Improving Level Design Through Game User Research: A Comparison of Methodologies

Marcello A. Gómez-Maureira, Michelle Westerlaken, Dirk P. Janssen, Stefano Gualeni, Licia Calvi

NHTV University of Applied Sciences,
Monseigneur Hopmansstraat 1,
4817 JT Breda, The Netherlands

Abstract

In this article we compare the benefits for game design and development relative to the use of three Game User Research (GUR) methodologies (user interviews, game metrics, and psychophysiology) to assist in shaping levels for a 2-D platformer game. We illustrate how these methodologies help level designers make more informed decisions in an otherwise qualitative design process. GUR data sources were combined in pairs to evaluate their usefulness in small-scale commercial game development scenarios, as commonly used in the casual game industry. Based on the improvements suggested by each data source, three levels of a Super Mario clone were modified and the success of these changes was measured. Based on the results we conclude that user interviews provide the clearest indications for improvement among the considered methodologies while metrics and biometrics add different types of information that cannot be obtained otherwise. These findings can be applied to the development of 2-D games; we discuss how other types of games may differ from this. Finally, we investigate differences in the use of GUR methodologies in a follow-up study for a commercial game with children as players.

Keywords:
Games User Research, Quality Assurance, Level Design, Platformer, Game

Email addresses: ma.gomezmaureira@gmail.com (Marcello A. Gómez-Maureira), michellewesterlaken@gmail.com (Michelle Westerlaken), janssen.d@nhtv.nl (Dirk P. Janssen), gualeni.s@nhtv.nl (Stefano Gualeni), calvi.l@nhtv.nl (Licia Calvi)

Preprint submitted to Entertainment Computing

1. Introduction

In 1983, the video game industry in North America, which had been buoyant up to then, collapsed because so many low quality products had entered the market that customers turned away [1, 2]. After this, game companies became more and more aware of the importance of quality testing. Nintendo was one of the first companies to adopt Quality Assurance (QA) as part of the game development phase in 1985: before releasing a game, they would undergo an iterative process whereby players' feedback of the game design and mechanics are reported back to the designer and used to optimize the game design itself [3].

In this article, we will concern ourselves with one particular type of QA, which is also called Game User Research (GUR). The term GUR is mainly used in academic research, but industry practice also distinguishes between for example fault testing ("Is the product bug free?") and user testing ("Do players like it?") as well as the usage of methods to provide feedback directly on the design [4].

Within GUR, there are three major types of information available [5]: Data from interviews (the users opinion voiced in a structured conversation with the researcher); data from player metrics (the in-game behavior measured and tracked by the computer itself), and data from psychophysiology (the bodily responses caused by the game as observed by sensors applied to the players). In keeping with industry terminology [6], the terms 'psychophysiology' and 'biometrics' are used interchangeably throughout this article.

There has been some previous work on the relative value of the different types of information. It has been suggested that biometric testing is useful for adjusting level design and difficulty [7]. Comparing interviews and psychophysiological data, these authors found that implementing changes based on both data sources made the game experience more pleasant and satisfactory for the target audience. On a few other dimensions, implementing the suggestions from psychophysiological data increased the quality of the game by a small but significant amount, while implementing the changes suggested by interview data did not raise the game above a non-GUR method [8]. Mirza-Babaei and colleagues conclude that a study into the combined effects of data sources would be prudent.

In this article we look at three methodologies, using three different sources of information, and compare which combinations are most productive in terms of the quality of the changes and the user evaluation of these changes. Through this comparison we want to illustrate how designers can gather and use GUR data to make informed decisions in their games. To simplify matters, we focus on 2-D level design: This is modular, fast and relatively easy to produce and iterate, and provides a clear basis for comparison among level-sets.

In the first data collection, we will use these three methodologies (interviews, metrics, biometrics) to get as much insight in the players' game experience as possible. All three measurements will be collected on each player. We will then combine the findings from these measurements to create improved versions of the game. Recall that the result of the three methodologies will identify possibilities for level improvement and we will derive design recommendations from them. We will combine the recommendations from two out of three methodologies. Doing this three times for each possible pair-wise combination will results in three different level implementations that correspond to the three possible combinations of methodologies.

In the second and final data collection, the improved versions of the game are compared in terms of player feedback. We can then decide which combination of two methodologies leads to levels that are evaluated best by a group of independent players.

We chose to use a clone of the well-known 2-D platformer *Super Mario Bros.*[1] [9], called *SuperTux*. This meant that all players were familiar with the objectives, the gameplay, the mechanics, and the metaphors used in this type of game. At that point we could look at the effect of level design while excluding any effect of emotional experience with this type of game. We also controlled how many computer games our participants played in general, to avoid any generic effects of experience with games.

Before any data were collected, three of the levels provided with *SuperTux* were selected and partially modified by the first author to create levels of equal length and increasing difficulty. These levels were evaluated by a group of five game designers in respect to aspects such as difficulty progression, level flow [10] and clarity. Recommendations made by the designers included changes in level geometry, obstacle placements and similar parameters to

[1]Super Mario is a registered trademark of Nintendo

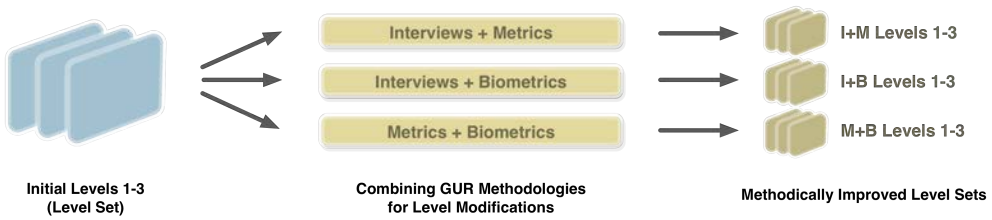

Figure 1: Graph illustrating the combination of methodologies in this study.

strike a balance between challenge and accessability. All recommendations that were supported by the majority were implemented.

The three levels were then presented to 20 participants as part of data collection one. The experience of each participant was measured with the forementioned three methodologies:

1. Participant *interviews* with player observation by researchers. Players were interviewed for about ten minutes, using a standardized script. They also filled out a 50 item questionnaire.
2. Data collection through *metrics*; the game was modified to log data about user behavior and user-game interaction [11]. We logged a large number of events such as all types of movements, attacks (including attacker and target), collection of bonus items, upgrades, downgrades and game deaths, and each single key press made by the participant.
3. Data collection through *biometrics*; this data was gathered from the play tester by using sensors to monitor heart rate, skin conductivity and the activity of the two facial muscles, the zygomaticus major and the corrugator supercillii [12].

In our game improvement phase, data from two methodologies were combined to create a new, methodically improved version of the levels. This was done three times to cover all possible pair-wise combinations (see Fig. 1).

As mentioned earlier, the methodologies tested included metrics and biometrics, both of which are technologically facilitated GUR methods that have recently become more popular. Metrics has risen with the advent of mobile and web-based games [13], while hardware and software advances have made biometrics accessible enough for game companies to include them in the QA procedure [6, 14]. Substantial research on how useful biometrics is compared to the traditional evaluation methods is, however, still missing.

In the final part of this article we present a follow-up study involving a commercial game, in which combinations of GUR methodologies were used to identify problematic aspects in the level design. While the same methodologies were involved in the evaluation of the levels as for our study on *SuperTux*, here we looked at the combination of all GUR data and reported the results to the designers. Due to the differences of the games, as well as the request of the involved company to not disclose details about their game, we focus on how differences in target groups and game mechanics can influence the acquisition and evaluation of the individual GUR methodologies. We also look at similarities between the two studies to reflect on the results of our *SuperTux* experiment.

2. Related Work

As a young research field building upon Human-Computer Interaction (HCI) and Experimental Psychology, Game User Research (GUR) studies the player experience from a player(user)-centered perspective. However, contrary to another user-centered design discipline like HCI for which methodologies and standards are already widely accepted, GUR is still working on the validation and standardization of procedures around data collection and analysis methods. In particular, what is felt as missing is a comparison and better understanding of the different data sources and analysis: What is best suited to which part of the game design analysis? What is their relative efficacy and effectiveness compared to each other? What is their relation to traditional testing methods like interviews and player observations?

In [15], data collected by player observation were compared with data collected using biometrics, particularly measuring Galvanic Skin Response (GSR, also called skin conductivity). The study aimed to identify which specific types of game user elements each method would single out for improvement, if any. This comparison demonstrated that these two methodologies (player observation and biometrics) reveal different issues: Player observations mainly identified usability problems and issues related to game mechanics, while biometrics identified issues with the player experience as such, and connected to the gameplay in terms of engagement, immersion, and emotional reactions. This specificity and complementarity suggests the adoption of a mixed method in testing games.

A recent study looked at the combination of biometrics with a think-aloud protocol [16]. These authors used four types of biometric data (GSR,

heart rate, and activity of the facial muscles responsible for smiling and frowning). They concluded that think-aloud protocols and biometric data provided different and mostly independent sources of information. Like us, they found that there were various practical hurdles in combining data from a such a large number of sources with different timing characteristics.

A follow-up study by Mirza-Babaei, et al., [8] focused more specifically on the differences between a game improved by using player interviews only, to a game improved by means of a combination of interviews, biometric and metric data. From the player's perspective the two improved games did not differ much. However, the designers could develop better visuals and a more engaging gameplay using mixed method data. The designers were also guided to implement many more changes than was suggested on the basis of interviews alone.

Our current study starts from a similar premise: identifying which combinations of GUR methodologies provide better player experience and satisfaction. However, we will not look at the designer perspecive but compare the resulting modified games in terms of player evaluation. Additionally, we will not compare two partially unbalanced conditions (interviews versus interviews with metrics and biometrics), but do a systematic comparison among all three possible pair-wise combinations of the three data sources.

3. The Game

As basis for our game research we chose *SuperTux* [17], a side-scrolling 2-D platforming game developed by the open source community (see Fig. 2). The game follows the design mechanics of the early Super Mario, a franchise on the Nintendo Entertainment System. As is the case in Super Mario, the player has to maneuver an avatar (the penguin *Tux*) through a two-dimensional game environment (a level) by means of running and jumping until the end is reached. In the course of the game, the player has to avoid obstacles such as pits or enemies. The level typically features ground surfaces to jump to and from, various items to collect and enemies to avoid, and platforms in mid-air that can be traversed. It is the occurrence of such platforms that give the genre its name, Platform games. Because of its pedigree *SuperTux* and similar games are often referred to as 'Super Mario Clones'.

We chose *SuperTux* specifically since it is freely available and can be modified by anyone due to its open source nature. The game under test will have to log its game state so that the data collection framework can link

Figure 2: A screenshot of *SuperTux* showing *Tux* - the protagonist - in its upgraded form (with red helmet), three enemies, several bonus coins, and four platforms.

events in the game to events recorded for the user [18]. The game comes with a tile-based level editor, which allowed for easy and modular modification of levels. Compared to other freely available clones, we consider *SuperTux* to adhere very closely to the tradition and the mechanics of the original *Super Mario Bros.* (e.g. jump height and related timings, which are not as close to the original in other clones). Finally, *SuperTux* features good quality graphic and sound assets, which makes the experience of the game close to what is expected of playing a commercial game.

4. The Experiment

The experiment we conducted comprised several phases (five in total) in which different GUR methodologies were used in combination to improve three separate levels that were designed in conformity to the design styleguide for *SuperTux*, and that were characterized by increasing difficulty.

The flowchart in Fig. 3 highlights the procedure that was adopted. According to this approach, a focus group of game designers initially evaluated the three levels and defined a benchmark (phase 1). The levels were then played by players (data collection 1), while we collected user experience data using all three GUR methodologies (phase 2). The data was analysed and improvement suggestions were derived from the data (phase 3). Next, the levels were modified according to the results of different pair-wise combinations of

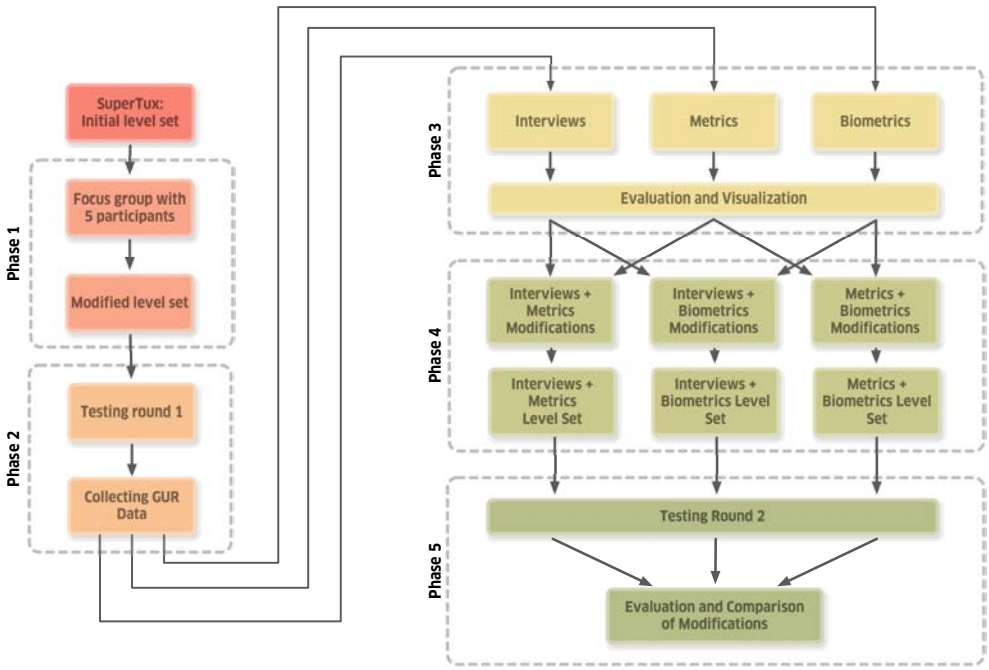

Figure 3: Flowchart illustrating the five phases of the experiment.

the three GUR methodologies (phase 4). These combinations were: interviews + metrics data, interviews + biometrics data, and metrics + biometrics data. The application of these combinations resulted in three different level sets, of three modified original levels each. These level sets were then played by a different group of players (data collection 2) and the levels were compared to each other through a questionnaire about player experience (phase 5). Each phase will be discussed in detail in the next subsections.

4.1. Phase 1: Preparation of the Benchmark Levels

In phase 1, three of the levels provided with *SuperTux* were selected and partially modified by the first author to create levels of equal length and increasing difficulty. Next, five game design students play-tested the game and then took part in a focus group and discussed all aspects of the game that should be changed in accordance to their knowledge and experience as game designers. All suggestions for changes that were brought up and supported by the majority of the focus group were implemented. Noteworthy modifications were, for example, the increase of enemies in proximity to invincibility pick-ups (which provided players the satisfaction of knocking out several enemies

in quick succession) and the placement of high level geometry towards the end of a level (to let players finish with the feeling of excitement).

This phase was intended to set a benchmark, to define a point at which professional designers would release their work for internal quality assurance.

4.2. Phase 2: GUR Data Collection

In phase 2, the first round of playtests were conducted. A total of 20 participants (8 of which were female) between the age of 18 and 57 (median age of 25) played through all three levels. On average, it took players 7.2 minutes to play through the levels, spending 1.8 minutes in the first level, 2.2 minutes in the second level, and 3.2 minutes in the third level.

During each session, the following data were recorded:

General level ratings: each player had to rate the level they had played in terms of fun, length and difficulty using a 5-point Likert scale. These three ratings were collected immediately after completing each level and were later used as reference points for the three GUR combinations.

Interview data: Open-ended, semi-structured interviews with players were held at the end of each level. While during gameplay, players were recorded using a microphone and two video cameras to capture images of the participants and of the game screen. The interviews focused on identifying any confusing, frustrating, enjoyable, and surprising parts in the level. Users were shown a print out of each level, showing all obstacles, bonuses and enemies, to pinpoint the exact location to which their comments pertained.

Game metric data: To collect metric data in SuperTux, logging functionality was added to the game. This allowed for periodic tracking of the player position as well as relevant game events, such as defeating enemies, jumps, collecting bonus items, etc. Game metrics were stored in clear text and time-stamped to be in sync with audio and video recordings.

Biometric data: All participants were monitored with several biometric sensors during the test sessions. Based on prior research in this field, we used facial Electromyography (EMG) sensors to detect activity in the Corrugator Supercilii muscle group (associated with frowning), and the Zygomaticus Major muscle group (associated with smiling). Both muscles are commonly used to measure emotional valence [19]. Finger sensors were used to measure

Blood Volume Pressure (BVP) and Galvanic Skin Response (GSR), which have been correlated to excitement, fear, engagement and arousal [19]. Due to technical difficulties during the evaluation phase of the study, GSR data had to be excluded from the set of biometric measures. After the experiment had concluded, we were able to recover GSR data. While it was at that point too late to use the data to evaluate level design, it was added to visualizations for review purposes.

Test sessions ended with a demographic questionnaire that also asked how frequent participants played video games. Participants in this research phase were selected by using convenience sampling from the campus and immediate surroundings of the university. We did not include data from participants that had ever been involved in game development, game art, or game programming, and we excluded all who identified themselves as 'hardcore gamers'.

4.3. Phase 3: Data Evaluation and Visualization

In phase 3, all data gathered in the previous phase was analyzed and processed. More specifically:

Interview GUR data: The data collected during the interviews was filtered to remove irrelevant information and categorized to make the data match the topics covered by the research, such as confusing, frustrating, enjoyable and surprising instances for each of the three levels (Fig. 4). In order to determine what improvements to implement for the following phase, potential changes were classified as actionable changes and non-essential changes, and the actionable changes were prioritized depending on the demand for change (ie, the number of participants requesting it).

Metric data: An in-house system of scripts and analysis software [18] was used to parse the game logs and derive play statistics, such as amount of collected bonuses, defeated enemies, distance walked, number of jumps made. Where necessary, measurements were divided by the time spent in the level, since the play duration had a direct effect on many play statistics (for example, the more time people spent, the more enemies they defeated). We also calculated correlations of the acquired metric data with each other. While these correlations can uncover possibilities for level improvement, we did not find actionable correlations this time. Apart from acquiring play statistics for each participant, the logs were used to create heatmaps (see

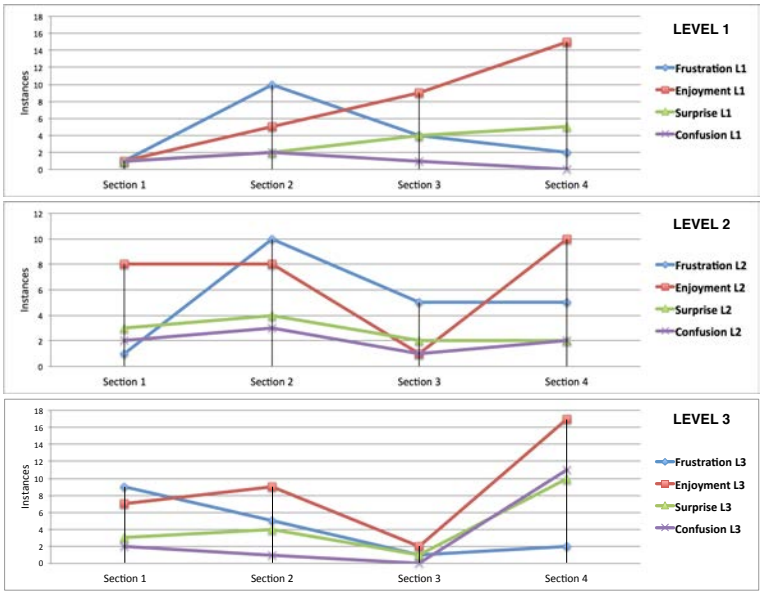

Figure 4: The graphs show aggregated counts of instances in each of the three levels that during interviews have been described as frustrating, enjoyable, surprising or confusing. Note that the graphs are not in the same scale (specifically level 2).

Fig. 5), which tied the position of the player as well as jumps, enemy kills, player deaths, and changes in direction to locations in the level.

Biometric data: For the evaluation of biometric data, each level was divided into 12 equally long sections. Since biometric data works best by averaging over time and participants, we chose this segmentation to strike a balance between having enough data and having localized data. Specifically players that moved through the level segments very fast yielded only limited biometrics data on, for example, making a jump in a particular segment. For the visualization of biometric data, the individual biometric measures were expressed in graphs and presented next to the corresponding level sections as shown in Fig. 6. Biometric data was further analyzed by linking each short stretch of biometric data to the game event that (likely) provoked it. However, as we were looking for actionable suggestions for changing our levels, these results are not discussed here.

As mentioned before, we were not able to include GSR measurements during the data evaluation stage because of a technical problem with the amplifier in the GSR sensor. We were able to recover a large amount of this

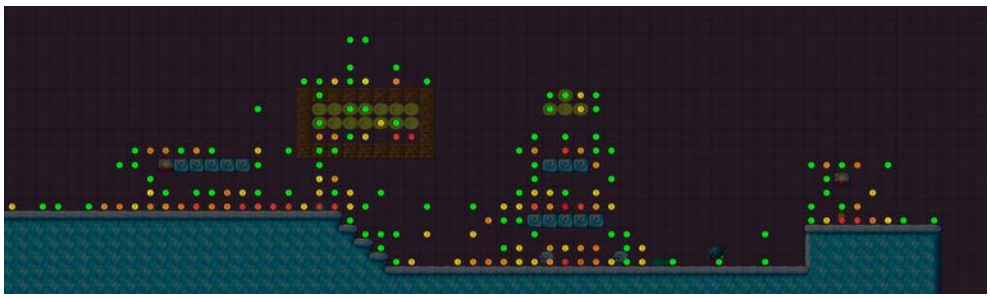

Figure 5: Heatmap example showing part of a level overlayed with colored markers that indicate where in the level players switched their movement direction. Marker colors range from green (indicating a single event) to red (indicating the maximum amount of events in a level). As the level geometry is tile-based, events were grouped to tiles and illustrated as heatmap marker per tile.

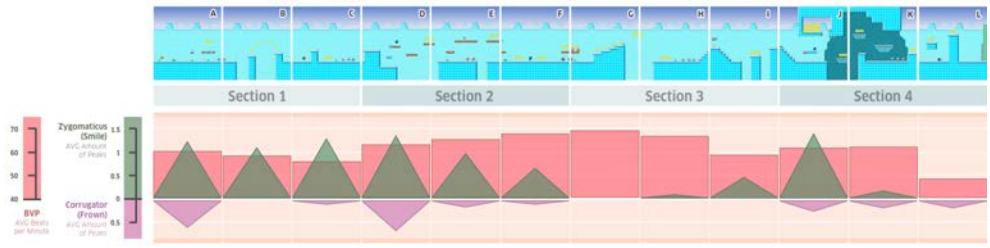

Figure 6: Biometric data superimposed over map sections of level 3. The top row shows the level graphics split into the four sections used during interviews and 12 sections used for biometric GUR data. The red bars show the mean BVP of all participants in each section. The height of the green triangles pointing upwards shows the mean of all smiles for each section while the purple triangles pointing downwards show the mean of all frowns for each section.

data at a later stage, albeit too late to include it in the level modifications. However, as GSR is a commonly used methodology it is of interest to see whether it would have added new information.

Due to the nature of the defect in the amplifier, we could retrieve partial data from 9 participants. For these participants, the signal would drop out intermittently with more signal loss as time progressed. The signal loss was intermittent and GSR curves could clearly be determined from the good data. Bad data was visually easy to spot and we wrote a heuristic script that removed bad data by looking for values that were repeated over and over again, often at a large distance from the previous sample. This script was rigorously tested and written to remove data when in doubt.

To deal with the increasing likelihood of signal loss over time, we decided to analyze the first of the three levels only. For each participant, the GSR signal over five regions of interest (ROIs, see Fig. 7) were computed by first standardizing the signal on a per-participant basis and then counting the number of times that the signal made an upwards jump of at least 0.002 z-value.

We did an additional analysis in which we counted the number of jumps over the threshold value but the results were identical to the ones reported here. Scores were averaged over participants to detect trends and remove random noise inherent to all biometric signals. It should be noted that despite our care in handling this data, the number of participants is low and there are regions with substantial data loss, which makes it important to interpret these results carefully.

The results of the GSR data aligned with the results we simultaneously recorded from the heart rate: GSR went up sharply in ROI-B, which features a number of hard to get bonus items and heart rate also goes up during this ROI. Both measures are in fact highest in ROI-B. GSR went down again for ROI-C, to then slowly climb over ROI-D and ROI-E. Again, heart rate (BVP) mimics this pattern of a sharp fall at ROI-C and steady increase after that.

Based on the limited data we have available, it seems that GSR would not have revealed any unique information or given any different suggestions in this particular case. This is because GSR and BVP show by and large the same pattern in this particular case. This is not surprising, as some of the factors that can cause an increase in GSR also cause an increased heart rate (surprise, upset). However, other emotions or mental states, such as concentration, can load differentially on these two sensors and games that

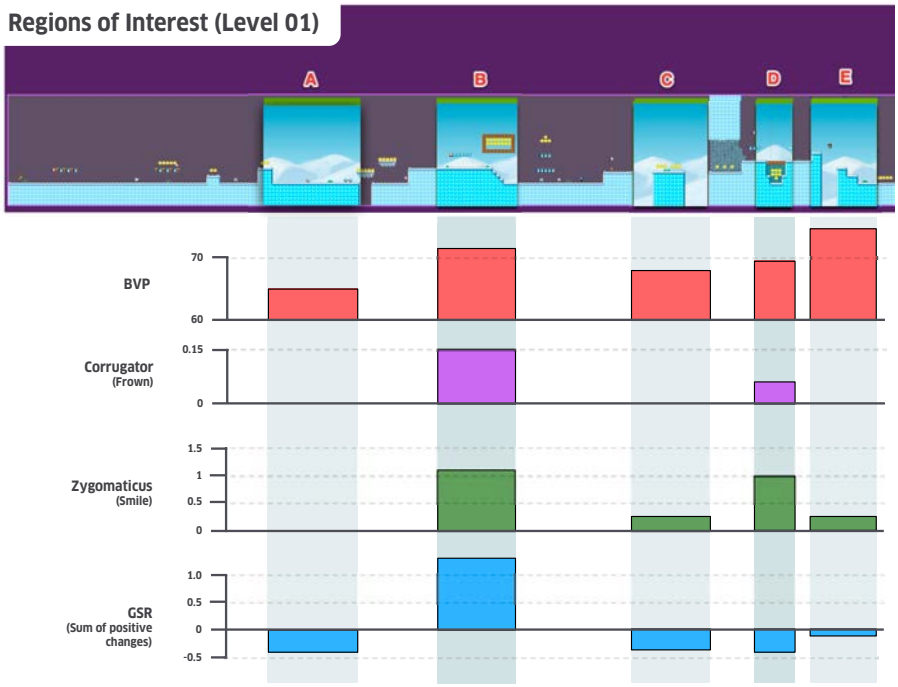

Figure 7: Biometric data superimposed over map sections of level 1. The top row shows the level graphics split into five regions of interest. The colored data bars below indicate the mean biometric measures for BVP, facial muscle activities and GSR across all participants for each region of interest in level 1. Note that the measures do not share the same vertical scale, as indicated by the vertical labels.

want to test those states or emotions would be best tested with both sensors. The results from the facial muscles were not related to the results from GSR and BVP.

4.4. Phase 4: Modifications

In phase 4, we used the evaluated data and their visualizations to make improvements in the level design. Each level was modified based on the data collected in phase 3 according to the three different pair-wise combinations of methodologies. These were: (1) interviews and game metrics; (2) interviews and biometrics; and (3) game metrics and biometrics. Each combination was additionally augmented with the general level ratings in terms of fun, length and difficulty to contextualize the gathered data. For each combination of methodologies, a total of six changes were implemented across the three

levels. Where there were more suggestions for change than our approach allowed, we chose the changes that were suggested by the data from the largest number of users.

The following example should illustrate how GUR data was used to modify levels with the goal of improving player satisfaction. Note that this process took place for each of the three methodology pairings. In this example, we were looking for potential changes through the combination of interview data and biometrics data in the first level. The biometric data we gathered showed very little player response throughout this level, especially in the beginning of it, as determined by a low and steady BVP as well as a lack of facial muscle activity. Interview data showed that the beginning area of the first level had the fewest comments from players (both in terms of enjoyment and surprise, as well as in frustration or confusion) compared to the remaining level areas. Here, with data from both methodologies, the designer (first author) concluded that change in this area was desirable to induce higher player excitement. In this case, platforms were repositioned and added to increase the amount of vertical exploration space, while the amount of pickups and enemies were kept the same. Acquiring all collectibles in the area would now require more effort from players. On the other hand, the risk of failure due to coming in contact with enemies was kept the same, as the beginning of the first level was intended to be low in difficulty. An illustration of the changes can be seen in Fig. 8.

While the decisions on which changes to implement was solely based on the collected data, their implementations in terms of level design were taken based on the professional experience, sensibility, and best efforts of the level designer. In all cases, changes were implemented locally at problematic areas without impacting the rest of a level.

4.5. Phase 5: Evaluation of Modifications

In phase 5, the last phase of our experiment, new participants, chosen through convenience sampling on the campus and in the environment of our university, were recruited to playtest one of the three modified level-sets that had been created in previous phase. A total of 40 participants (22 of which were female) in ages ranging from 15 to 27 years (median age of 23 years) took part in this second data collection session.

Players played one version of the modified level-sets, consisting of three levels. We did not want to repeat levels between players to avoid confusion and order effects. Players were asked to complete the Game Experience

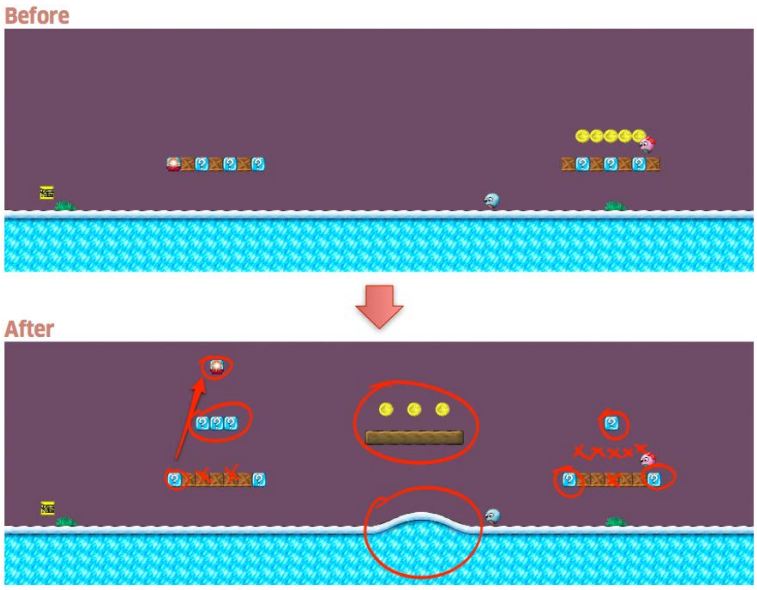

Figure 8: Example of changes in a level segment that have been performed in response to interviews and biometric GUR data.

Questionnaire (GEQ, [20]), a tool that is commonly used in the GUR field [7, 21] to quickly analyze player experience. Of the four modules described in [20], we administered the 'Core' and the 'Postgame' module. Using the factor weights listed in the source, the 50 answers were reduced to 11 dimensions.

5. Results

The graph shown in Fig. 9 illustrates the results of the GEQ questionnaire for each of the three methodology pairings. On the whole, the differences between the combined methodologies were much smaller than we expected. A consistent pattern can be seen across the variables *Positive Affect*, *Flow*, *Positive Experience*, and *Competence*: The levels modified by 'interview & metrics' were rated most positively. This result is replicated for the negative dimensions *Tension/Annoyance*, *Challenge*, and *Negative Affect*.

Levene's test for homogeneity of variances showed that each of the methodology pairings were heterogeneous, except in the GEQ dimensions *Tension/Annoyance* (p=.015) and *Returning to Reality* (p=.013). These dimensions were therefore not used for further statistical analysis, as the group results were considered to be too similar. ANOVA analysis of the remaining dimensions

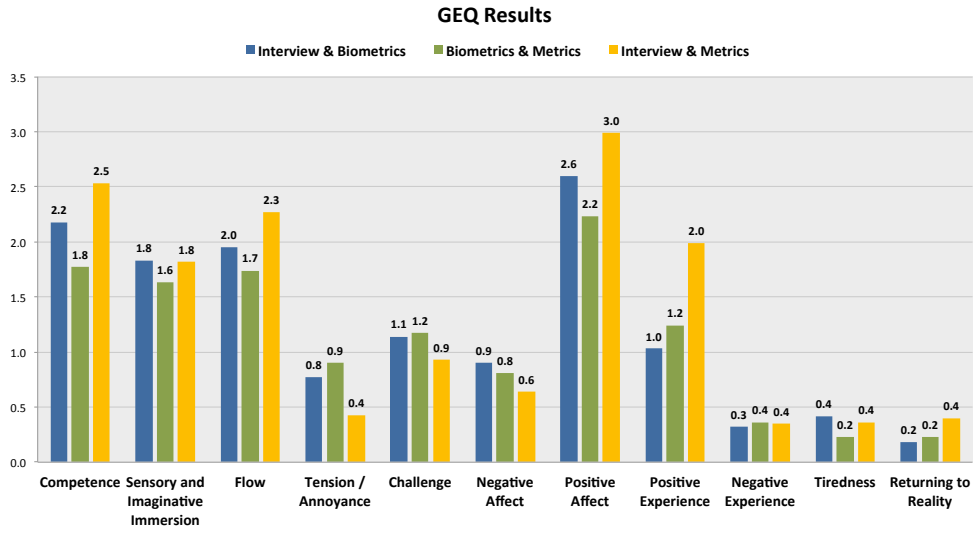

Figure 9: Graph showing the GEQ scores of the individual methodology testing groups divided by the aspects that are scored by the GEQ.

showed that three dimensions had statistically significant differences between the groups: *Positive Experience* (p=.008, F=5.595), *Competence* (p=.04, F=3.530), and *Positive Affect* (p=.048, F=3.290). Tukey's post-hoc test further showed that the significant differences within those dimensions are found between the pairings 'interview & metrics' and 'biometrics & metrics' for the dimensions *Competence* (p=.03) and *Positive Affect* (p=.038), and between the pairings 'interview & metrics' and 'interview & biometrics' for the dimension *Positive Experience* (p=.008).

The results indicate that the different methodology pairings produce different averages in the dimensions of the GEQ questionnaire, especially pronounced in *Positive Experience*, *Competence*, and *Positive Affect*. Given that the variance of the results was found to be largely heterogeneous, the sampling of participants seems to be acceptable. However, we still caution that there could be a Type II error as the ANOVA analysis was significant for only a few GEQ dimensions.

From the point of view of a level designer, each pair of GUR methodologies was able to provide actionable indications regarding locations or situations that should be modified to improve player satisfaction. With few exceptions, each pair of methodologies offered a unique change recommendation for the

designer to act on.

6. Follow-Up Study

The insights of our research on *SuperTux* lead to a follow-up study in which the combination of GUR methodologies, including in-game metrics, psychophysiological data, observations, and player questionnaires was further evaluated. For this study we tested several levels of a serious game that was developed to help children cope with cognitive challenges. The producer is a successful commercial studio, who is developing this game in close collaboration with the funder. The aim of this research was to collect and evaluate GUR data on children and subsequently report our findings to the designers with suggestions on how levels could be improved. We tested levels of two mini-games within the larger game. Both levels involved navigating over predefined pathways to reach an objective, while avoiding dangers and distractions. One mini game had an additional system of time limits, while the other involved mental rotation to make the game more challenging.

In this section we focus on the general question of how the target group of young children affects the use of GUR methodologies and the extent to which that resulted in changes of our follow-up study compared to the research on *SuperTux*.

6.1. The Game(s)

The game in this study consisted of several sections with differing goals and game mechanics, with an overarching narrative connecting them. Given the relative autonomy of each section, they can be considered as games within a game or what is known in the industry as 'mini-games'. Each of these mini-games featured several levels of increasing difficulty. For this study we were asked by the developer to provide GUR on the level design for two of the mini-games. The developer provided logging capabilities required for tracking metric and biometric measures.

The first mini-game asked players to collect a series of in-game items and return to the level start within a specific amount of time. Players controlled a player avatar through interconnected pathways, shown from an axonometric perspective and using the mouse as input. In the second mini-game players had to travel from the start of a level to a predefined end position while avoiding obstacles. Here the movements of the avatar (viewed from an oblique perspective) were not executed in real-time but rather had to be entered in a

movement queue that would then be executed. Generally, mini-game 1 can be considered more action oriented while mini-game 2 is closer to a puzzle game.

6.2. The Experiment

A total of 37 children (7 of which were female) between the age of 6 and 8 participated in the playtest sessions for this game. After a short introduction and the application of the biometric sensors, the participants could play the levels of one mini-game for 20 minutes. After playing, the children were asked to answer questions from the KidsGEQ [22], a questionnaire similar to the GEQ used in our *SuperTux* study that was specifically developed for children. It should be noted that especially younger children required some help of the researchers to answer the questions. Where help was necessary, researchers rephrased the questions and helped children answer while remaining neutral and objective.

The procedure of acquiring psychophysiological data during the test sessions was identical to our *SuperTux* study, including the use of sensors and the method of capturing the data. In contrast to the *SuperTux* experiment, we were able to include GSR in the evaluation process. Another important difference was that the experiment took place outside the laboratory at several different after-school facilities.

The collection of in-game metric data naturally differed between the two mini-games, as they involved different game mechanics. Parameters that were tracked for both of them included player position for the creation of heatmaps, level completion times, and level attempts when objectives were not met at the first attempt.

6.3. Results

On the whole, the type of conclusions we were able to draw from the two mini-games mimicked the conclusions we drew from SuperTux: On the basis of the combined data sources, we were able to recommend helpful changes to the designers about the levels, tutorials, and some tweaks to the overall gameplay. In contrast to *SuperTux*, we mostly needed the results from all three methodologies to come to well-grounded recommendations that were based on observable facts. Furthermore, different from *SuperTux* was that the biometrics data gave us enough spatial resolution to make statements about certain pathways in mini-game 1: Children moved through the levels

more slowly and the general speed of these mini-games was much slower, this made our biometric measurements more accurate.

7. Discussion

Given the insight gained in the five phases of our *SuperTux* study and the GEQ results that came out of it as a result, we feel that player interviews are essential to the success of improving level design and should therefore always be involved. Each methodology added its own recommendations, so none of them fully substitutes the others.

The combination of interviews and metric methodologies puts both subjective and objective information into context. While interviews are great to uncover problems in a level, we found that metric data provides useful information regarding how to solve these problems, for example on the basis of heatmaps. We feel that the biggest challenge for the use of metric data is the complexity (and consequent time consumption) of its evaluation. Furthermore, it ideally requires designers to a priori establish goals that can be expressed in metric parameters.

The addition of biometric methodologies into QA processes remains a promising possibility, especially for the exploration of qualitative aspects in design that are hard to evaluate through other means. As of the time of writing however, we believe that further efforts need to go into making the addition of this methodology less intrusive, less time intensive and therefore less costly. Only then can biometric methodologies be a viable addition to the QA processes of commercial game development.

Whether the 'interview & biometric' combination will be ranked second to 'interview & metrics' outside of the domain of level design for a 2-D platformer is an open question. In the level design problem considered here, the strongest limitation of the biometrics data was its lack of spatial precision. Because most participants completed a level in about 2 to 3 minutes, the amount of biometric data per game tile is smaller, and worse so when investing specific action - per game tile or group of tiles. When aggregating over many tiles, the data becomes insightful, however it is difficult to derive specific level design recommendations from this aggregated data that has lost its spatial specificity.

7.1. Discussing the Follow-Up Study

In our follow-up study, we gained further insights in the process of working with multiple, and partly simultaneous measures of GUR data. In the following paragraphs we discuss the most important differences and their impact as far as they extended on our *SuperTux* study.

7.1.1. From Controlled Laboratory to Controlled Field

Our user tests with the children took place in an after-school environment. Here, the research environment was set up in a separate and quiet area of the building where the participant was alone with the researchers and caretakers to avoid distractions and/or peer pressure. Collecting user data in the field introduced some logistical challenges, especially in regards to the collection of biometric data, as the sensor equipment had to be set up anew at each location. On the other hand, the ability to collect GUR data at a location that was familiar to the participants meant that the testing environment was close to that of a regular, unmonitored play session. Given the fact that games are ultimately played in such environments, it stands to reason that the ability to gather GUR data in the field holds the potential of acquiring data with higher ecological validity.

Especially in test sessions with young children, a familiar environment can help in providing conditions that lets them focus on the game content rather than on the unknown surroundings.

7.1.2. The Challenges of Interviewing Children

In the *SuperTux* experiments, we included questionnaires as well as open-ended interviews. However, in this follow-up study, we experienced that asking children between 6 and 11 years old to reflect on how much they liked a game can be a challenging task: For example, in some cases we noticed how a child was clearly confused during the play-session, but did not report this when the researchers asked questions about it. For this study, we therefore relied more on observations and added think-out-loud protocol to support interview data. In addition to this, the visual references in the form of handouts we used during the *SuperTux* study interviews could have helped the children in remembering the different elements and situations of the game.

7.1.3. Game User Research in a Commercial Setting

Whereas our *SuperTux* research study simulated the involvement of GUR methodologies in an ongoing development cycle, the outcomes of our follow-up study were communicated to an external team of designers at a point where the level design had already been considered final. An important difference between the two studies was also that no specific focus was given on which data source proved to be the most useful, as the specific value of each individual GUR methodology was not under review. As a result, the conclusions of this study proposed level design changes based on a combination of all the user research methodologies that were used, rather than the clear distinctions between different data sources that were used for *SuperTux*. From observations we noticed that a particular event in the game was perhaps too intimidating for children of this age and this argument was supported by the psychophysiological changes in the Blood Volume Pressure (BVP). Heatmaps that were produced through in-game metric data showed that the participants generally avoided these intimidating areas. A combination of these three results, mainly driven from an observational insight and supported by game metrics and biometric methods, resulted in suggested changes to the level design.

7.2. Does GUR Work for All Games?

As is the case with many forms of technically mediated evaluation, metric and biometric methodologies can be applied to games but do not always offer a thorough understanding of gameplay as experienced by the players. Both methodologies are by nature limited to understanding play as a quantifiable performance. While this quantification is somewhat inherent to all QA efforts, given that modifications are usually performed on quantifiable parameters (such as amount of enemies in *SuperTux*), interviews are sometimes closer to reflecting game experience from the players' perspective than metrics and biometrics. Metrics and biometrics approach and quantify the players' behavior to reach for a thorough understanding of a video game in terms of its design and its effects on the players. Is this a feasible goal?

Such a rational and quantifiable approach to 'play' does not account for aspects of engagement with games. What is lacking is for example a broader measure of player freedom and of participative experiences. A more qualitative measure would also be more open to negotiation and interpretation than a strictly quantitative procedure based on metrics and biometrics data.

However, even the interview methodology cannot fully reach into the domains that are not specific to video games, but apply to all games, such as the freely creative, ritual, social and transformative ones that constitute its 'myth domain' [23, 24].

Of the GUR methodologies, biometrics are specifically aimed at giving the experimenters a first peek at those internal user states, but this is a necessarily very limited one as biometrics only measure the changes in body state, not the underlying mental processes. Even brain imaging techniques like fMRI and EEG cannot currently capture these domains, which are traditionally researched using first-person methodologies shared by philosophers and anthropologists [25].

It is an open question how we can combine the results from these approaches to games, given that they have such different aims and domains of application. From a designer's point of view, a game concept cannot be characterized by counting the number of actions an average player takes, or the smiles it provokes. Similarly, from a player's perspective, the design philosophy or underlying message of a game design can be an abstract quantity that does not relate to their player experience at all: Media researchers have long known that media can be consumed by different audiences and for different end goals than what they were intended for [26]. And neither viewpoint includes more social and ritualistic aspects of games [23]. So evaluation of games on the basis of purely quantitative data should be done with an awareness that play is a complex activity which is deeply rooted in the very things that make us human, and that its experience may never be fully captured by questionnaires, interviews, observations or the statistical analysis of data.

Game developers as well as academics are aware of these opposing viewpoints to games. It may well be that most games have to be described at both levels to be properly characterized. It has been argued [27] that a purely quantitative GUR framework can be used to fully describe single-player video games that offer limited operative options to the players. Those are, in fact, games that more or less restrictively force the players to execute a specific set of actions in the ways the developers have envisaged them. This is the case of puzzle games, resource management games, point-and-click adventures, simple platformers, hidden object games, etc. So for these games at least, the two viewpoints coincide. But from a designer's point of view, these are the games of least interest.

SuperTux is a mechanically simple, single-player closed system. As benchmark for the comparison of GUR methodologies it allowed us to side step the

problem of describing highly complex games. The freely creative, socially relational and ritual dimensions of play are structurally absent by design. When focusing on such games, the work of GUR researchers can therefore not be accused of taking a derivative and impoverishing stance in relation to gameplay.

It should be noted too that improving games or levels from GUR research is actually not a fully qualitative undertaking. For one, there is the interpretation and subjectivity involved in correlating the game metrics and biometrics findings to the game at hand: An increased activity of the zygomaticus major can indicate the happy smile of a partial victory, or the relaxation and tension-release of an in-game death [22]. Second, the implementation of these findings is a highly subjective affair not any different from any other game design decisions.

7.3. Beyond 2-D Level Design

As a final point of discussion, we would like to address the possibility of using the processes described in this study for 3-D games. While the addition of a third spatial dimension raises the complexity in terms of visualizing data, there is no reason why the approaches we have taken would not work in 3-D space. Heatmaps in 3-D games are already part of metric evaluations and usually take an aerial perspective for the visualization of level geometry. Likewise we can imagine the use of such depictions of a level as visual aids during player interviews. In other words, while implementing GUR methodologies in 3-D games certainly raise the complexity compared to their use in 2-D games, we believe that such challenges can be overcome.

8. Limitations

8.1. Lack of a Baseline

A limitation of the current study is that we did not re-establish the baseline in the second data collection. While it would have been interesting to add a control group to the second testing round in form of an unmodified level-set, our research focused on the comparison of methodologies. In our study the assumption was taken that the implementation of GUR methodologies will raise player satisfaction.

8.2. Lead Researcher As Level Designer and Participant Observer

The author of this document was the lead researcher of the *SuperTux* study as well as acting game designer and was therefore involved in all steps of the *SuperTux* research. In being so, it becomes a challenge to remain objective over the course of the research. Also, while prior experiences as game and level designer have given the researcher insights into common design practices, it is ultimately difficult to prove a qualification in terms of level design. We have been aware of these limitations from the beginning of the study and attempted to mitigate these potential influences, for instance by providing the research with external input in form of a focus group and by requiring every level change to be based on research findings.

8.3. Level Designers Provide Subjective Influences

While there are many aspects of level design that follow certain logics and rules, the design of a level is highly dependent on the designer in terms of personal sensitivity, experience and interpretation of the development objectives for that particular product. Consequently it is inherently difficult to compare the quality and the merits of a design decision objectively. It should however be noted that a certain subjectivity of the designer is found in real world scenarios and is therefore always a factor in dealing with modifications due to GUR methodologies [8].

8.4. Combining Methodologies

Combining all methodologies or testing them separately in our *SuperTux* study could potentially have yielded different results. While we argue that the combination of GUR methodologies is a common practice and partly necessary depending on the methodology, it is likely that a combination of all methodologies would have given slightly different results. At the same time, it would have been interesting to see the individual influences of each data source. Our follow-up study therefore involved the combination of all GUR data sources.

9. Conclusion

In sum, we can conclude that QA efforts regarding improvements in level design benefit strongly from the involvement of player interviews and direct player observations. It stands to reason that having access to all three of the methodologies discussed in this paper has strongest benefit for designers, as

each methodology offers unique insights that can often not be accessed by other means. However, given the constraints of time and resources, studios may well be looking to add only one additional data source.

From our research, in-game metrics seem to be the most useful addition for 2-D platformers. This should be qualified by the observations that psychophysiological data may be less applicable to (2-D platformer) level design than to game design at large because of its relatively low spatial resolution. We think that the most important take-away point is that we found complementary benefits when combining methodologies: each methodology offers unique insights that can often not be accessed by other means. It is for this reason that the addition of biometric GUR as design evaluation method remains promising, despite the challenges in the evaluation and implementation.

Acknowledgements

Acknowledgements have been redacted from this version to meet the requirements of the double-blind review process.

References

[1] M. Ernkvist, Down many times, but still playing the game: Creative destruction and industry crashes in the early video game industry 1971-1986, History of Insolvency and Bankruptcy (2008) 161.

[2] D. Wesley, G. Barczak, Innovation and Marketing in the Video Game Industry: Avoiding the Performance Trap, Gower Publishing, Ltd., 2012.

[3] D. Sheff, Game Over: How Nintendo Zapped an American Industry, Captured Your Dollars, and Enslaved Your Children, Diane Publishing Company, 1993.

[4] M. Seif El-Nasr, A. Drachen, A. Canossa, Introduction, in: M. Seif El-Nasr, A. Drachen, A. Canossa (Eds.), Game Analytics, Springer-Verlag London, 2013, pp. 3–13.

[5] M. Seif El-Nasr, A. Drachen, A. Canossa, Game analytics: Maximizing the value of player data, Springer, 2013.

[6] M. Ambinder, Valve's approach to playtesting: The application of empiricism, in: Game Developer's Conference, 2009.

[7] S. Gualeni, D. Janssen, L. Calvi, How psychophysiology can aid the design process of casual games: A tale of stress, facial muscles, and paper beasts, in: Proceedings of the International Conference on the Foundations of Digital Games, ACM, 2012, pp. 149–155.

[8] P. Mirza-Babaei, L. E. Nacke, J. Gregory, N. Collins, G. Fitzpatrick, How does it play better? exploring user testing and biometric storyboards in games user research, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 2013, pp. 1499–1508.

[9] Nintendo, Super Mario Bros., NES/Famicom Cartridge (1985).

[10] J. Chen, Flow in games (and everything else), Communications of the ACM 50 (4) (2007) 31–34.

[11] A. Drachen, A. Canossa, Analyzing user behavior via gameplay metrics, in: Proceedings of the 2009 Conference on Future Play on GDC Canada, ACM, 2009, pp. 19–20.

[12] L. E. Nacke, An introduction to physiological player metrics for evaluating games, in: M. Seif El-Nasr, A. Drachen, A. Canossa (Eds.), Game Analytics, Springer-Verlag London, 2013, pp. 585–621.

[13] A. Canossa, Interview with nicholas francis and thomas hagen from unity technologies, in: M. Seif El-Nasr, A. Drachen, A. Canossa (Eds.), Game Analytics, Springer-Verlag London, 2013, pp. 137–143.

[14] V. Zammitto, M. Seif El-Nasr, User experience research for sports games, presented at GDC Summit on Games User Research.

[15] P. Mirza-Babaei, S. Long, E. Foley, G. McAllister, Understanding the contribution of biometrics to games user research, in: Proc. DIGRA, 2011.

[16] C. T. Tan, G. Studio, T. W. Leong, S. Shen, Combining think-aloud and physiological data to understand video game experiences, in: Proc. CHI, 2014.

[17] Open Source Community. Supertux development wiki [online] (2013) [cited 2014-04-06].

[18] D. P. Janssen, L. Calvi, S. Gualeni, A framework for biometric playtesting of games, Proceedings of the 8th International Conference on the Foundations of Digital Games (FDG 2013), 2013.

[19] J. T. Cacioppo, L. G. Tassinary, G. Berntson (Eds.), Handbook of Psychophysiology, 3rd Edition, Cambridge University Press, 2007.

[20] W. IJsselsteijn, Y. de Kort, K. Poels, A. Jurgelionis, F. Bellotti, Characterising and measuring user experiences in digital games, in: International conference on advances in computer entertainment technology, Vol. 2, 2007, p. 27.

[21] L. Nacke, Affective ludology: Scientific measurement of user experience in interactive entertainment.

[22] K. Poels, W. Ijsselsteijn, Y. de Kort, Development of the kids game experience questionnaire, Proceedings of Meaningful Play 2008.

[23] J. Huizinga, Homo Ludens: A study of the play element in culture, Beacon Press Boston, MA, 1992 (original work published 1938).

[24] B. DeKoven, The Well-Played Game: A playful path to wholeness, iUniverse, 2002.

[25] B. Nardi, My life as a night elf priest: An anthropological account of World of Warcraft, University of Michigan Press, 2010.

[26] T. E. Ruggiero, Uses and gratifications theory in the 21st century, Mass communication & society 3 (1) (2000) 3–37.

[27] M. Sicart, Against procedurality, Game Studies 11 (3).