' *Research in the Teaching of English* 15(4): 293-309.

Koznac, L. (2002). 'Computer game criticism: A Method for computer game analysis.' *Proceedings of the Computer Games and Digital Culture Conference*, Tampere, Finland, 89-100.

Kücklich, J. (2003). 'The playability of texts vs. the readability of games. Towards a holistic theory of fictionality.' In: M. Copier and J. Raessens (eds.), *Level Up: Digital Games Research Conference*. [CD ROM]. Utrecht University: Faculty of Arts.

Kücklich, J. (2004). *Play and Playability as Key Concepts in New Media Studies*. Research report, Dublin City University. Available at <http://www.playability.de/Play.pdf> [17 August 2010].

Laurel, B. (2004). (ed.) *Design Research: Methods and Perspectives*. Cambridge, MA: The MIT Press.

Malliet, S. (2007). 'Adapting the Principles of Ludology to the Method of Video Game Content Analysis.' *Game Studies* 7(1). Available at <http://gamestudies.org/0701/articles/malliet> [17 August 2010].

Muller, M.J. (2007). 'Participatory Design: The Third Space in HCI.' In: J. Jacko and A. Sears (eds.), *Handbook of HCI*, 1051 - 1068. Mahway, NJ, USA: L. Erlbaum Associates.

Nacke, L.E., A. Drachen, K. Kuikkaniemi, J. Niesenhaus, H. J. Korhonen, W. M. van den Hoogen, K. Poels, W. A. IJsselsteijn & Y. A. W. de Kort (2009). 'Playability and Player Experience Research.' Panel Presented at DiGRA 2009: Breaking New Ground: Innovation in Games, Play, Practice and Theory, Sept. 1-4, London UK.

Nacke, L.E. (2009). *Affective Ludology: Scientific Measurement of User Experience in Interactive Entertainment*. Blekinge Institute of Technology: PH.D Dissertation.

Poels, K., W.A. IJsselsteijn, Y.A.W. de Kort, & B. Van Iersel (in press).

Digital Games, the Aftermath. Qualitative insights into Post Game Experiences. In: R. Bernhaupt, R. (Ed.). *Evaluating User Experiences in Games*. Berlin: Springer.

Prensky, M. (2001a). 'Digital natives, digital immigrants.' *On the Horizon* 9(5): 1–2.

Prensky, M. (2001b). *Digital Game-Based Learning*. New York: McGraw-Hill.

Ribbens, W. & S. Malliet (2009). 'Perceived realism in digital games: a quantitative exploration of its structure.' Paper presented at the 59th conference of International Communication Association (ICA), May 21-25, Chicago, USA.

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

Schaffer, N. (2007). 'Heuristics for usability in games.' White Paper, Rensselaer Polytechnic Institute. Available at <http://friendlymedia.sbrl.rpi.edu/heuristics.pdf> [17 August 2010].

Van Eck, R. (2006). 'The effect of contextual pedagogical advisement and competition on middle-school students' attitude toward mathematics and mathematics instruction using a computer-based simulation game.' *Journal of Computers in Mathematics and Science Teaching* 25(2): 165-195.

Table 1 *List of Games that were Developed*

	Educational topic	Game play mechanisms	Nb Developers	Nb players
G1	Biology	Physical interface - driven game. Player-game interactions against background of human skin structure.	5	1
G2	Biology	Strategic quiz game against background of human skin structure. Complex rule structure to instigate tactical game play.	3	4 – 8
G3	French	Narrative game. Learning French through conversations, with multiple possibilities for correcting mistakes and looking up correct answers.	1	1
G4	Mathematics	Hybrid game, relying upon character design and memory-based missions to learn using statistical summary methods.	1	1
G5	Mathematics	Physical interface – driven game. Learning about statistical methods through physical responses to quiz questions.	3	2 – 6

Table 2 *Overview of evaluation procedures*

	Nb Evaluators	Nb players / session	Total Nb of evaluations	Average duration of evaluations	Evaluation type
G1	5	1	15	20 min.	Assignment-based
G2	3	4 – 8	12	25 min.	Assignment-based
G3	1	1	12	15 min.	Free play
G4	1	1	3	50 min.	Hybrid
G5	3	2 – 6	5	30 min.	Free Play

Table 3 Challenges met during game evaluations and methodological suggestions

Challenge	Response
Enhanced group dynamics may result in loss of focus with participants	<ul style="list-style-type: none"> • evaluator/participant ratio should be larger than 1/1 • strict assignment of evaluator roles is advised • scenario-based evaluations work better than free evaluations • involving teachers in evaluation process enhances quality of testing
Use of physical interfacing in combination with competition elements may result in distraction of participants	<ul style="list-style-type: none"> • scenario-based evaluations work better than free evaluations • number of evaluations should be larger than 5 • evaluation methods should be complemented with player interviews afterward or with field notes and video/audio recordings of playing session.
Participants tend to consider game evaluations a moment of distraction in between classes.	<ul style="list-style-type: none"> • scenario-based evaluations work better than free evaluations • specific testing of education-related aspects is advised • quantitative post-hoc questionnaire helps complement qualitative insights
Providing evaluator feedback during evaluations may result in loss of focus with participants	<ul style="list-style-type: none"> • number of evaluations should be larger than 5 • combination of free evaluations and scenario-based evaluations provides significant results
Both free evaluations and scenario-based evaluations result in significant insights	<ul style="list-style-type: none"> • Longer duration of evaluation sessions is advised • Complementing observations with structured post-hoc interviews is advised

i This investigation was performed in the context of the *Gimme, Gimme, Gimme a Game* project, in collaboration with the centre for *Education & ICT (ED+ict)*, University College of Limburg. Part of this research was financially supported by the PWO (*Projectmatig Wetenschappelijk Onderzoek*) program.