Evaluating the Onboarding Phase of Free-to-Play Mobile Games: A Mixed-Method Approach

Falko Weigert Petersen
Department of Communication Aalborg University
Copenhagen, Denmark
falkowp@hotmail.com

Line Ebdrup Thomsen
Department of Communication Aalborg University
Copenhagen, Denmark
let@hum.aau.dk

Pejman Mirza-Babaei
University of Ontario Institute of Technology
Oshawa, Canada
Pejman.Mirza-Babaei@uoit.ca

Anders Drachen
Digital Creativity Labs, University of York
York, United Kingdom
anders.drachen@york.ac.uk

ABSTRACT
The first few minutes of play, commonly referred to as the onboarding phase, of Free-to-Play mobile games typically display a substantial churn rate among new players. It is therefore vital for designers to effectively evaluate this phase to investigate its satisfaction of player expectations. This paper presents a study utilizing a lab-based mixed-methods approach in providing insights for evaluating the user experience of onboarding phases in mobile games. This includes an investigation into the contribution of physiological measures (Heart-Rate Variability and Galvanic Skin Conductance) as well as a range of self-reported proxy measures including: a) stimulated recall, engagement graphs, b) flow state survey and c) post-game experience questionnaire. These techniques were applied across 28 participants using three mobile Free-to-Play titles from different genres. This paper makes two important contributions to the games user research (GUR) domain: 1) evaluates different research techniques (e.g. physiological measures and experience graphs) in the context of mobile games; 2) provides an empirically based recommendation for design elements that result in high arousal.

ACM Classification Keywords
H.5.2 Information Interfaces and Presentation: User interfaces – Evaluation/ methodology

Author Keywords
Mobile Games; GUR; Games User Research; Onboarding phase; User Experience; Free-to-Play; Freemium; F2P.

INTRODUCTION
In recent years, alongside technological advancements, the Freemium business model with the Free-to-Play (F2P, FtP), revenue model has come to dominate the mobile game market [1]. F2P functions by initially offering the game to the player for free, generating revenue via in-game advertisements and purchases [2,3,4]. Although the model is attractive, due to the low financial barrier of entry for new players, developers face many challenges, such as discoverability issues caused by saturation in the F2P market across Android and iOS platforms. Moreover, the first few minutes of play, also referred to as the onboarding phase, is highly critical for F2P games. As they are also characterized by low player retention rates (i.e., most players that leave within a few minutes of play never return) [4,5]. While specific rates vary from game to game, and there are few verified data sources available on the topic, one general estimate suggests that only 28.60% of players return to a game after the first day, and retention rates drop exponentially as a function of the time since the player was last in contact with the game [6]. Many other studies (i.e., Sifa et al. [7] and Hadiji et al. [4]) also reported similar rapid attrition, however, the causes of the low retention rate in F2P mobile games varies based on factors such as competition, traffic sources, and failing to meet user expectations. There are also studies that argued one potential key issue could be poorly designed onboarding experiences (i.e., the starting experience does not foster engagement) [6,1]. For example, Seufert [1] describes the user's first session with a product as being critical in determining the player's lifetime experience with the game in question and it is therefore "worthy of the product team's attention when optimizing the user experience".

In addition to development costs, rises in User Acquisition Cost (UAC - describing the cost of acquiring new players via marketing), market research, and advertising demand that F2P games rapidly engage and retain new players in order to generate profit. Hence, the onboarding phase is arguably more critical in F2P games because players have not yet made any investment in the game, unlike premium games, where the full game is paid for in advance. The F2P revenue model will thus only generate an adequate return on investment (ROI) if users continue to play after the onboarding phases has ended.

Research addressing the onboarding phase of mobile F2P titles, and the development of evaluation frameworks

CHI PLAY '17, October 15–18, 2017, Amsterdam, Netherlands

ACM ISBN 978-1-4503-4898-0/17/10.
DOI: https://doi.org/10.1145/3116595.3125499

suitable for mobile titles, has the potential to address these challenges. There are, however, to the best of our knowledge, limited resources available for the evaluation of mobile games – including the applicability and adaptation of different user research methods in such contexts, work targeting the onboarding phases of games, or design implications of such work. This means that any attempt at investigating the onboarding phase of mobile F2P titles will need to start by considering related work for other, more documented categories of games. A more detailed description of the onboarding phase as a concept and how it is used in this paper will be described in the related work section.

The current approaches for evaluating User Experience (UX) in digital games are largely based on methods that have been adapted from other fields, repurposed in the domain of Game User Research (GUR) [8]. User research has been embraced by the game industry, as it can generate meaningful user insights, which could provide a competitive edge for game companies [9]. However, the success of conducting user research in the context of games is largely dependent on appropriately applying methods which are traditionally reserved for productivity applications to game features [10]. Approaches for evaluating player experience are grounded in a variety of fields and feature-structured research protocols. Evaluation practices vary between developers, and elements like game genre, platform, and target audience are some factors that can affect evaluation techniques throughout the development process [8].
This paper will investigate a range of UX proxy measures for evaluating the onboarding phase of F2P mobile games, including the contributions of physiological measures in evaluating small-factor gaming platforms.

RELATED WORK

Mobile games have been the target of Human-Computer Interaction (HCI) and games research, but mainly in the context of application-specific requirements or the establishment of design guidelines [11], and rarely in terms of evaluating one or more elements or factors impacting UX in mobile games. This is evident in the few references that exist on mobile games and UX. For example, Duh et al. [12] examined the lack of UX research on mobile games, and highlighted a need for research on UX and form factors in mobile games. They investigated three different games played on a mobile platform and discussed control elements such as complexity, motor skill, and interface mapping. Similarly, Engl & Nacke [13] suggested that the context of play was important for the user experience in mobile gaming, and adapted contextual models to suit the mobile situation, noting the need for further research. Moreover, Paul et al. [14] studied social aspects of mobile gaming, which identified socialization as one of the key contextual drivers in playing mobile games, and emphasized the lack of available knowledge on testing methodologies for this format.

A substantial amount of attention in GUR has been directed at exploring the use of physiological (or psycho-physiological) measures to evaluate user experience in games [15,16]. Parallel to similar developments in academia, game studios have started using physiological measures. For example, Chalfoun [17] reported on the use of physiological measures within the User Research Team at Ubisoft. Ambinder [18] reported on similar use of physiological measures – such as Heart Rate (HR), Galvanic Skin Response (GSR), and facial expressions – as possible inputs for games that respond to physiological signals.

Physiological measures require extensive knowledge to implement and can be difficult to analyze, complicating its use in comparison with other user research methods such as observation and interview. Collecting and evaluating physiological data is one out of many methods for evaluating UX, but many studies have reported the fundamental advantage of providing continuous and unconscious recordings of UX [15,19].

Despite this interest and the arguably recent rapid growth of the mobile game industry, the use of physiological measures has not been utilized for evaluation of mobile games; or, such knowledge has not yet been made available in the public domain. Furthermore, there has been little research on using physiological measures to evaluate mobile applications in general, with few exceptions. For instance, Yao et al. [20] reported on a preliminary study examining the possibility of including physiological responses for task performance testing on mobile platforms. They investigated GSR data on failed vs. successful task completions, correlating this response with simple self-reported UX measures. The results indicate that there is potential value in using physiological measures to evaluate UX in mobile applications.

It is thus unclear how well current evaluation techniques translate from large form factor PC and console games, to the smaller form factor screens of mobile phones and tablets. Furthermore, divergent use contexts and different control schemes (often solely touch-based) can also impact the insights generated via different methods, which in turn may require adaptation to suit the mobile product – similar to how user research techniques from productivity applications were modified to suit the evaluation of video games [8,19].

GUR Methods: Towards a Framework for Mobile Games

This section focuses on key related work regarding methodologies for measuring UX in mobile games.

Evaluating F2P mobile games

The existing body of literature available on F2P games is mainly focused on the revenue model [2] and the use of analytics [21], which are applied in the prediction of purchase decisions [7]. The relationships between critical acclaim and commercial success in mobile free-to-play games have also been investigated. These studies often

focus on factors relevant to the success of specific games (e.g. marketing budget or strategy) [22]. Our focus, however, is to establish methods for UX evaluation of onboarding phases. The challenge is that existing literature in this particular area is sparse and largely outdated, as the mobile platform has undergone vast technological advances since its establishment as a gaming platform. For example, Korhonen [23] conducted a study comparing user testing methods (think-aloud, questionnaire, and post-session interview) with expert reviews (using playability heuristics [11]). Although the study was conducted in 2010, the device and the game used are almost incomparable with today's systems. A newer study by Alha et al. [24] also used playability heuristics to evaluate issues in games based on the F2P revenue model. However, the games studied were played through social networks, and are therefore not directly comparable with traditional mobile games. Expert review methods have also been compared to user testing, indicating that expert review is able to discover usability problems, but that user testing is able to reveal very detailed and specific usability problems, which were not adequately identified by expert review [23]. Moreover, expert reviews often suffer from experts' subjective opinions [25].

GUR Measures

When conducting GUR, selecting the right evaluation method depends on several variables (e.g., which kinds of data researchers want to collect). Questionnaires are frequently used for user testing, but often as a supplement to other methods, and can be used before, during, or after a play session. This method focuses on collecting players' self-report impressions and attitudes regarding factors including experience, engagement, and motivation [26,27]. However, the use of self-reporting measures for data collection incurs challenges such as self-selected and time-dependent biases. Hence, using questionnaires or interviews to measure UX elements suffers from the underlying problem of relying on participants' discussion decisions (they may only refer to game events that were meaningful to them) and their ability to recall their experiences of those events [28]. One way to minimize biases in interviews is by using video recordings of the participant's gameplay session to facilitate improved memory, also known as stimulated recall [28]. Another relatively new technique supporting participant memory is the use of experience graphs, where researchers instruct players to draw a curve visualizing their experience and describe it afterwards [29]. Studies show that these graphs accurately reflect overall experience, but are limited in detail [19]. The graphs also appear to address the perception issue by tying the player's impressions of play to their memory of actual in-game events, which will very likely be the issues or successes that participants may communicate to other potential players [19].

Physiological research emphasizes measurement of physiological signals, enabling the identification of associated mental processes [19]. Human experiences are theorized to be highly associated with the emotions encountered during an interaction. The use of physiological measures to identify and understand emotional reactions is common in a range of academic fields [30].

Two of the most widely used methods in academic GUR literature are Galvanic Skin Response (GSR) and Heart Rate Variability (HRV). Both are considered lightweight measures, feature portable equipment, and are seeing use in industrial contexts [18]. GSR is viewed as a reasonably robust measure of arousal if not misused or directly abused [31]. The HRV score is a widely used measure for assessing arousal by looking at the activity of the autonomic nervous system. Photoplethysmogram (PPG) sensors can be used to measure the interbeat interval (IBI) of the heart, by monitoring changes in the blood volume. IBI is than used to calculate the HRV score [32]. Although these measures have been used with both PC/console games and other media and productivity applications, they have not, to the best knowledge of the authors, been previously explored in the context of mobile games.

The Onboarding Phase in F2P Mobile Games

As discussed earlier, the onboarding phase is a critical component in the design of F2P mobile games due to their high attrition rates. Despite this, there are limited resources and research available on this topic.

A related concept in game analytics [33] is First-Time User Experience (FTUE). FTUE refers to the experience of playing a game for the first time. Available knowledge is mainly based on industry research [34,35] and aims to provide design tips on constructing the optimal FTUE, for example, via the application of design principles [36].

Chou [5] presented the Octalysis (within the domain of gamification), a theoretical framework describing eight core player motivations connected to four phases of a player's journey [5]. The first phase is Discovery, which features the player's motivation for wanting to try out the experience. The next phase is Onboarding, where players learn the rules and tools required to play the game [5]. The last two phases are Scaffolding and Endgames, which are concerned with repeated actions taken towards a specific goal, and how current players' motivation can be retained. Hence, the onboarding phase, as defined by Chou [5], begins as soon as the players download the game and ends when the players have mastered the core mechanics and fundamental skills required for the early stages of the game.

On the other hand, Seufert [1] introduces the onboarding funnel, which is an event-based graph of player churn during the initial set of game interactions. He also refers to the "onboarding period" and elaborates that it can vary from being "a very explicit sequence of levels or events" to "a more vague sequence of events employed over multiple game sessions". However, his description of the purpose of onboarding is very similar to Chou's [5], *"the purpose of the onboarding process is to introduce a new user to the product and equip the user with the knowledge necessary to*

interact completely with the product's feature set" [1] (P98).

This paper adapts a definition of the onboarding phase in mobile games based on related research and available industry sources, including reviewing the onboarding of 25 top-ranked F2P mobile games available on Android and iOS App stores, as well as discussions with developers from two game studios (King and Norsfell). Based on these discussions, we tracked the onboarding phase as defined by Seufert [1]. However, we also utilized Chou's [5] definition, as we needed an end time for the onboarding phase, in order to evaluate it effectively. Chou [5] describes the end of the phase as when "users are fully equipped and they are ready to take on the journey on their own". For our purposes, the onboarding phase starts from the first time a player interacts with the game client, and lasts until the completion of a learning phase covering the core game mechanics. Using this definition we examined the onboarding phase of three different games, detailed in the following section. We determined that the onboarding phase of the games used would last around seven minutes (during which players complete the tutorial and/or learn the core mechanics). The designers at King and Norsfell (studios who developed the games in this study) validated this definition as fitting for all three games used in the studies.

EVALUATION

To investigate the contributions of different GUR methods in evaluating the mobile onboarding experience, we designed a mixed-methods experimental setup to study three different F2P titles. The games selected were *Candy Crush Jelly Saga* [37], *WinterForts* [38], and *Pogo Chick* [39]. The measures utilized include two physiological measures, HRV obtained via PPG sensors [32] and GSR, combined with self-report proxy measures related to UX and engagement: questionnaires, stimulated recall interviews, and engagement graphs. GSR and HRV were both chosen based on related literature (as discussed earlier) describing them as lightweight measures with a low level of intrusiveness that can be collected and analyzed without extensive specialized knowledge regarding physiological measures.

Setup

The experimental setup was based on a within-subject design [40] where each participant was exposed to three different mobile game onboarding phases. All test sessions were held in the same location, on weekdays between 10:00 and 19:00, with each test session lasting approximately one hour and 30 minutes. After a short introduction including information about the test session, the participant was asked to sign a consent form informing them about the purpose of the test session, their rights, and how the collected data would be handled and stored.

The Games

The participants played the onboarding phase of three different mobile games, *Candy Crush Jelly Saga*, *WinterForts*, and *Pogo Chick* (see Figure 1). Each game had unique onboarding phases, play styles, and genre features (puzzle, strategy, and arcade). Diverse genres and onboarding styles were chosen to make the sample representative of the variety available in the mobile games market, with the goal of yielding more generalizable results. These titles were also selected based on an ongoing collaboration with the developers. This was an important factor, as the study relied on input from the developers for their intended design.

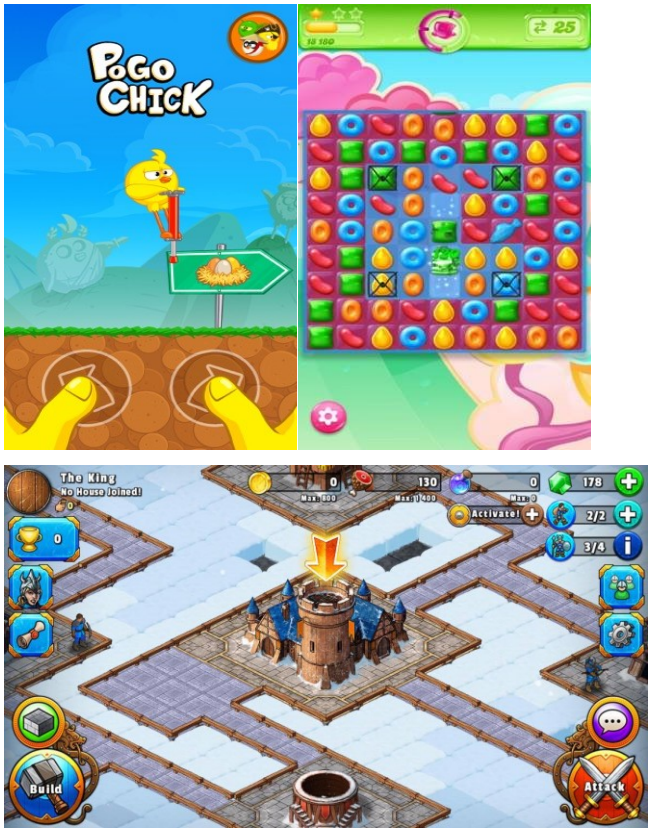

Figure 1: In-game screenshots of the onboarding phases of *Pogo Chick* (Top left), *Candy Crush Jelly Saga* (Top Right), and *WinterForts* (Bottom) [41].

Participants

The participants included 21 females and seven males. The age of the participants ranged from 20-37 years (mean = 25.25; SD = 3.63). The gender inequality in our sample is due to the uneven distribution among people who volunteered to participate in the study. Participants were recruited voluntarily through email and social networks. Demographic data was collected as well as information on prior game experiences. All participants played mobile games at least 1-2 days a week and the majority considered themselves casual players. During the pre-session interview, the participants were asked if they had prior experience with the games in this study. While most participants had played earlier versions of *Candy Crush*, none had played *Candy Crush Jelly Saga*. Additionally,

none of the participants had prior experience with *WinterForts* or *Pogo Chick*.

Apparatus

Participants sat in a comfortable chair while electrodes and the PPG sensor were applied. The participants were asked to rest for three minutes while a baseline for physiological measures was recorded. After the rest period, the participants were given a tablet with a running game. The participant played each game for approximately seven minutes, which represented the onboarding experience, as discussed in the previous section. To reduce the potential of carryover effects [40] affecting the data collection, a counterbalancing approach was utilized. The six different gameplay orders were assigned evenly amongst participants.

Pre-session and physiological measures

Physiological data (GSR and IBI) was recorded during the play sessions. After each session, questionnaires, interviews, and experience graphs were collected. Skin conduction levels were measured using a Bitalino sensor and the included recording and visualization software. Pre-gelled Skintact F-401C silver-silver chloride (Ag-AgCl) electrodes, were attached to the participant's ring and little fingers on the non-dominant hand, to reduce interference during play. The electrodes where disposed after each test session to reduce the potential for electrode polarization [42]. The participants reported feeling no particular hindrance from this attachment method. IBI data was recorded using a Merlin-digital Heart Rate Monitor PRO (PPG sensor) attached to participants' left earlobe. An iPad Mini 3 was used for the test sessions. The iPad was placed in a stable mount, which enabled the participant to comfortably play the three different games while allowing screen recording. The same iPad was used to ensure that every participant had the same game experience and in an attempt to control the extraneous variables that the use of different devices could add to the study. Two video cameras, a GoPro Hero3 and a Panasonic HC-V700 HD Camcorder, were used for this study. The HC-V700 HD was placed behind the participants to record the screen of the tablet. The GoPro Hero3 was placed in front of the test participant in order to record movement and other events that could have introduced bias into the collected data

Post-session and self-reported measures

The post-play sessions consisted of five steps: 1) drawing first experience graph [19]; 2) FSS questionnaire [26]; 3) stimulated recall interviews [43]; 4) drawing second experience graph; 5) PGQ questionnaire [44]. Immediately after playing each onboarding phase, the participants were asked to draw an experience graph and to subsequently explain the graph. This is a method previously used by other researchers [19] and serves as a tool for helping participants visualize and reflect upon their recent experience. The participants were provided with a pen and paper with pre-drawn X and Y axes. The participants were instructed that the x axis represented time, with the y axis symbolizing the level of experience (positive to negative), and directed to draw a graph resembling their game experience. After the participants finished drawing the first graph, they completed the FSS, a multiple choice Likert scale item questionnaire. The questionnaire was printed out and presented to the participants. Following this, participants completed a stimulated recall interview. Stimulated recall can be an alternative to think-aloud techniques, as it is conducted post-play and therefore does not disturb the collection of physiological data. Similar to the think-aloud method, participants were instructed to speak their mind during a video of their gameplay. This allowed for the participant to comment freely while the facilitators are able to ask follow up questions when deeper insight is needed. This method also enabled players to point out exactly where they had an encountered issues, as they might otherwise remember only an emotional impression (e.g., feeling annoyed or confused) without sufficient details of the interaction leading to this impression. After finishing both the FSS questionnaire and stimulated recall interview, the participants were asked to draw a second experience graph to clarify whether they had a different recollection of their experience after watching their play session. Lastly, the participants were asked to answer the PGQ, another multiple choice Likert scale item questionnaire. This approach was followed for each onboarding phase condition. Before the first test session, several pilot tests were conducted to optimize the setup and detect issues related to participant fatigue following guidelines by Bordens & Abbot [40].

DATA ANALYSIS

Data analysis consisted of multiple steps: 1) post-session processing of physiological data; 2) evaluation of the FSS and PGQ surveys; 3) explorative content coding of the post-session interviews; 4) evaluation and comparison of the experience graphs with GSR and HRV; 5) comparative and correlational analysis across the qualitative and quantitative measures. This approach was used for all three evaluated games. For the physiological measures, the recorded GSR data was first visually inspected to check for logging gaps [45]. Logging gaps were identified in the GSR data for four test sessions and were excluded from the analysis, while no logging gaps were identified in the IBI data. The next step in the GSR analysis was a normalization of the data, since skin conductance level (SCL) can vary from participant to participant and absolute GSR values are therefore not comparable between participants. GSR data was normalized using a rescaling technique that was used in similar previous analysis [15]:

$$GSR_{Normalized} = \frac{\left(GSR_{(t)} - GSR_{Min}\right)}{\left(GSR_{Max} - GSR_{Min}\right)}$$

In the GSR normalization above, *GSRMin* and *GSRMax* refer to the minimum and maximum GSR values in a certain time frame. *GSRt* is a GSR data point contained

within the time frame [46]. After normalizing the physiological data, it was divided into smaller time windows that were based on the game design (e.g. levels for *Candy Crush*). This was done in order to identify how participants' physiological responses fluctuated throughout the play session. This furthermore provided context for the data and enabled the creation of composite graphs for both GSR and HRV for all three games. The GSR and HRV graphs were compared to the experience graphs created by the game designers and participants (an example of these graphs can be seen in Figure 2-4). This was done to investigate whether physiological data would mirror the indented design of the developers, which could then help to indicate if the onboarding design is experienced as intended.

The recordings of the test sessions were used for multiple purposes. A meaning condensation was created for the stimulated recall interviews and subsequently compared to the experience graphs drawn during the test session. The meaning condensations were then also coded through an open approach to identify emerging patterns and issues discovered during the test sessions. Recordings were also used for the second part of the physiological data analysis, examining whether peaks in GSR or HRV data was event-related or non-specific. The GSR and HRV data was normalized into seven-second time frames [46]. This allowed for the identification of small game events that caused a fluctuation in the physiological responses across participants. The game events were then identified using the screen recordings for each game and logged to determine whether similar events caused changes in arousal across different participants (see Table 1).

RESULTS

Several findings were observed throughout this study: 1) insights on events that caused arousal in the players during the onboarding phase; 2) Contribution of using physiological measures and other user research methods in a mobile game context; 3) a relationship between the proxy measure of UX (experience graph), and the physiological composite graphs; 4) a set of onboarding heuristics, which is discussed in our previous article [47].

Although physiological measures have previously been used for the evaluation of video games, no study has discussed the contribution of these measures in evaluating an onboarding phase of mobile games. One of our main goals was to investigate if mobile games had the capability to cause changes in the participant's arousal levels. The analysis of the physiological data showed that it was possible to pinpoint specific game events that caused fluctuations in the participant arousal levels. Table 1 shows events across all three games yielding the most notable increases in arousal. In *Candy Crush* Jelly Saga, the level change event caused high arousal among 24 of the participants. For *Winter Forts*, the most notable event was the appearance of the advisor character; for *Pogo Chick*, the most consistently arousing event was player death.

When comparing the physiological data (coded in table 1) to the meaning condensed interview data, it was found that the participants described their game experience on a more general level when interviewed, when compared to the physiological data. During the stimulated recall, and while explaining their experience graphs, the participants described feelings they had during specific moments. During the stimulated recall of *WinterForts*, 19 participants expressed annoyance during battles, where players were forced to wait for several minutes while the game performed all actions automatically, and 24 participants felt frustration after dying several times in *Pogo Chick*. In addition, the interview data revealed game elements that were hard to identify by the means of physiological measures. For example, during the stimulated recall, participants stated that the music added a positive element to the games (*WinterForts* 79%, *Pogo Chick* 63%, and *Candy Crush* Jelly Saga 45%). The high percentage of participants indicating this during *WinterForts*, indicates that the music contributed positively to the participant's game experience. This element would be challenging to identify if only inspecting the physiological measures.

The analysis of the physiological measures also indicates that they are event-specific and more objective than self-reported measures when evaluating mobile games, specifically the onboarding phase of these games. On the other hand, not all game elements that formed the overall user experience could be identified by the means of physiological measures alone, for example, the feeling of success that participants described after playing *Candy Crush Jelly Saga*. Furthermore, during the analysis of the physiological measures, counter-attacking was identified as a game event in *WinterForts* that caused high levels of arousal. During the stimulated recall however, it was found that a feeling of annoyance instigated by a lack of autonomy caused this change. The physiological measures used for this study demonstrated fluctuations in the arousal level of the participants, however, they do not indicate the valence (i.e., positive or negative quality) of these changes, and may be influenced by external disturbances, not related to the test sessions. This made the analysis of the physiological data time-consuming, which the literature review identified as another general drawback of using physiological measures.

Candy Crush Jelly Saga		*WinterForts*		*Pogo Chick*	
Game Events	Count	Game Events	Count	Game Event	Count
Frame Change	24	Meeting advisor	21	Death	18
Considering what move to make	20	Mining	19	Restarting after death	16
Using Special candy	17	Counter attack	15	Challenging map element	14
Level completed	14	Building roads	8	Jump without visible track	9
Chain Reaction	13	Naming fort	8	Close to dying	8
Instruction Tutorial	13	Upgrading castle	8	Exploring menus	7
Level Overview	13	Add	7	Finding balance	3
Positive feedback	13	Building workers	7	Add	2
End Level Explosions	11	Completing quest	7	Selecting new skin	2
Illegal Move	10	Browsing menus	6	Unlocking new skin	2
Level begins	9	Building solider	6	Browsing skins	1
Special Candy Creation	6	Clicking during attack	5		
Fast Move	5	Browsing quests	4		
Notification	5	Attacked by enemy	3		
External events (not game related)	3	Enemy commander	3		
Restarting level (after fail)	2	Game bug	3		
		Browsing Buildings	2		
		Browsing enemy base	1		
		Finding treasure chest	1		

Table 1: Game event identified in the physiological data (seven-second normalization) for the three evaluated games and the number of times events were identified across all participants. Each event was only counted once per participant [41].

The stimulated recall interviews and the experience graphs also generated different levels of insights. While the experience graph is more concerned with general concepts and the overall game experience, the stimulated recall interviews fostered more event-specific descriptions of how emotional responses developed, and thus provided more specific insights. Participants would often pinpoint events drawn on the graph during the interview, indicating that the recall of some of the emotions and events described during interviews may have depended on participants' ability to draw and refer to a graph. When compared with physiological measures, which can locate many events connected to specific design elements, the stimulated interview identifies fewer events, but affords more in-depth descriptions of game experience.

Combining stimulated recall interview and experience graphs with physiological measures was found to be a useful data collection method, as the data was collected post-play and thereby did not add noise to the GSR and HRV data. The stimulated recall also reminded participants of issues they might not have disclosed or remembered during their description of the experience graph.

Additionally, this approach also proved to be useful for the participants to describe playing habits, context, and how they valued playing games in general. The act of playing a game and subsequently being interviewed about their experience appeared to trigger a need for sharing details 'about their general relationship with and use of games. The participants reported: 1) usually playing with sound off to avoid disturbing their surrounding;, 2) that games helped them relax in their stressful everyday life; 3) if they could relate to the theme of the game.

- "*The chick sounds are funny, also the sound that the chick makes when I fall. Although I do not usually have the sound turned on when I play.*" *(P3, Pogo Chick)*
- *"I am thinking constantly and it is just a way to get away from your thoughts." (P18, Candy Crush Jelly Saga)*
- *"It was the same and I was not engaged at all. The medieval theme is not me at all." (P12, WinterForts)*

The questionnaire analysis started by investigating the reliability of the PGQ and FSS answers by calculating Cronbach's Alpha. Both the FSS and PGQ had acceptable Cronbach's alpha scores (.7 and above in all scales for both the PGQ and FSS). While analyzing the PGQ it was found that the participants rated *Candy Crush Jelly Saga* as the most positive, followed by *WinterForts*. *Pogo Chick* was rated highest on the negative game experience scale, in line with the players descriptions of feeling frustrated because of the difficulty of the game.

	Positive	Negative
Candy Crush Jelly Saga	2.27 (SD=1.19)	1.52 (SD=.94)
Pogo Chick	1.74 (SD=.95)	2.04 (SD=1.40)
WinterForts	2.08 (SD=1.40)	1.83 (SD=.99)

Table 2: Average scores for the positive and negative scales of the PGQ [41].

In addition to *Candy Crush Jelly Saga* being rated most positive in the PGQ, it also managed to get the highest Flow score, followed by *WinterForts* with the second highest, and *Pogo Chick* with the lowest FSS score.

Game	Average flow scores
Candy Crush Jelly Saga	3.84 (SD = 1.03)
WinterForts	3.11 (SD = 1.26)
Pogo Chick	3.04 (SD = 1.30)

Table 3: Average FSS score for the three tested games

In addition to coding the physiological data, GSR and HRV were also visualized by composite average graphs that showed the average experience of all participants. These composite average graphs were created for the onboarding phase of all three games and compared to the experience graphs created by the game developers, visualizing the intended experience during the onboarding phases, and the experience graphs created by the participants.

When comparing the designers' intended experience graphs to the composite player experience graphs, it was found that the two graphs for *Candy Crush Jelly Saga* matched well, in the sense that both graphs increased and decreased at the same time. This indicated that the onboarding experience of *Candy Crush Jelly Saga*, in relation to arousal and experience, follows the intended design.

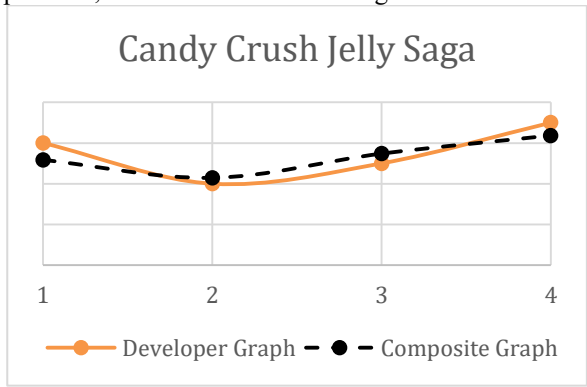

Figure 2: Comparison between designer-intended user experience graph and average player experience graph for *Candy Crush Jelly Saga*. The graphs show the experience for each of the levels that the participated played during the test session, meaning that each number on the X-axis equates to an in-game level [41].

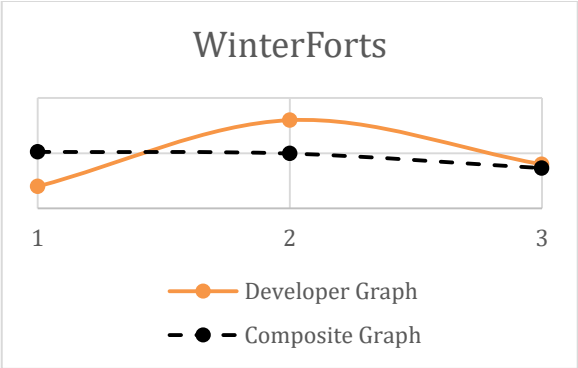

Figure 3: Comparison between designer-intended user experience graph and average player experience graph for *WinterForts*. The first point on the X-axis is the moment where the participants started playing the onboarding phase. The second point is where the tutorial ended, and the third is the end of the onboarding phase

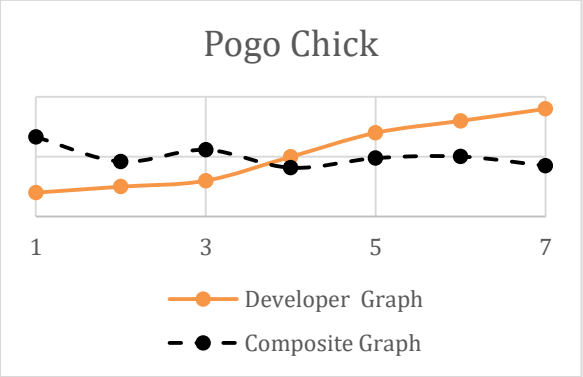

Figure 4: Developer and average player experience graph comparison for *Pogo Chick*. Points on the X-axis delineate minutes of gameplay time. This scaling was chosen due to the loosely structured nature of *Pogo Chick's* onboarding phase [41].

During the interview with the developers of *WinterForts*, it was stated that the experience of the players should increase during the course of the tutorial, as players build their fort and engage in battles with enemies and decrease when the tutorial ended. When inspecting the average graph however, it was visible that the arousal level did not increase during the tutorial phase, but follows the intended design after the tutorial ended. While analyzing the data, collected through stimulated recall and the explanations of the participant created experience graphs, it was found that this lack of engagement was caused by the highly structure tutorial phase, where participants missed the feeling of autonomy. Figure 4 shows the developer and average player experience graphs for *Pogo Chick*. The game was designed to be a challenging action game, and the goal of the designers was that the engagement should only increase during the onboarding phase. The average graph however shows that, even though the arousal level increases at several points, the overall arousal level is decreasing over time.

In addition to comparing player and designer experience graphs, a comparison between the participants' created experience graphs and the average physiological data graphs, was performed. The analysis showed that 43.4% of the participant graphs for *Candy Crush Jelly Saga* and 42% of the graphs for *Pogo Chick* had a strong visual resemblance to the physiological graphs. However, only 17.3% of the participant-created graphs shared characteristics with the average physiological graph for *WinterForts*. While inspecting the self-reported data, no common explanation could be found for this low resemblance compared to the other two games.

DISCUSSION AND CONCLUSIONS

The focus of the work presented here was to evaluate the adaption of physiological, qualitative, and self-report techniques to evaluate UX in the onboarding phase of three F2P mobile games. A number of experiences were drawn from the methodological implementation that bears importance on future work regarding the adoption of physiological measures in mobile game contexts.

Applying Physiological Measures in Mobile Games User Research

The results of this study indicate that physiological measures can be applied to evaluate mobile games, similar to PC and console game contexts [48], as well as for mobile productivity applications, though evidence for this use case is highly limited [20]. While the event-dependent activation of GSR and HRV signals appears similar to those for other game formats, it is worth noting that the arousal-related signals were generally of lower amplitude than reported for other formats [19,48]. The comparison was made by looking at the results from other studies but even though an educated guess could be made on why PC or console horror games tend to generate higher amplitude responses than F2P mobile games, it is still unknown which elements cause this disparity (e.g., screen size, sounds, genre etc.).

The mixed-methods approach used for this study showed how the techniques used could provide insights on specific game events. The engagement graphs visualized the main spikes in participants' remembered experience, and explanation provided information about the reason behind the spikes. This method relied on participant memory and ability to recall specific game events, however, this means that this technique identified arguably the most meaningful events. On the other hand, physiological measures could highlight specific game events that cause fluctuations in player arousal levels, without the need for the participants to remember the specific event, and are thereby more objective then both stimulated recall and the experience graphs [16]. The physiological measures taken alone did not, however, explain the reason behind these fluctuations in the arousal levels. The stimulated recall interviews provided insights about the participant's feelings and reasons behind specific game events, but like the experience graphs, require the participant to remember and interpret their experience. While each of these three methods alone could provide information about the onboarding phase of mobile games, and potentially be used to improve the user experience, the three methods used together can provide a more holistic picture about player's game experience. Moreover, since physiological measures represent unconscious responses, they are less contaminated by variables such as answering style or interpretation of questionnaire items. Physiological measures have previously been perceived as expensive in relation to equipment and training [48], but with today's advancements in technology and the increasing focus on smart health care, equipment has become more advanced, less intrusive, and cheaper. The rise of smart personal health care has, over the years, produced new and more capable devices, such as sports watches with GSR and PPG sensors, while still maintaining relatively high precision and minimizing intrusiveness. It was also found during the literature review, that off-the-shelf equipment enables the collection of physiological data, to be utilized in a much wider range across research fields [30].

Furthermore, during the planning phase of this project, it was decided that physiological data would be analyzed post-session. However, because GSR and HRV were relatively lightweight and easy to analyze, the data could potentially have been preliminarily analyzed while the participants complete the first experience graph and questionnaire, providing a tool for the interviewer to ask more precise probing questions during the stimulated recall interviews. Future work could investigate if this alteration of the test setup, where preliminary analysis of physiological data acts as an interview tool, could improve the setup and provide deeper and more precise information about the player experience during the onboarding phase of mobile games.

The Onboarding Phase

During the initial research for this study, poorly designed onboarding phases, amongst others, were identified as one of the main reasons for the typical high churn rate of F2P mobile games. One of the goals of this study was to create a test setup that could be used to collect FTUE data and provide game designers with a tool to improve this phase. During the analysis of the physiological data, average player arousal graphs were created and compared to the designers' intended user experience graphs, visualizing player experience. This comparison has the potential to help game designers improve the onboarding phase of their games highlighting differences between actual player arousal and the intended design. More research is needed to refine and improve this method in order to maximize the value that these graphs provide for game designers. Future studies could focus on improving the use of physiological measures and experience graphs by combining them with data analytics, which are already widely used by mobile game companies to improve their games.

The analysis of the data collected during the test sessions showed that all three games use highly diverse onboarding styles, and all three games contained some game elements

that were perceived negatively by the participants (e.g., lacking autonomy in *WinterForts*, waiting times during end level explosions in *Candy Crush Jelly Saga*, and the challenging controls in *Pogo Chick*). Our methods highlighted these and several more game elements that could be improved to create an even more enjoyable onboarding experience for players and thereby potentially increase player retention rates. In general, if players are exposed to the same stimuli, comparing results across multiple players is comparatively simple; comparison can be more challenging if many different first-time experiences exist. For example, *Candy Crush Jelly Saga* presents almost identical onboarding experiences across players, whereas *Pogo Chick* opted for more variety.

Even though the techniques used provided deep information about player experience, one of the main drawbacks of the testing setup is that it used a combination of data collection techniques, which makes the test sessions relatively complex, and requires extensive planning. This complexity could potentially be problematic for small game companies without a dedicated UX or GUR department. Future studies could therefore also focus on how this method could be improved to make it more accessible for smaller organizations. This is also the case for other situations where onboarding experiences are particularly important, such as ensuring successful learning in virtual labs and educational applications. A potential reduction in complexity could be achieved by removing the two questionnaires, as the physiological measures together with experience graphs and stimulated recall interview provided a rich data source even without the two questionnaires. Another benefit of removing the two questionnaires from the test setup would be reduced length of the test sessions.

In conclusion, physiological measures can be utilized to evaluate the UX of mobile games, when supported by the qualitative measures, such as stimulated recall interviews and experience graphs, to provide insights into the impact of design on player experience. This finding thus broadly corresponds with conclusions from other game formats [49], although the focus for onboarding phases demands a particular level of detail, as high F2P attrition rates mean that mobile developers must refine action-response cycles on a literal second-by-second basis to optimize UX. The apparent relationship between the physiological measures (arousal), average player experience graphs, and self-reported engagement measures indicates a connection between physiological arousal responses, and qualitative proxy measures of engagement, which is a key UX component for F2P mobile games [33,7]. The full relationship between the measures requires additional research to explore, ideally including additional dimensions of UX, and introducing the valence dimension of physiological measures, even if these measures are not yet economically viable in industrial contexts [50,18].

LIMITATIONS

This study has only been conducted with the onboarding phases of three mobile games, and additional empirical research is thus needed in order to establish the quality of this methodological approach. By repeating the study using other genres of mobile games and even on other elements than the onboarding phase, the quality of the method can be further investigated. As for most laboratory-based user research, the sample size was also too small to draw statistical inference to the parent population of F2P mobile game players worldwide. However, combining small-scale user research with large-scale behavioral analytics [33,7] forms a potential pathway towards solving this problem. Finally, physiological measures are highly dependent on the context in which the data is collected [50,15,28]. Both setup and context will thus have an impact on the interpretation of the data; hence, further studies in different lab contexts would be useful to validate the findings presented here.

ACKNOWLEDGEMENTS

The authors extend their warmest gratitude towards the developers at King and Norsfell for their contribution to this study. We thank Samantha Stahlke for informative discussion and valuable feedback in preparing the final draft of this paper. Furthermore, we would like to thank the participants who volunteered for this study.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731900. Part of this work was conducted in the Digital Creativity Labs (www.digitalcreativity.ac.uk), jointly funded by EPSRC/AHRC/InnovateUK under grant no EP/M023265/1.

REFERENCES

1 Seufert, Eric. *Freemium Economics - Leveraging Analytics and User Segmentation to Drive Revenue*. Morgan Kaufmann, Burlington, Massachusetts, 2014.

2 Alha, Kati, Koskinen, Elina, Paavilainen, Janne, Hamari, Juho, and Kinnunen, Jani. Free-to-play games: Professionals' perspectives. *In Proc DiGRA Nordic 2014* (2014).

3 ESA. *Indystry Facts*. http://essentialfacts.theesa.com/, 2016.

4 Hadiji, F, Sifa, S, Drachen, A, and Thurau, C. Predicting Player Churn in the Wild. *Player Churn in the Wild. In Proceedings of the IEEE Computational Intelligence in Games* (2014).

5 Chou, Yu-Kai. *Actionable Gamification - Beyond Points, Badges, and Leaderboards*. Lean Publishing, 2014.

6 Even, Alon. 2015/10/13/gaming-app-user-retention-only-22-return-after-one-month/ (Oct. 13, 2015).

7 Sifa, R., Hadiji, F., Drachen, A., and Runge, J.

Predicting Purchase Decisions in Mobile free to play Games. In *AAAI Artificial Intelligence in Interactive Digital Entertainment* (Santa Cruz, CA 2015).

8 White, Gareth R and McAllister, Graham. Video Game Development and User Experience. In *Game User Experience*. Springer International Publishing, Switzerland, 2015.

9 Pagulayan, R., Keeker, K., Wixon, D., Romero, R. L., and Fuller, T. User-centered design in games. In *The Human-Computer Interaction Handbook*. L. Erlbaum Associates Inc. , 2003.

10 Smeddinck, Jan David, Markus, Krause, and Kolja, Lubitz. Mobile Game User Research: The World as Your Lab? *Proceedings at CHI'13,* (April 27, 2013).

11 Korhonen, Hannu and Koivisto, Elina M, I. Playability heuristics for mobile games. In *Proceedings of the 8th conference on Human-computer interaction with mobile devices and services* (Helsinki 2006), ACM Digital Library, 9-16.

12 Duh, Henry Been-Lirn, Chen, Vivian Hsueh Hua, and Tan, Chee Boon. Playing Different Games on Different Phones: an empirical study on mobile gaming. In *Proceedings of the 10th international conference on Human computer interaction with mobile devices and services* (Amsterdam 2008), ACM Digital Library, 391-394.

13 Stephan Engl, Lennart E. Nacke. Contextual influences on mobile player experience – A game user experience model. *Entertainment Computing*, 4, 1 (2013), 83-91.

14 Paul, Sheila A, Jensen, Marianne, Wong, Chui Y, and Khong, Chee W. Socializing in mobile gaming. In *Proceedings of the 3rd International Conference on Digital Interactive Media in Entertainment and Arts* (New York 2008), ACM, 2-9.

15 Mandryk, Regan L, Atkins, M Stella, and Inkpen, Kori M. A Continuous and Objective Evaluation of Emotional Experience with Interactive Play Environments. *CHI 2006 Proceedings - Novel Methods: Emotions, Gestures, Events* (April 2006), 1027-1036.

16 Mirza-Babaei, Pejman, Long, Sebastian, Foley, Emma, and McAllister, Graham. Understanding the Contribution of Biometrics to Games User Research. (2011), DiGRA Conference.

17 Chalfoun, Pierre. Biometrics at Ubisoft Montreal (March 15, 2016).

18 Ambinder, M. Biofeedback in Gameplay: How Valve Measues Physiology to Enhance Gaming Experiences. In *Game Developers Conference* (San Francisco, CA 2011).

19 Mirza-Babaei, Pejman. *Biometric Storyboards: A Games User Research Approach for Improving Qualitative Evaluations of Player Experience*. Sussex, 2013.

20 Yao, Lin, Liu, Yanfang, Li, Wen, Zhou, Lei, Ge, Yan, Chai, Jing, and Sun, Xianghong. Using Physiological Measures to Evaluate Use Experience of Mobile Applications. In *Engineering Psychology and Cognitive Ergonomics*. Springer International Publishing, Switzerland, 2014.

21 Drachen, Anders, El-Nasr, Magy Seif, and Canossa, Alessandro. Game Analytics – The Basics. In *Game Analytics - Maximizing the Value of Player Data*. Springer, London, 2013.

22 Alha, Kati, Koskinen, Elina, Paavilainen, Janne, and Hamari, Juho. Critical Acclaim and Commercial Success in Mobile Free-to. play Games. *DiGRA/FDG '16 - Proceedings of the First International Joint Conference of DiGRA and FDG*, 13, 1 (August 2016).

23 Korhonen, Hannu. Comparison of playtesting and expert review methods in mobile game evaluation. *10 Proceedings of the 3rd International Conference on Fun and Games* (2010), 18-27.

24 Alha, K, Paavilainen, J, and Hamari, J. Domain-specific playability problems in social network games. *In proceedings of DiGRA FDG Conference*, Dundee, Scotland (2016).

25 White, Gareth R, Mirza-Babaei, Pejman, McAllister, Graham, and Good, Judith. Weak inter-rater reliability in heuristic evaluation of video games. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems* (2011), ACM.

26 Jackson, Susan A. and Marsh, Herbert W. Development and Validation of a Scale of Measure Optimal Experience: The Flow State Scale. *Journal of sport & exercise psychology* (1996), 17-35.

27 Brockmeyer, Jeanne H., Fox, Christine M., Curtiss, Kathleen A., McBroom, Evan, Burkhart, Kimberly M., and Pidruzny, Jacquelyn N. The development of the Game Engagement Questionnaire. *Journal of Experimental Social Psychology* (2009), 624-635.

28 Nacke, Lennart E. Game user research and physiological evaluation. In *Game user experience evaluation.* Springer, Toulouse, 2015.

29 Bromley, Steve, Mirza-Babaei, Pejman, McAllister, Graha, and Napier, Jonathan. Playing to Win? In Quandt, Thorsten, ed., *Multiplayer: The Social Aspects of Digital Gaming*. Routledge, Oxon, 2013.

30 Silveira, Fernando, Eriksson, Brian, Sheth, Anmol, and Sheppard, Adam. Predicting Audience Responses to Movie Content from Electro-Dermal Activity Signals. *UbiComp'13, ACM international joint conference on Pervasive and ubiquitous computing September and Behavior* (September 8-12, 2013), 707-716.

31 Boucsein, Wolfram. *Electrodermal Activity*. Springer US, New York, 2012.

32 Lu, G., Yang, F., Taylor, J. A., and Stein, J. F. A comparison of photoplethysmography and ECG recording to analyse heart rate variability in healthy subjects. *Journal of Medical Engineering & Technology* (2009), 634–641.

33 Seif El-Nasr, M., Drachen, A., and Canossa, A. *Game Analytics: Maximizing the Value of Player Data*. Springer, London, 2013.

34 Mcalmont, T. Tips for a Great First Time User Experience (FTUE) in F2P Games (September 3, 2015), http://blog.gameanalytics.com/blog/tips-for-a-great-first-time-user-experience-ftue-in-f2p-games.html.

35 Higgins, Krystal. First time user experiences in mobile apps (April 15, 2012), www.kryshiggins.com/first-time-user-experiences-in-mobile-apps/.

36 Novik, David G and Ward, Karen. Why don´t people read the manual? *Departmental Papers (CS)* (2006), Paper 15.

37 King Digital Entertainment. 2016. Candy Crush Jelly Saga. Game [mobile] (Jan 6, 2016) King Digital Entertainment, Dublin, Irland. Played january 2016.

38 Norsfell Games Inc. 2014. WinterForts: Exiled Kingdom. Game [mobile] (Oct 1, 2014).Execution Labs, Montréal, Canada.Played january 2016.

39 Norsfell Games Inc. 2015. Pogo Chick. Game [mobile] (Aug 13, 2015). Norsfell Games Inc, Montréal, Canada.Played january 2016.

40 Bordens, Kenneth S. and Abbott, Bruce B. *Research Design and Methods A Process Approach*. McGraw-Hill, New York, NY 10020, 2011.

41 Accessed on 2017-07-19, Wikidot. (2017). Retrieved from http://mgur.wikidot.com/start.

42 Khawaji, Ahmad , Zhou, Jianlong, Chen, Fang, and Marcus, Nadine. Using Galvanic Skin Response (GSR) to Measure Trust and Cognitive Load in the Text-Chat Environment. In *SIGCHI Conference on Human Factors in Computing Systems* (Crossings, Seoul, Korea 2015), ACM Digital Library, 420-423.

43 Lyle, John. Stimulated recall: a report on its use in naturalistic research. *British Educational Research Journal* , Volume 29, Issue 6 (2003), 861-878.

44 Poels, Karolien, de Kort, Yvonne, and Isselsteijn, Wijnand. Identification and Measurement of Post Game Experiences. *Westminster paper* (April 2009), 109 -129.

45 Nacke, Lennart and Lindley, Craig A. Flow and Immersion in First-Person Shooters: Measuring the player's gameplay experience. In *Proceedings of the 2008 Conference on Future Play: Research* (Toronto, Ontario, Canada 2008), ACM, 81-88.

46 Mirza-Babaei, Pejman, Wallner, Günter, McAllister, Graham, and Nacke, Lennart E. Unified visualization of quantitative and qualitative playtesting data. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems* (New York 2014), ACM, 1363-1368.

47 Thomsen, Line Ebdrup, Petersen, Falko Weigert, Drachen, Anders, and Mirza-Babaei, Pejman. Identifying Onboarding Heuristics for Free-to-Play Mobile Games: A Mixed Methods Approach. In *Entertainment Computing - ICEC 2016*. Springer, 2016.

48 Kivikangas, Matias J., Ekman, Inger, Chanel, Guillaume et al. Review on psychophysiological methods in game research. In *Nordic DiGRA 2010* (Stockholm 2010), Authors & Digital Games Research Association (DiGRA)., 181-199.

49 Ravaja, N, Turpeinen, M, Saari, T , Puttonen, S, and Keltikangas-Jarvinen, L. The psychophysiology of James Bond: Phasic emotional responses to violent video game events. *Emotions vol. 8* (Feb 2008), 114-120.

50 Cacioppo, John T., Tassinary, Louis G., and Berntson, Gary G. *The Handbook of Psychophysiology*. Cambridge University Press, New York, 2007.

51 Julkunen, J. 3 Things to KNow About Session-Length Restriction When Designing a Free2play Game (Februray 19, 2015), http://www.bluecloudsolutions.com/articles/mobile-gaming-sessions-now-longer/.

52 Saari, Timo, Ravaja, Niklas, Salminen, Mikko, Laarni, Jari, and Kallinen, Kari. Phasic Emotional Reactions to Video Game Events: A Psychophysiological Investigation. *Media psychology* (2006), 343-367.