Comparing Game User Research Methodologies for the Improvement of Level Design in a 2-D Platformer

Marcello Andres Gómez Maureira, Dirk P. Janssen, Stefano Gualeni,
Michelle Westerlaken, and Licia Calvi

NHTV University of Applied Sciences,
Monseigneur Hopmansstraat 1, 4817 JT Breda, The Netherlands
http://www.nhtv.nl

Abstract. In this paper we compare the effects of using three game user research methodologies to assist in shaping levels for a 2-D platformer game, and illustrate how the use of such methodologies can help level designers to make more informed decisions in an otherwise qualitative oriented design process. Game user interviews, game metrics and psychophysiology (biometrics) were combined in pairs to gauge usefulness in small-scale commercial game development scenarios such as the casual game industry. Based on the recommendations made by the methods, three sample levels of a Super Mario clone were improved and the opinions of a second sample of users indicated the success of these changes. 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.

Keywords: Games, Games User Research, Quality Assurance, User Testing, Level Design, Platformer, Game Industry, Casual Games, Combined Methodologies, Biometrics, Physiological Measures.

1 Introduction

In the late 1970s and early 1980s, when video games were still in their infancy, developers and programmers produced very personal and sometimes low-quality games as fast as they could [1]. This led to the North American video game crash of 1983, which demonstrated what it means if low quality products saturate a market [2]. In 1985 Nintendo started a strategy of far reaching quality testing and became the most successful console system [3]. Since then, quality assurance (QA) has become an essential phase of commercial video game releases. QA is almost always part of an iterative production process, with test results being reported back to the designers for evaluation. The objective of this process is to ensure that the intentions and goals of the underlying game design are successfully conveyed to the player, that the players understand the metaphors and new

D. Reidsma, H. Katayose, and A. Nijholt (Eds.): ACE 2013, LNCS 8253, pp. 77–92, 2013.

concepts that the game introduces, and finally that the positive and negative feedback is successful in motivating the player.

In this paper, 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?"*) and the usage of methods to provide feedback directly on the design [4].

Within GUR, there are three major types of information available: Data from interviews (the user's opinion); data from player metrics (the in-game behavior), and data from psychophysiology (the bodily responses caused by the game).

There has been some previous work on the value of the different types of information relative to each other. It has been suggested that biometric testing is useful for adjusting level design and difficulty [5]. Comparing interviews and psychophysiological data, it was found that 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 [6]. Mirza-Babaei and colleagues conclude that a study into the combined effects of data sources would be prudent.

In this paper 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. The choice for a clone of a well-known 2-D platformer *Super Mario Bros.* meant that almost all players know the objectives, mechanics and metaphors used in this type of game, so we can look at the effect of level design while excluding other variables.

The methodologies tested were:

1. Participant *interviews* with player observation by researchers.
2. Data collection through *metrics*; The game was modified to log data about user behavior and user-game interaction [7]. We logged a large number of events such as movements, attacks, collection of bonus items and key presses.
3. Data collection through psychophysiology (also called *biometrics*); This data was gathered from the play tester by using sensors to monitor heart rate, skin conductivity and the activity of the facial muscles [8].

In our game improvement phase, data from two methodologies were combined to create a new version of the levels. This was done three times to cover all possible combinations.

The first phase of the research involved evaluation of the initial three levels by a team of level designers. In the second phase we gathered GUR data on these three levels. The third phase is the evaluation and the processing of the gathered data by means of various (statistical) methods. The result of phase 2 and 3 was a clear set of problems and recommendations that should be dealt with. Each methodology rendered its own results. The fourth phase involved the qualitative implementation of these recommendations in changes to the levels by a level designer. There were three implementations, corresponding to the three possible combinations of GUR methods (we did not make changes based on all three recommendations combined). The fifth phase compared the different level-sets created by the three combinations.

Why include metrics and psychophysiology in this comparison? The collection of metrics data has gained an enormous popularity with the advent of web-based games, mobile gaming and consoles that are connected to the internet permanently [9]. In its wake, psychophysiological (biometric) testing has become available to companies within the game industry and several development studios have added the methodology to their QA efforts [10,11]. So far, there is little actual research into just how useful biometrics can really be, what other parts of game design it can be used for and how it compares to traditional testing methods such as interviewing or observing players.

2 Related Work

Game user research (GUR) is a relatively recent field of research, which draws upon theories and methodologies from Human Computer Interaction and Experimental Psychology to study digital games [12]. Research in this field may also be called ‘player experience research’ or research into ‘user-centered game design’. It involves studying the interaction between users and games with the aim of understanding, and ultimately improving, the user experience. While the body of research grows, there is currently no universally accepted methodology. Many questions remain about validity and procedure, about data collection and analysis methods [12].

Recent GUR studies have highlighted the need for research into a better understanding of the value of the different testing methodologies relative to each other. A 2011 study [13] compared the data obtained from traditional observation-based methods with that of biometric methods only (using input from galvanic skin response (GSR) as a data collection measurement). The results showed that different types of issues are revealed by the two approaches: Observation-based methods mainly exposed issues related to usability and game mechanics, while biometric research analysis was more suited to discovering issues related to gameplay and emotional immersion. Both methods uncovered unique issues that the other method did not reveal. The study concludes that using a mixed-methods approach allows for greater confidence and validation of issues. The approach has received positive feedback from game developers and producers that the researchers have collaborated with [13].

Mirza-Babaei et al. [6] performed another study in the same direction, which is strongly in line with the aims of our research. Their experiment aimed to “identify the strengths, weaknesses and qualitative differences between the findings of a biometrics-based, event logging approach and the results of a full, observation-based user test study”. The authors compared a game modified with the help of ‘Classic User Testing’ (Classic UT) to one modified through ‘Biometric Storyboards User Testing’ (BioSt UT). They found that “BioSt can help designers deliver significantly better visuals, more fun, and higher gameplay quality than designing without UTs and that classic UTs do not provide this significant advantage”. From the point of view of the players, however, BioSt UT and Classic UT did not differ from each other in terms of the ratings given to the resulting games. It is important to highlight that the two approaches compared are already mixed method approaches: the Classic UT consisted of interview and observation, while the ‘Biometric Storyboards’ included a blend of interview, metric, and biometric data.

The paper points out that “the usefulness of **user tests** for game designers has not been studied sufficiently” [6, p. 1]. The authors attempt to remedy this by evaluating how the game designers in their study approached and used the data from the user tests, and how the generated design recommendations differed qualitatively. Their results show that designers working with BioSt UT generated the largest number of game changes, and had the highest confidence ratings about changes compared to the designers working with Classic UT or no UT.

Where this previous paper provides valuable insights and findings in the use of BioSt UT to provide more nuanced game design improvement, our study attempts to separately pair the three introduced GUR methodologies to find out which combination leads to improved player satisfaction. In our experiment we look at participants with a casual player profile rather than experienced PC gamers.

3 The Game

As basis for our game research we chose *SuperTux*, a side-scrolling 2-D platforming game developed by the open source community (see Figure 1). The game follows the design mechanics of the early Super Mario franchise on the Nintendo Entertainment System. As is the case in Super Mario, the player has to maneuver an avatar 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 not only ground surfaces to jump to and from, but also platforms in mid-air that can be traversed. It is the occurrence of such platforms that give the genre its name. Platform games that largely imitate the game mechanics of Super Mario are often referred to as ‘Super Mario Clones’. SuperTux is one of such clones.

We chose SuperTux specifically since it is freely available and can be modified by anyone due to its open source nature. Since we logged game states as part of the metric and biometric data collection (described in later chapters), this was a necessity in absence of a collaborating game development team. The game comes with a tile-based level editor, allowing for easy and modular modification of levels.

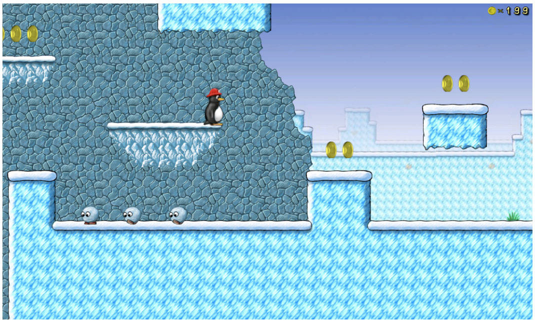

Fig. 1. A screenshot of SuperTux showing Tux - the protagonist - in its upgraded form (with red helmet), three enemies, several bonus coins, and four platforms

4 Experiment

The comparative design of this study consisted out of 5 different phases and entailed a comparison between levels that have been modified with a different combination of GUR methodologies. The levels were part of a sequence (level-set), which consisted of a tutorial (which was not changed) and three levels of increasing difficulty.

A flowchart of the approach is shown in Figure 2.

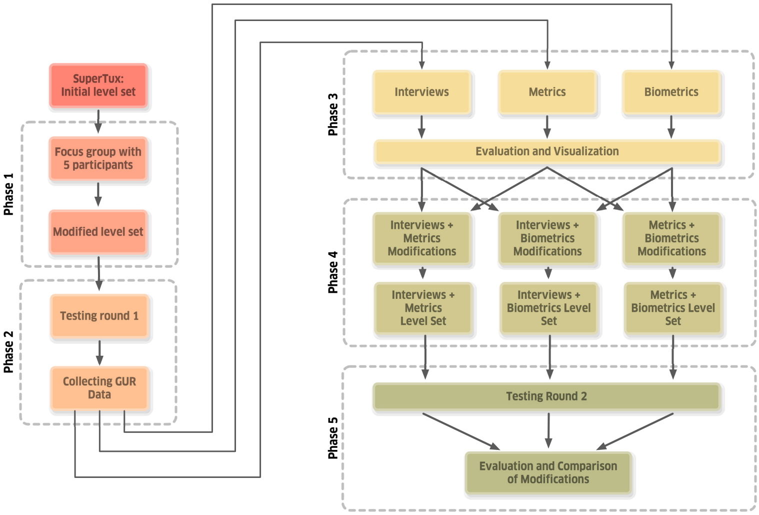
SuperTux: Initial level set
Phase 1
Focus group with 5 participants
Modified level set
Phase 2
Testing round 1
Collecting GUR Data
Phase 3
Interviews
Metrics
Biometrics
Evaluation and Visualization
Phase 4
Interviews + Metrics Modifications
Interviews + Biometrics Modifications
Metrics + Biometrics Modifications
Interviews + Metrics Level Set
Interviews + Biometrics Level Set
Metrics + Biometrics Level Set
Phase 5
Testing Round 2
Evaluation and Comparison of Modifications

Fig. 2. Flowchart illustrating the five phases of the experiment

In phase 1 a focus group of designers evaluated the initial benchmark level-set which incorporated the first three levels of the game SuperTux and suggested modifications as a basic quality assurance. This level-set was then tested in phase 2, where the first round of GUR data was collected. In phase 3, the collected data of the first testing round was evaluated and visualized according to the three GUR methodologies that were used in this study. Phase 4 involves implementing changes using three different combinations of GUR methodologies (including modifications based on results derived from interviews + metrics data, interviews + biometrics data, and metrics + biometrics data). This resulted into three different level-set versions. These level-sets formed the basis for Testing Round 2 in phase 5 and the modifications were then evaluated and compared with each other. Each of these phases will be further described in the next section of this paper.

5 Preparation of the Benchmark Levels (Phase 1)

The first phase of the experiment focused on providing an initial quality assurance of the level-set. By evaluating the quality of the original first three levels in the game before conducting further experiments, we intended to emulate a point during level development at which professional designers would release their work to internal quality assurance. The decision to base this on the original first three levels of the game was taken under the assumption that the creators of the game intended these levels to be playable without requiring prior knowledge or advanced skills while still providing a progression in difficulty. The individual levels are unique and require the player to understand game mechanics that were introduced in preceding levels.

To ensure a high quality standard of the level-set in terms of level design, five game design students play-tested the game as part of a focus group and discussed problematic aspects that should be changed in accordance to their knowledge and experience as game designers. After the transcription of the focus group, a list of suggested modifications was made and subsequently implemented in the level design of the initial levels. These modified levels were used in phase 2.

6 GUR Data Collection (Phase 2)

In the second phase of the experiment we conducted test sessions. A total of 20 participants (8 of which were female) between the age of 18 and 57 (median age of 25) played through all levels in the level-set. On average, it took players 7.2 minutes to play through the level-set, spending 1.8 minutes in the first level, 2.2 minutes in the second level, and 3.2 minutes in the third level. During these test sessions, we recorded the following data:

General Level Ratings. After each level, participants were asked to rate the level regarding fun, length, and difficulty. Ratings were given based on a 5-point Likert scale with fun ranging from 'Not at All' to 'Very Fun', length ranging from 'Too Short' to 'Too Long', and difficulty ranging from 'Easy' to 'Difficult'. The ratings of a level were used in all three GUR methodology combinations as a basic reference point.

Interview Data. During the test sessions, the researchers monitored participants with two video cameras and a microphone through which they recorded both the participant and the game screen. Observations were noted down and peculiar situations were brought up during open-ended interviews.

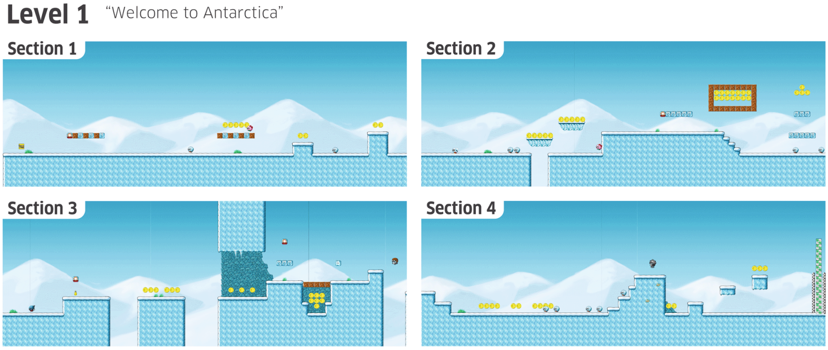

Fig. 3. Example of a level sheet used to aid participants in recalling details about their play experience. Levels are divided into four uniformly sized sections and are numbered chronologically.

The interviews took place after each level and asked participants to answer a set of semi-structured questions, which inquired about confusing, frustrating, enjoyable, and surprising parts in the level. While answering these questions, the participants could refer to a visual aid that divided each level into four equally sized and numbered sections (see Figure 3) to allow locating the source of the remark. Participants were asked to give any remaining comments or feedback after all questions.

Game Metric Data. Due to the open source nature of SuperTux, the researchers were able to add logging functionality to the game, which periodically tracked the position of the player character as well as relevant game events, such as defeating enemies, jumps, collecting of bonus items, etc. Game metrics were stored in clear text and time-stamped to be in sync with audio and video recordings.

Biometric GUR 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 [14]. 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 [14]. Due to technical difficulties, we had to exclude the skin conductivity data.

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 [15] in the immediate surroundings of the University. We decided to not include participants that had ever been involved with game development and all those who would identify themselves as 'hardcore gamers'.

7 Data Evaluation and Visualization (Phase 3)

In the third phase of the research, we analyzed and processed the GUR data that we acquired from test sessions in phase 2.

Interview GUR Data. After transcribing observation notes and interview feedback, we filtered the data to exclude information that was considered irrelevant for the goals of this research, such as requests for additional game mechanics. The filtered data was then divided into different general themes in correlation to the topics of the interview questions ('Confusing Instances', 'Frustrating Instances', 'Enjoyable Instances' and 'Surprising Instances'). By visualizing the

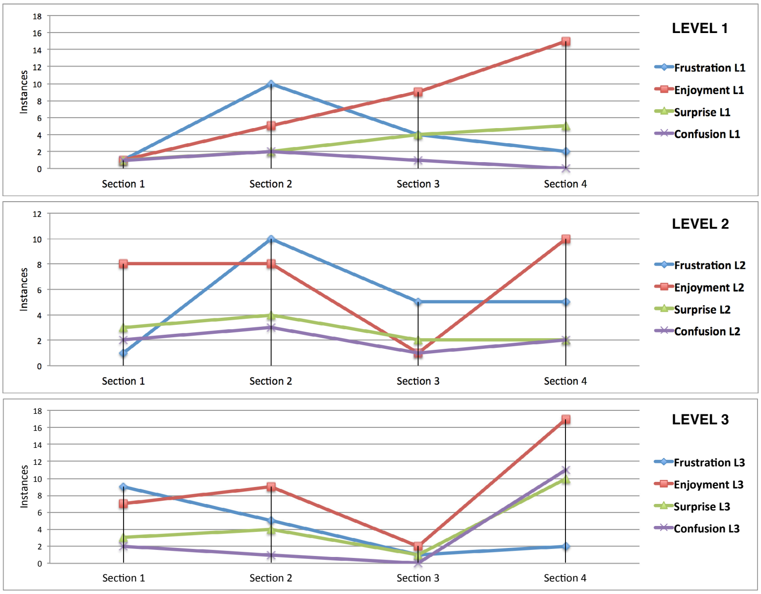

Fig. 4. The graphs show aggregated counts of instances in each of the three levels that during interviews have been described as frustrating, enjoyful, surprising or confusing. Note that the graphs are not in the same scale (specifically level 2)

frequency of themed instances (see Figure 4) for each level, data from the interviews could be used to highlight the need for improvements within the four sections of a level.

In order to determine which improvements to conduct, the filtered data was divided into actionable changes and sorted based on demand for change. Here we encountered situations that were mentioned positively by some participants and negatively by others. In general, uncontested changes were prioritized when looking for potential modifications.

Metrics Data. For the evaluation of metric GUR data we developed scripts that analyzed logs from the test sessions to derive aggregated measures of play statistics, such as amount of collected pick-ups, defeated enemies, etc. Where necessary, measurements were normalized in terms of time that was spent in the level, since the play duration had a direct affect on many play statistics. In addition to deriving information from the individual measurements, we calculated correlations of the acquired metric data. While these correlations could have been useful to uncover possibilities for the improvement of a level, we did not find actionable correlations. Apart from acquiring play statistics for each participant, the logs were used to create heatmaps (see Figure 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.

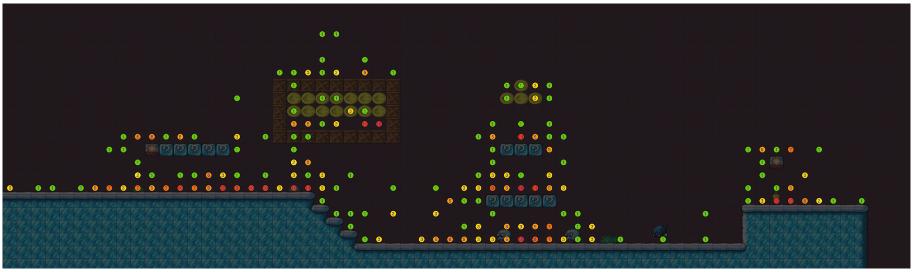

Fig. 5. Heatmap example showing part of a level overlayed with colored markers that indicate where in the level players switched their movement direction. Marker colors range from green (indicating a single event) to red (indicating the maximum amount of events in a level). As the level geometry is tile-based,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

Biometrics Data. For the evaluation of biometric GUR data, each level was divided into 12 sections of equal size in terms of quantity of horizontal modules. Since biometric data works with averages, we chose this number to keep a balance between getting useful as well as localized data. We programmed scripts to analyze the data and remove noise. 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 Figure 6. Similar to [6], we analyze biometrics data by investigating the signals and their relation to game events. Usually, we investigate relative changes in biometrics signals compared to the participant average. Although an initial classification of user mental states from combined biometric signals has been shown in the research literature [16] we do not think that this method is applicable to small samples of widely different games that we analyze.

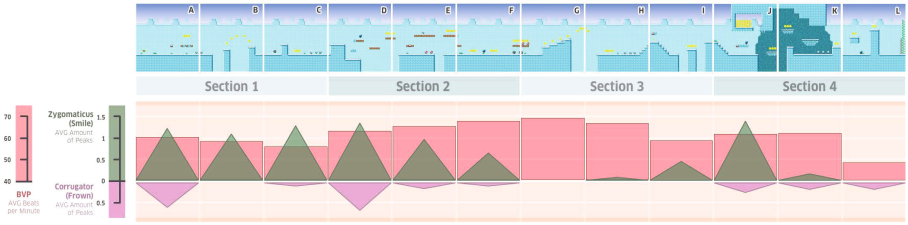

Fig. 6. Biometric data superimposed over map sections of level 3. The top row shows the level graphics split into the four sections used during interviews and 12 sections used for biometric GUR data. The red bars show the mean BVP of all participants in each section. The height of the green triangles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

8 Modifications (Phase 4)

In this phase we used the evaluated data and their visualizations to facilitate improvements in the level design. The first author, who has an industry background in game and level design, made all the level changes based on his perception on the shortcomings and the sub-optimal level design choices made for the game in relation to the target audience. This allowed for a consistent skill in terms of implementing modifications in the level. Furthermore, all modifications had to be connected to data that supports a change.

We looked at the three possible combinations of two GUR methodologies: (1) Interviews and Metrics; (2) Interviews and Biometrics; and (3) Metrics and Biometrics.

General level ratings (fun, length and difficulty of a level rated by participants) were added to each of the three combinations to contextualize the gathered data. For each of the three combinations we then decided to make the six most important changes across the three levels. Where we made the changes depended on where the data would support change.

We discuss the recommendations from the methodologies and the choice of implementation in detail elsewhere [17]. In this paper we explain the process of implementing the changes by one representative example. This particular change was made in the first level and was based on the combination of interview and biometric data:

For each change to a level, we started by inspecting the results of one of the two methodologies for issues that attracted comments from or were causing unwanted effects on the players. In this case biometric data showed very little player response throughout the level, especially in the beginning, as illustrated by low and steady BVP and lack of discernible facial emotion. We then checked this with the data from the other methodology to see if there was a common basis for a change. Interview data showed that the start of the level was rated consistently low across all measures (frustration, enjoyment, surprise and confusion). With the data from both methodologies, the designer (first author) concluded change in this area was desirable to induce higher player excitement and changed the level in a way that would solve the perceived problem without impacting the rest of the level: We decided to add more platforms and visual elements to give the section a more interesting look. The changes also encourage more action from the player if they want to collect all collectables, yet there is no increased risk of death, which would not be desirable at the beginning of the first level. An illustration of the changes can be seen in Figure 7.

The decision of how modifications should be implemented in terms of level design were taken based on the professional experience, sensibility, and best efforts of the level designer. All level changes are therefore of qualitative nature yet based on data from GUR methodologies that has been evaluated in a quantitative approach.

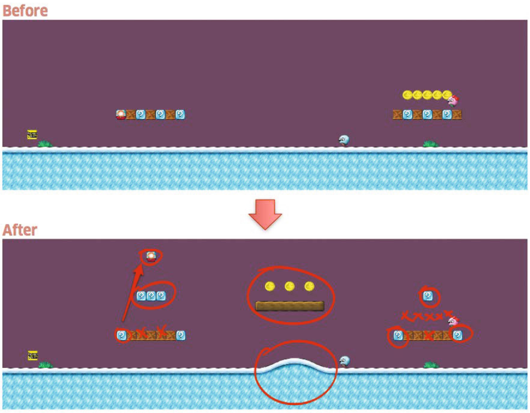

Fig. 7. Example of changes in a level segment that have been performed in response to interviews and biometric GUR data

9 Evaluation of the Modifications (Phase 5)

In the final phase of the experiment, new participants, chosen through convenience sampling 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 concluding test session.

Since some modifications in the level design tended to be subtle, we decided that participants should only play one of the level-sets with randomizing which of the level-set would be played. To compare the player experience between level-sets we used the Game Experience Questionnaire (GEQ, [18]), which has been developed to classify player experience in multiple fields. The questionnaire has been used in several publications within the academic field [19,5]. In our research we used two of the five GEQ modules: The 'Core' module and the 'Postgame' module, both of which were administered after having played through the levels. The GEQ contains a large number of questions, which are reduced to 11 dimensions through averaging. This procedure removes the noise that is associated with every single question and produces more reliable results. With the conclusion of the final test session, the results of the GEQ were calculated. We discuss these results in the next section.

10 Results

The graph shown in Figure 8 illustrates the results of the GEQ. 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 Interview & Metrics levels were rated most positively, with the Interview & Biometrics levels in second place and the Biometrics & Metrics levels last. This pattern is replicated for some of the negative dimensions Tension/Annoyance, Challenge and Negative Affect.

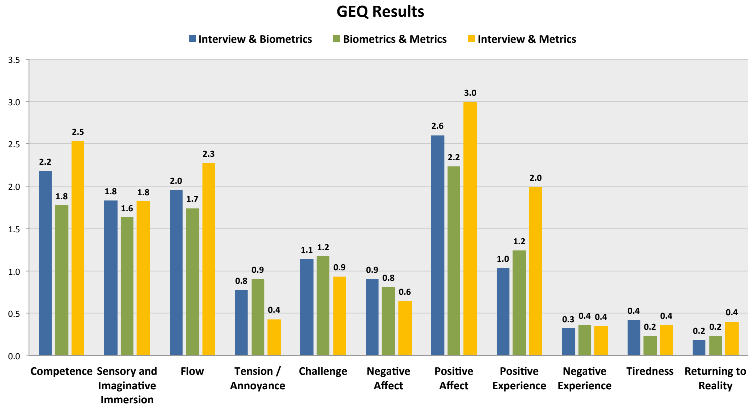

Fig. 8. Graph showing the GEQ scores of the individual methodology testing groups divided by the aspects that are scored by the GEQ

From the point of view of a level designer, each pair of GUR methodologies was able to provide action-able 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. As a result none of the combined methodologies could fully substitute another.

11 Discussion

Given the insight gained in the five phases of the research and the GEQ results of the individual methodologies, we feel that player interviews are essential to the success of improving level design and should therefore always be involved.

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 establish rough 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. From our own experience, 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 minimal. When aggregating over many tiles, the data becomes insightful, however it is difficult to derive specific level design recommendations from this.

While it would have been interesting to add a control group to the second testing round in form of a 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. For future comparisons of methodologies we do however advise to include such a control group in order to better evaluate the magnitude of improvement.

As a final point of discussion, we would like to address the possibility of replicating the processes of this study in 3-D games, which are arguably the majority of game titles nowadays. 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.

11.1 Limitations

Lead Researcher As Level Designer and Participant Observer. The author of this document was the lead researcher of this study as well as acting game designer and was therefore involved in all steps of the 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.

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 [6].

Combining Methodologies. Combining all methodologies or testing them separately could 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 methodology. However, it was beyond the scope of this study to compare seven versions of the same game.

12 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 method. From our research, in-game metrics seem to be the most useful addition. 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.

Acknowledgments. The authors wish to thank Isabelle Kniestedt, Meggy Pepelanova, the game design students participating in the focus group and all the testers who took part in our experiments; the RAAK International Funding body, NHTV Breda University of Applied Sciences, and the SuperTux development community for their support in this research.

References

1. Ernkvist, M.: 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, 161 (2008)
2. Wesley, D., Barczak, G.: Innovation and Marketing in the Video Game Industry: Avoiding the Performance Trap. Gower Publishing, Ltd. (2012)
3. Sheff, D.: Game Over: How Nintendo Zapped an American Industry, Captured Your Dollars, and Enslaved Your Children. Diane Publishing Company (1993)
4. Seif El-Nasr, M., Drachen, A., Canossa, A.: Introduction. In: Seif El-Nasr, M., Drachen, A., Canossa, A. (eds.) Game Analytics, pp. 3–13. Springer, London (2013)
5. Gualeni, S., Janssen, D., Calvi, L.: 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, pp. 149–155. ACM (2012)
6. Mirza-Babaei, P., Nacke, L.E., Gregory, J., Collins, N., Fitzpatrick, G.: How does it play better? exploring user testing and biometric storyboards in games user research (2013)
7. Drachen, A., Canossa, A.: Analyzing user behavior via gameplay metrics. In: Proceedings of the 2009 Conference on Future Play on GDC Canada, pp. 19–20. ACM (2009)
8. Nacke, L.E.: An introduction to physiological player metrics for evaluating games. In: Seif El-Nasr, M., Drachen, A., Canossa, A. (eds.) Game Analytics, pp. 585–621. Springer, London (2013)
9. Canossa, A.: Interview with nicholas francis and thomas hagen from unity technologies. In: Seif El-Nasr, M., Drachen, A., Canossa, A. (eds.) Game Analytics, pp. 137–143. Springer, London (2013)
10. Ambinder, M.: Valves approach to playtesting: The application of empiricism. In: Game Developer's Conference (2009)
11. Zammitto, V., Seif El-Nasr, M.: User experience research for sports games. Presented at GDC Summit on Games User Research
12. Seif El-Nasr, M., Desurvire, H., Nacke, L., Drachen, A., Calvi, L., Isbister, K., Bernhaupt, R.: Game user research. In: Proceedings of the 2012 ACM Annual Conference Extended Abstracts on Human Factors in Computing Systems Extended Abstracts, pp. 2679–2682. ACM (2012)

13. Mirza-Babaei, P., Long, S., Foley, E., McAllister, G.: Understanding the contribution of biometrics to games user research. In: Proc. DIGRA (2011)
14. Cacioppo, J.T., Tassinary, L.G., Berntson, G. (eds.): Handbook of Psychophysiology, 3rd edn. Cambridge University Press (2007)
15. Singleton, R.A., Straits, B.C.: Approaches to Social Research, vol. 4. Oxford University Press, New York (2005)
16. Mandryk, R.L., Atkins, M.S.: A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies. International Journal of Human-Computer Studies 65(4), 329–347 (2007)
17. Maureira, M.G.: Supertux a song of ice and metrics: Comparing metrics, biometrics and classic methodologies for improving level design. Master's thesis, NHTV University of Applied Sciences, Breda, the Netherlands (February 2013)
18. IJsselsteijn, W., de Kort, Y., Poels, K., Jurgelionis, A., Bellotti, F.: Characterising and measuring user experiences in digital games. In: International Conference on Advances in Computer Entertainment Technology, vol. 2, p. 27 (2007)
19. Nacke, L.: Affective ludology: Scientific measurement of user experience in interactive entertainment (2009)