Gameplay experience testing with playability and usability surveys – An experimental pilot study

Lennart Nacke
Postdoctoral Research Associate
Department of Computer Science
University of Saskatchewan
Canada
Lennart.Nacke@acm.org

Jonas Schild
Research Assistant
Department of Computer Science and Applied Cognitive Sciences
University of Duisburg-Essen
Germany
jonas.schild@uni-due.de

Joerg Niesenhaus
Research Assistant
Department of Computer Science and Applied Cognitive Sciences
University of Duisburg-Essen
Germany
joerg.niesenhaus@uni-due.de

ABSTRACT

This pilot study investigates an experimental methodology for gathering data to create correlations between experiential factors measured by a gameplay experience questionnaire and player quality measures, such as playing frequency, choice of game, and playing time. The characteristics of two distinct games were examined concerning the aspects of game experience, subjective game quality, and game usability. Interactions within the three aspects were identified. The results suggest that gameplay experience dimensions flow and immersion are similarly motivating in different game genres, which however might not be equally enjoyable. On the one hand, usability ratings may be positively influenced when a game provides immersion and flow or on the other hand, flow and immersion may be negatively influenced by poor usability ratings. These results emphasize the

need for an approach to classify games based on correlation patterns involving game experience, quality, and usability.

Keywords

Playability, gameplay experience, usability, immersion, fun, flow, player.

1. Introduction

Prior studies of digital games have often focused on the negative effects of digital gaming, such as violent content and its impact (Carnagey, Anderson, & Bushman, 2007; Gentile & Stone, 2005) or addiction to playing (Grüsser, Thalemann, & Griffiths, 2007). However, there has been a recent focus on trying to understand aspects central to gameplay experience (Nacke, 2009b; Poels, de Kort, & IJsselsteijn, 2007). For example, Fernandez (2008) proposed a gameplay experience model, which focused on temporal influences before, during, and after gameplay experience in player-game interaction. Fun is the main component of player experience in this model. It further proposes that game evaluation should concentrate on emotional and cognitive player reactions. IJsselsteijn, Poels, and de Kort theorized that immersion, tension, competence, flow, negative affect, positive affect, and challenge are important elements of gameplay experience and developed a game experience questionnaire (GEQ) to assess these elements (IJsselsteijn, Poels, & de Kort, 2008). It is the goal of this study to investigate correlations between the experiential factors measured by the *GEQ* and player quality measures such as playing frequency, choice of game, and playing time.

This test also requires investigation of whether the underlying usability of a game implementation has an influence on gameplay experience. As Nacke (2009a) notes, usability research on the other hand has taken the ISO 9241-11 standard, defining usability as “effectiveness, efficiency, and satisfaction in a specified context of use” (ISO/IEC 9241-11, 1998). Relating more directly to game developers, Sánchez, Zea, & Gutiérrez (2009) tried to map usability to playability for evaluating UX in games by deconstructing playability and integrating methodological considerations from game development practice.

Following this argumentation, one could see digital games merely as software with the same interaction requirements as other products. However, the interactive experience in games is focused on the progression inside the game system rather than the outcome achieved by playing the game. This is one essential difference in interaction design for digital games and software tools: You play games for the experience itself, thus you have creative freedom in designing the experience itself, while in software tools you are trying to design a pleasant way of achieving a goal efficiently. Game development practice must account for this interaction design aspect. If we see experiential factors like flow, immersion, and enjoyment as constructs of game experience, then these can be facilitated by:

1. Choices in gameplay design – which is essentially the *social*, *psychological* and *cognitive* construction of an enjoyable, interactive, goal-driven experience – or
2. Underlying *technical* prerequisites for this interactive experience to unfold – this relates to the usability of the technology, interface, and interaction devices, which facilitate gameplay.

While in general, both of these factors contribute to overall game usability, here we use the term *game usability* to refer to factor 2: technology, interface, and interaction. We also assume for our pilot study that we can measure game usability of games using a modified system usability scale (*SUS*) (Brooke, 1996) – which will be introduced in the methods section below – as well as measures like playing time, playing frequency, and game quality evaluation. The idea of using playing time and frequency as additional measures for this study came from the discussion of usability metrics in Seffah, Donyaee, Kline, & Padda (2006), where behavior over time was discussed as a usability metric. For games, it is especially interesting to look at frequencies and play-session times, because these metrics could indicate a preference of a certain game, just in the same way as one would prefer software tools that take the least amount of time for achieving certain tasks (Seffah, et al., 2006). However, game preference may also come from aesthetic factors that could enhance how

people perceive the quality of a game, which is why we also chose to measure the quality of gameplay with a separate questionnaire item. In summary, we hope to gain a richer picture of the experiences and preferences evoked by gameplay with this combination of usability metrics, game quality evaluations, and subjective experience assessments. The main contribution of this paper is therefore more in the discussion of its methodological approach than in its initial results presented.

Our approach is the examination of game experience ratings' impact on quality measures such as rated quality, playing time and playing frequency. More specifically, we used the following research questions as a basis for our hypotheses: "Do immersion and flow influence play behavior? What is the effect of usability aspects on these quality measures?" Finally: "How do such aspects differ for games in different genres with different interaction and play styles?" In order to investigate these research questions, we formulate the following experimental hypotheses:

H1: The two games from different genres invoke a different gameplay experience measured by the GEQ.

H2: Game quality correlates with game usability as measured by the SUS.

H3: A high game quality correlates with longer playing time and/or frequency.

2. Material and Methods

2.1 Participants

Participants were 12 right-handed Swedish young adults (between 20 and 33 years old). On average, participants played video games 17 hours and 30 minutes per week (M = 17.5, SD = 21.62). Of the total, one third (N = 4) were female and two thirds (N = 8) were male. On average they have played digital games for 16.18 years (Min = 3, Max = 27, SD = 6.56) in their life. Only one

third of the participants preferred to play games in multiplayer mode (*MMOGs*[1], Local Multiplayer or Clan) in contrast to two thirds who favored single player mode (either alone or with other people in the room). The most popular genres were adventure games (including action-adventures) (33.3%) and role-playing games (33.3%). One participant decided to abort the experiment after the initial session and was excluded from analysis.

2.2 Games Used in the Study

In this experiment, we used two games of similar style and quality, but from different genres and using different interfaces. The games were chosen to be equally pleasurable for advanced and novice players, without factoring in the date the games were developed or released. One game was a remake of a classic commercial game, while the other was a downloadable commercial game:

- Maniac Mansion Deluxe [MMD] (originally released in 1987 by Lucasfilm Games) (LucasFan Games, 2004) – Adventure game, see Figure 1.
- Zuma [Zuma] (PopCap Games, 2003) – an Action Puzzle game, see Figure 2.

Figure 1. Screenshot from the adventure game Maniac Mansion Deluxe [MMD] (LucasFan Games, 2004)

The games were chosen with the two genres being very different from each other so that there should be a clear preference for one of them for each player. Both games are in a similar comic style, non-realistic and simplistic in their interface. While one is more complex in its narrative, the other focuses on the timely solution of puzzle challenges, thus challenging the player in a different way, which could lead to different game experiences. Concerning the usability,

[1] *MMOG* is short for Massively Multiplayer Online Game.

the different forms of interaction in both games lead to the question if the usability criteria of the *SUS* apply well for both games. *Maniac Mansion Deluxe* uses a command system to control the game, which is shown in a separate graphical user interface (GUI) window. This type of GUI becomes outdated as many current games, like *Zuma*, use game elements, graphical and acoustic feedback mechanisms within the game to give feedback to the user. Despite the different interfaces, both games have received excellent reviews indicating an equal quality of the very different gameplay mechanics.

Figure 2. Screenshot from the action puzzle game Zuma [Zuma] (PopCap Games, 2003)

For example, *Maniac Mansion Deluxe* was hailed by *Computer Gaming World* as "a clever and imaginative game" (Ardai, 1988), receiving an average reader rating at IGN.com of 9.6[2], and *Zuma* has a *Metacritic*[3] score of 77% and an average reader score of 9.2.

2.3 Methods

We used several different questionnaires for assessing tendencies of game players to experience flow or immersion during gameplay. For the initial assessment of game experience (when playing the games in the laboratory) we used the *GEQ*. These results were later compared with logged playing time and frequency data and a game quality and game usability assessment completed by each player for each of the games. After the three-week experimental period, an *assessment of game quality* was made using our own questions and questions from the *SUS*. The scale consists of ten items, which are then used to derive a total usability score in a range from 0 to 100. We altered these

[2] IGN.com is an online magazine focused on games, see also http://pc.ign.com/objects/006/006749.html

[3] Metacritic combines review scores from a carefully-screened group of well-respected critics into an overall grade. Its game section is available at http://www.metacritic.com/games

questions to account for game systems instead of general systems, by replacing the word "*system*" with "*game system*" and the word "*use*" with "*play.*" Regarding the assessment of *game quality*, we asked the participants to rate overall quality of each game on a scale from *1* (*worst quality*) to *10* (*excellent quality*). In addition, participants were asked to rank their games (1st and 2nd place) according to their playing preference.

2.4 Procedure

We recruited participants from a gaming background, also from within a Master's program in Digital Game Design. An initial assessment questionnaire was sent out to all participants willing to take part in the experiment. All participants were invited to a game laboratory. After a brief description of the experimental procedure, each participant filled in a compulsory "informed consent" form (with a request not to take part in the experiment if suffering from epileptic seizures or game addiction). Each participant had to complete an initial demographic and psychographic assessment questionnaire prior to the experiment, which was checked for completion. The first phase of the experiment consisted of a game session in the laboratory. The participants played each of the games for 15 minutes with a laptop computer. After each game, participants reported their game experience (using the *GEQ*). The games were played in a counterbalanced order. Then, the functionality of the game logging software was explained to each participant individually and the software and games were handed out via direct transfer, USB-stick or secure download. Each participant was asked to play the two stimulus games, emphasizing that they can freely play the games over a period of three weeks (i.e. freely choose which game to play for any play session, for how long, and how many play sessions they would like to play), with the exception that each of the two games should be tried at least once. It was also explained that it is preferable to play only those games contained in the study during the three weeks. Each participant was thanked for taking part and escorted out of the lab.

In the second phase of the study, the participants played the two games at home. Their game-playing behavior (game selections, playing time) was recorded using simple, custom-made software that is used to launch the games and to record starting times, ending times and durations to a log file. After the three-week experiment period, participants were allowed to keep the games and asked to send back the log file by email.

2.5 Data Reduction

Data entries showing two minutes or less of playing time before a longer playing session were deleted from the logs. This was done after discussing the playing experiences with the participants and after many of them indicated that they had trouble starting the games through the logging software for the first time. It was checked whether the data collected was distributed normally using the *Kolmogorov-Smirnov* test. This was important in order to evaluate whether *non-parametric* methods had to be used for the analysis if the distribution was not normal.

3. RESULTS

3.1 Game experience questionnaire (GEQ) results after initial play session

The data were parametric except for the component *tension* under the *MMD* condition, D (11) = 0.36, $p < .0001$. A dependent *t-test* showed significant differences between *Zuma* and *MMD* for *flow* ($t(10) = -2.23$, $p < .05$), *positive affect* ($t(10) = -2.67$, $p < .05$), *competence* ($t(10) = -3.96$, $p < .01$), and *immersion* ($t(10) = 2.68$, $p < .05$). Differences for *challenge* and *negative affect* were insignificant. Significance for *tension* was checked using a *Wilcoxon signed-rank test* and no significant differences were found between the games.

As expected, both games scored low in *tension* (*MMD*: $M = 2.4$, $SD = 1.06$; Zuma: $M = 2.2$, $SD = 0.89$) and *negative affect* (*MMD*: $M = 1.8$, $SD = 1.04$; Zuma: $M = 1.6$, $SD = 0.80$), suggesting that both titles were equally suited for longer gameplay with this participant group. This could indicate that *MMD* would be more suited to immersive gameplay than *Zuma*, which would be more suited to flow gameplay, *MMD* scored significantly higher on *immersion* ($M = 2.34$, $SD = 0.70$) than *Zuma* ($M = 1.41$, $SD = 0.79$). In addition, *Zuma* scored significantly higher on *flow* ($M = 2.55$, $SD = 0.77$) than *MMD* ($M = 1.71$, $SD = 0.86$). Interestingly, this does not affect the ratings on *challenge*, which have almost equal average values: for *Zuma*, $M = 1.76$, $SD = 0.81$, and for *MMD*, $M = 1.73$, $SD = 1.0$. In addition, we have to mention that there was a strong positive correlation between *flow* and *immersion* scores, $r = .62$, $p < .05$, potentially suggesting a mutual influence of these concepts for the games that were measured in this study. For *Zuma*, we also found that the ratings for *positive affect* ($M = 3.12$, $SD = 0.90$) and *competence* ($M = 2.82$, $SD = 1.01$) were much higher than for *MMD* (PA: $M = 2.06$, $SD = 0.85$; C: $M = 1.18$, $SD = 0.78$). Figure 3 shows a comparison of all game experience scores for each game. In summary, these differences support our first hypothesis (H1) that the two different games lead to distinct game experiences.

Figure 3. GEQ scores (Likert scale 0- 4) for each GEQ dimension of the two games in comparison

3.2 Playing time and frequencies during a three-week period

Both games were played almost equally frequent, *Zuma* being preferred a little (53.6%, $N = 45$) over *MMD* (46.4%, $N = 39$). On average *Zuma* was played 4.09 times ($SD = 0.81$) and *MMD* 3.55 times ($SD = 1.40$). The frequency for *MMD* was nonparametric ($D(11) = 0.37$, $p < .0001$) and there were no significant differences in playing frequencies between *MMD* and *Zuma*. The mean playing time for *Zuma* was 1748.16 seconds (~ 29 minutes),

which was not significantly different[4] from the mean playing time invested in *MMD* of 1905.08 seconds (~ 32 minutes). When looking at correlations for *Zuma*, we did not find any correlation between *flow* and *playing time*, between *immersion* and *playing time*, between *flow* and *playing frequency*, and between *immersion* and *playing frequency* using Pearson's correlation coefficient (r). In contrast to this in *MMD*, we found a strong positive correlation between *flow* and *playing time*[5], $\tau = .58$, p (one-tailed) $< .01$, and between *immersion* and *playing time*, $\tau = .46$, p (one-tailed) $< .05$. Both *flow* and *immersion* were correlated with *playing frequency* (flow: $\tau = .53$, p (one-tailed) $< .05$; immersion: $\tau = .48$, p (one-tailed) $< .05$). This suggests that high flow and immersion ratings correlate with longer gaming sessions in higher frequencies. Similarly to what we observed in the *Zuma* condition, there was a strong positive correlation between *flow* and *immersion* scores in the *MMD* condition, $r = .62$, $p < .05$.

3.3 System (Game) Usability Scale

After the subjects had played the games for three weeks, we used the slightly modified *SUS* to assess the perceived usability of the games. According to Tullis (2008), an average *SUS* score under 60% is relatively poor and one over 80% can be considered as good. The average *SUS* score for *MMD* was quite low ($M = 58.41$, $SD = 24.32$, *Cronbach's* $\alpha = .92$) whereas the *SUS* score for *Zuma* can be considered to be very good ($M = 86.59$, $SD = 9.83$, *Cronbach's* $\alpha = .56$). In addition, we let the participants rate the overall quality of each game on a scale from *1* (bad quality) to *10* (excellent quality). Again, *Zuma* scored relatively high on this scale ($M = 7.32$, $SD = 1.19$) in comparison to *MMD* ($M = 5.09$, $SD = 2.43$). When asked to rank the games in order of preference, 72.7% ranked *Zuma* and only 27.3% ranked *MMD* as their favorite game. When we looked for a correlation between quality rating, *SUS* score, *flow* and

[4] Neither the dependent *t-test* nor the *Wilcoxon signed-rank* test showed significant differences.

[5] Which was non-parametric as indicated by the *Kolmogorov-Smirnov* test.

immersion scores for *Zuma*, we found a positive correlation between *flow* rating and *SUS* score ($r = .67$, $p < .05$) as well as a positive correlation between *immersion* score and quality rating ($r = .72$, $p < .05$), but no significant correlation between quality rating and *SUS* score. Correlations with playing time were also insignificant. When looking at the same correlations for *MMD*, we found a positive correlation between *SUS* score and quality rating ($r = .60$, $p < .05$).

4. DISCUSSION

For one of the games (*Maniac Mansion Deluxe*) a significant positive correlation, between *flow* and *immersion* on one hand and *playing time* and *frequency* on the other, was found. This indicates that, at least for such a narrative-based game, the amount of time invested in the game as well as the frequency of playing it have a positive relationship with the self-reported feelings of flow and immersion. Another factor for this positive relationship regarding time and frequency is the fact that playing an adventure-genre game usually requires more time investment per session than a puzzle game does. The completely different game mechanics of these different genres lead to different playing times. However, this does not explain the correlation with playing frequency. Despite, our results for *Zuma* do *not* replicate this effect which is interesting since *Zuma* was ranked as the more favorite game for many participants, which would led one to believe an increased correlation of flow and playing frequency. A possible reason is that *Zuma*'s game concept of challenge-based action puzzle gameplay might contain other aspects with a strong impact on frequent play besides flow and immersion.

As for our second hypothesis (H2), we did find a significant correlation between game quality rating and usability score for the game *MMD*. Thus, the lower quality ranking of this game might be explained by the usability (*SUS*) score of the game, suggesting that aspects of interface and functionality lead

the game to be perceived as being as of less quality than the *Zuma* game. It could also be that the *SUS* rating indicates a poor gameplay experience although its questions mainly focus on interface and functionality. There is obviously a need for a different measure that does discriminate more clearly between *usability* issues and *gameplay* quality than our modified version of the *SUS* can account for. In addition, no such correlation between game quality rating and usability score exists for the *Zuma* game. For this game, the flow and immersion ratings both had a strong relationship to the reported usability score, suggesting that either when a game is perceived as providing an experience of immersion and flow, this does affect how the game's usability is rated or that a game with poor usability does not support flow and immersion. At least for an action puzzle game this is a very interesting relationship between *game experience* and *usability*. As a result, we draw the conclusion that – while standard usability criteria may apply in studies, which focus on games with a clearly separated GUI (like the command system in *MMD*, which is separated from the game screen) – other games with more hybrid forms of interfaces and feedback mechanisms may be better evaluated by using evaluation methods for playing quality instead of applying technical usability standards.

Although the described difference in quality ratings of the two games was not statistically significant, we do conclude a higher quality rating of *Zuma* based on the explicit preference of nearly 73% of the players. Interestingly, the participants' perceived differences in quality and experience between the two games did not lead to either game being played less often. Thus, we have to reject our last hypothesis (H3) that the preferred game would be played longer and more often. The games may have complementary strengths and weaknesses: *Zuma* shows a clearly higher flow score along with positive affect and competence, while *MMD* scores higher in *sensory* and *imaginative immersion*. This suggests that flow and immersion are two experiential constructs, which may not be equally enjoyable (as indicated by the different quality and positive affect ratings), but nevertheless prove to be equally motivating to play the game. It also leads us to question whether playing

frequency and time are suitable as dependent measures to investigate positive gameplay experience and subjective quality of a game.

In summary, these results show clear differences in how the various aspects of game experience, quality, and usability influence each other for the two different game genres. This raises the question if these characteristics remain constant for games that are each very similar to the two tested games and whether these differences apply to other game genres. With this being the case, the combination of methods in the present study, GEQ, SUS and quality assessment through gameplay logging, could lead to patterns of correlations that help to support game genre classifications. For further investigating this topic, we suggest to create groups of similar games in a larger evaluation setup involving more subjects and a broader variety of genres.

5. CONCLUSION AND FUTURE WORK

In this study, we examined the characteristics of two games in distinct genres concerning the aspects of game experience, subjective quality, and game usability. This combination of methods led to the identification of strong interactions within the three aspects. Gameplay experience dimensions flow and immersion may not be equally enjoyable, but seem to be equally motivating to play a game. On the one hand, usability ratings may be positively affected when a game provides experiences of immersion and flow or on the other hand flow and immersion may be negatively affected by poor usability ratings. These results strongly differ for the two different games, implicating an approach for game genre classification based on correlation patterns involving game experience, quality, and usability. This pilot study provides a basis for more comprehensive future research defining the mutual influence of gameplay experience, quality, and usability. By formalizing gameplay experience in coherence with other aspects of gameplay, we might be able to better evaluate, categorize, and design games in the future.

6. ACKNOWLEDGMENTS

We would like to express our gratitude to our colleagues Sophie Stellmach, who helped to conduct this study, and Craig Lindley, who contributed with helpful discussions. We would also like to thank Ian Livingston and Regan Mandryk from the interaction lab at the University of Saskatchewan. Part of this research was funded by the European Commission under the 6th Framework Programme: New and Emerging Science and Technology (NEST), Project FUGA - The Fun of Gaming: Measuring the Human Experience of Media Enjoyment (Contract: FP6-NEST-28765). We also thank all our participants.

7. REFERENCES

Ardai, C. (1988). The Doctor is in: An Appointment with Terror in Activision's Maniac Mansion. *Computer Gaming World, May,* 40-41.

Brooke, J. (1996). SUS - A quick and dirty usability scale. In P. W. e. a. Jordan (Ed.), *Usability Evaluation in Industry* (pp. 189-194). London: Taylor & Francis.

Carnagey, N. L., Anderson, C. A., & Bushman, B. J. (2007). The effect of video game violence on physiological desensitization to real-life violence. *Journal of Experimental Social Psychology, 43*(3), 489-496.

Fernandez, A. (2008). Fun Experience with Digital Games: A Model Proposition. In O. Leino, H. Wirman & A. Fernandez (Eds.), *Extending Experiences: Structure, Analysis and Design of Computer Game Player Experience* (pp. 181-190). Rovaniemi, Finland: Lapland University Press.

Gentile, D. A., & Stone, W. (2005). Violent video game effects on children and adolescents. *Minerva Pediatrica, 57*(6), 337-358.

Grüsser, S. M., Thalemann, R., & Griffiths, M. D. (2007). Excessive Computer Game Playing: Evidence for Addiction and Aggression? *CyberPsychology & Behavior, 10*(2), 290-292. doi: DOI: 10.1089/cpb.2006.9956

IJsselsteijn, W., Poels, K., & de Kort, Y. A. W. (2008). *The Game Experience Questionnaire: Development of a self-report measure to assess player experiences of digital games.* Eindhoven: TU Eindhoven.

ISO/IEC 9241-11. (1998). Ergonomic requirements for office work with visual display terminals (VDTs) -- Part 11: Guidance on usability. Retrieved February 9, 2008, from http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=16883

Nacke, L. (2009a). *Affective Ludology: Scientific Measurement of User Experience in Interactive Entertainment.* Unpublished Ph.D. Thesis, Blekinge Institute of Technology, Karlskrona.

Nacke, L. (2009b). *From Playability to a Hierarchical Game Usability Model.* Proceedings of the Conference on Future Play @ GDC Canada, Vancouver, BC, Canada. doi: 10.1145/1639601.1639609

Poels, K., de Kort, Y., & IJsselsteijn, W. (2007). *"It is always a lot of fun!": exploring dimensions of digital game experience using focus group methodology.* Proceedings of the 2007 Conference on Future Play, Toronto, Canada. doi: 10.1145/1328202.1328218

Sánchez, J. L. G., Zea, N. P., & Gutiérrez, F. L. (2009, July 19-24). *From Usability to Playability: Introduction to Player-Centred Video Game Development Process.* Proceedings of First International Conference, HCD 2009 (Held as Part of HCI International), San Diego, CA, USA. doi: 10.1007/978-3-642-02806-9_9

Seffah, A., Donyaee, M., Kline, R., & Padda, H. (2006). Usability measurement and metrics: A consolidated model. *Software Quality Journal, 14*(2), 159-178.

Tullis, T., & Albert, B. (2008). *Measuring the user experience: Collecting, Analyzing, and Presenting Usability Metrics*. Burlington, MA, USA: Morgan Kaufmann Publishers.

Figure 4 *Screenshot from the adventure game Maniac Mansion Deluxe [MMD] (LucasFan Games, 2004)*

Figure 5 *Screenshot from the action puzzle game Zuma [Zuma] (PopCap Games, 2003)*

Figure 6 *GEQ scores (Likert scale 0- 4) for each GEQ dimension of the two games in comparison*