SuperTux
A Song of Ice and Metrics

Comparing Metrics, Biometrics and Classic Methodologies for Improving Level Design

Bachelor Thesis

Marcello Gómez Maureira

March 2013

SuperTux – A Song of Ice and Metrics
Comparing Metrics, Biometrics and Classic Methodologies for Improving Level Design

Thesis Report

Marcello Gómez Maureira

Student at NHTV Breda University of Applied Sciences,
International Game Architecture and Design

Abstract

In this bachelor thesis the author explores the question of how the combination of individual game user research methodologies can be used to improve level design in a side-scrolling 2D platform video game. Three methodologies - classic methods, metrics and biometrics - are studied regarding their impact and viability in real world development scenarios. The study takes levels of the game SuperTux through five phases which function as milestones in an iterative quality assurance process. The study concludes that classic methodologies should be part of any effort regarding quality assurance. Other methodologies provide additional insights that can be used by designers to instigate further improvements. Especially biometric game user research methodologies require further development to make them viable in real world game development.

Acknowledgements

This thesis and the connected research has been my first expedition into the academic world. Coming from a design-centered background, many customs and procedures were unfamiliar to me. Luckily, I was not alone on my journey. Many people have offered their help along the way. It is for this reason that I have decided to use the first person plural in this document, highlighting the fact that this thesis has not been created in a proverbial bubble.

First and foremost I would like to thank **Stefano Gualeni**, my supervisor. Not only has he introduced me to the field of game user research, but also to game design at the beginning of my education at the NHTV in Breda. Without him I would not be where I am today, both on a professional and personal level.

One of the biggest influences in this thesis was without a doubt **Dirk Janssen** who has provided countless hours of time, care and general advice not only for the project but also for countless other aspects in my life. He nurtured my interest in technology and programming like no one else and the best thing I can do to honor him is to deepen my understanding of programming in general and of Python in particular.

A big thanks also goes to **Licia Calvi** who has supervised me prior to my graduation for more than a year and introduced me the research environment at the NHTV in Breda. Her positive nature has always been a great influence in that time.

I want to further thank **Meggy Pepelanova**, closest friend, soon to be Master, frequent comrade-in-arms and sometimes fierce discussion partner. Without her help this research would have continued well into the next decade and would still not meet the quality it does right now.

More thanks go out to **Isabelle Kniestedt**, my silly alter-ego and fellow course surfer. She was at my side for each step of this research - which is even more impressive given the fact that her fondness of the academic field should better remain unexpressed.

I also want to thank **Michelle Westerlaken** for her help and for being the best work partner I ever had. Her genuine kindness combined with the ease with which she can be scared with cheap parlor stunts ensure that each day around her is a great day.

Thanks also go to **Marta Clavero** with whom I initiated my journey into the academic world. While it is unfortunate that we could not travel together, I wish her all the best on her way.

Before coming to an end, I want to include **Bodo Schütze** and thank him for his presence in the world. We shared all the major milestones of our education - and I can only hope that this will continue to be the case for a while longer.

Finally my thanks go out to all participants that provided time and attention to this research. A special thanks goes to the participants of the focus group: **Alex Camilleri, Elroy Aarts, Jesse Ravensbergen, Mattia Traverso** and **Thomas Buijtenweg**.

Thank you everyone!

TABLE OF CONTENTS

1. INTRODUCTION 7

2. LITERATURE REVIEW 9

2.1 GAME USER RESEARCH (GUR) 9

2.1.1 Classic User Research Methods 10

2.1.2 Metrics 11

2.1.3 Biometric Methods 12

2.1.4 Combining Methods 15

2.2 CONSIDERATIONS FOR GAME DEVELOPMENT 17

2.3 MEASURING USER EXPERIENCE 18

3. RESEARCH GOALS 20

3.1 BRIDGING ACADEMIA AND GAME DEVELOPMENT 20

3.2 NEUTRAL EVALUATION OF INDIVIDUAL GUR METHODOLOGIES 20

4. RESEARCH FOCUS 21

4.1 2D LEVEL DESIGN ORIENTED 21

4.2 PLATFORM VIDEOGAME: SUPERTUX 21

5. RESEARCH DESIGN: COMPARATIVE DESIGN WITH MIXED METHODS 23

5.1 QUASI-EXPERIMENT 25

5.2 FOCUS GROUP 25

5.3 GAME USER RESEARCH (GUR) METHODOLOGIES 26

5.3.1 Classic 26

5.3.2 Metrics 26

5.3.3 Biometrics 27

5.4 GAME EXPERIENCE QUESTIONNAIRE (GEQ) 28

6. PHASE 1: PREPARATIONS FOR SIMULATING A QA ITERATION 30

6.1 LEVEL DESIGN OF THE INITIAL LEVEL-SET 31

6.2 INITIAL QA - A FOCUS GROUP WITH DESIGNERS 31

6.2.1 Focus Group Design and Environment 31

6.2.2 Sampling: Focus Group Participant Selection 32

6.2.3 Focus Group Procedures 32

6.2.4 Focus Group Results 33

6.3 ADDING LOGGING FUNCTIONALITY TO THE GAME 34

6.4 FORMULATING A QUESTIONNAIRE FOR PLAYER INTERVIEWS 35

6.5 CONCLUSION AND DISCUSSION OF PHASE 1 36

7. PHASE 2: TESTING ROUND 1 - COLLECTING GUR DATA 37

7.1 TESTING DESIGN AND ENVIRONMENT FOR ROUND 1 37

7.1.1 Biometric Sensors 39

7.2 SAMPLING: PARTICIPANT SELECTION FOR TESTING ROUND 1 40

7.3 TESTING PROCEDURES FOR ROUND 1 41

7.4 CONCLUSION AND DISCUSSION OF PHASE 2 42

8. PHASE 3: DATA EVALUATION AND VISUALIZATION 44

8.1 A CASE FOR VISUALIZATION OF DATA FOR DESIGNERS 44

8.2 CLASSIC GUR DATA 45

8.2.1 Interpretation of Classic GUR Data 45

8.2.2 Visualization of Classic GUR Data ... *46*
8.2.3 Results of Classic GUR Data ... *47*
8.3 METRIC GUR DATA ... 48
8.3.1 Interpretation of Metric GUR Data ... *48*
8.3.2 Visualization of Metric GUR Data ... *48*
8.3.3 Results of Metric GUR Data ... *50*
8.4 BIOMETRIC GUR DATA ... 50
8.4.1 Interpretation of Biometric GUR Data ... *50*
8.4.2 Visualization of Biometric GUR Data ... *51*
8.4.3 Results of Biometric GUR Data ... *52*
8.5 CONCLUSION AND DISCUSSION OF PHASE 3 ... 53

9. PHASE 4: LEVEL-SET MODIFICATIONS ... 55
9.1 MODIFICATIONS BASED ON CLASSIC AND METRIC GUR DATA ... 55
9.1.1 Classic & Metric Modification 1 (Level 1 - Section 2) ... *55*
9.1.2 Classic & Metric Modification 2 (Level 2 - Section 1) ... *59*
9.1.3 Classic & Metric Modification 3 (Level 2 - Section 2) ... *60*
9.1.4 Classic & Metric Modification 4 (Level 2 - Section 3) ... *63*
9.1.5 Classic & Metric Modification 5 (Level 2 - Section 3) ... *65*
9.1.6 Classic & Metric Modification 6 (Level 3 - Section 2/3) ... *66*
9.2 MODIFICATIONS BASED ON CLASSIC AND BIOMETRIC GUR DATA ... 67
9.2.1 Classic & Biometric Modification 1 (Level 1 - Section 1) ... *67*
9.2.2 Classic & Biometric Modification 2 (Level 1 - Section 1/2) ... *68*
9.2.3 Classic & Biometric Modification 3 (Level 2 - Section 2) ... *70*
9.2.4 Classic & Biometric Modification 4 (Level 2 - Section 2) ... *71*
9.2.5 Classic & Biometric Modification 5 (Level 2 - Section 4) ... *72*
9.2.6 Classic & Biometric Modification 6 (Level 2 - Section 4) ... *73*
9.3 MODIFICATIONS BASED ON BIOMETRIC AND METRIC GUR DATA ... 74
9.3.1 Biometric & Metric Modification 1 (Level 1 - Section 1/2) ... *74*
9.3.2 Biometric & Metric Modification 2 (Level 1 - Section 4) ... *76*
9.3.3 Biometric & Metric Modification 3 (Level 2 - Section 1) ... *77*
9.3.4 Biometric & Metric Modification 4 (Level 2 - Section 2) ... *78*
9.3.5 Biometric & Metric Modification 5 (Level 2 - Section 4) ... *79*
9.3.6 Biometric & Metric Modification 6 (Level 2 - Section 4) ... *80*
9.4 CONCLUSION AND DISCUSSION OF PHASE 4 ... 80

10. PHASE 5: TESTING ROUND 2 - EVALUATION AND COMPARISON OF LEVEL-SET MODIFICATIONS BASED ON GUR ... 82
10.1 TESTING DESIGN AND ENVIRONMENT FOR ROUND 2 ... 82
10.2 SAMPLING: PARTICIPANT SELECTION FOR TESTING ROUND 2 ... 83
10.3 TESTING PROCEDURES FOR ROUND 2 ... 83
10.4 CONCLUSION AND DISCUSSION OF PHASE 5 ... 84

11. RESEARCH RESULTS ... 85

12. RESEARCH LIMITATIONS ... 86

13. RESEARCH CONCLUSION AND DISCUSSION ... 88

14. BIBLIOGRAPHY ... 90

15. APPENDICES ... 93

1. INTRODUCTION

With the North American video game crash of 1983, the video game medium had its historical precedence for what it means if low quality products saturate a market (Wesley & Barczak, 2010). The 1983 crash refers to a massive recession in the North American video game market that has been connected to a general lack of quality control. Since then, the game industry has come a long way and quality assurance (QA) is typically an essential phase of every commercial video game release. Often it is part of an ongoing iterative process with the goal of ensuring that the intentions of the underlying game design, its rules and framework are successfully conveyed to the player.

While the fundamental purpose of QA in video games is simple, the ways in which it can be implemented are not. Traditionally, QA in video games includes observing and interviewing players after letting them playtest the game (Ambinder, 2009). Additional information can be acquired by applying metrics, which, at its core, is the process of quantifying game states, player interactions or similar parameters.

Game user research (GUR) is the academic treatment of QA. A relatively new addition to GUR methodologies is biometric testing. In this method of research data is gathered from the play tester by using sensors, which can monitor for example muscle activity, heart rate and skin conductivity. In recent years biometric testing has become available to companies within the game industry and several development studios have added the methodology to their QA efforts. While it has proven useful in balancing the structure and difficulty of a game, 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.

This study aims to give an insight into three GUR methodologies: the 'classic' QA methods of observation, interview and questionnaire, data collection through metrics and data collection through biometrics, and how these can be used to improve level design in a 2D platform game. A combined reading of these measures in relation to game events can be used to shed light on the experience of the player.

For this study the decision was made to focus on level design. This does not mean though that the GUR methodologies tested in this research cannot be used for other areas within the game industry. The reason level design was chosen for the research is because it is modularly adjustable (especially in a tile-based game), easy to produce, and provides a clear basis for comparison among level-sets. It is still not possible to eliminate all variables, but it provides a more controlled testing environment since the levels can be relatively tightly controlled. There are only a few studies that have used

GUR specifically for level design purposes, and this is part of the contribution to knowledge that this research aims for.

This study aims to answer the question:

How can combinations of GUR methodologies be used to improve level design in a side-scrolling 2D platform game?

While the focus of this study is on evaluating the selected GUR methodologies, it is also intended to serve as a guide to any game developer planning to implement GUR in their development process. Therefore, this study also seeks to assess the methodologies in terms of the practical considerations involved in data acquisition and interpretation. We outline the shortcomings as well as the strengths of the different methods, as well as how they could contribute to the game development process.

Working on the casual game project *Gua-Le-Ni: The Horrendous Parade* (Gualeni & Double Jungle SaS, 2011) demonstrated to the author the potential of implementing GUR in general, and biometric testing in particular. It became apparent, however, that there is a need for more knowledge on how this methodology can effectively be used. Rather than assuming that biometric testing will supersede other methods, the aim of this study is to provide a balanced examination of possible combinations of methods.

2. LITERATURE REVIEW

Answering the research question in this study requires building on the current academic framework of GUR. The following literature review lays out the known characteristics and uses of the classic GUR methodologies that have been selected to reflect common industry practice. It then offers a brief overview on what research has been done in applying metric and biometric testing to games. This chapter of the report aims to demonstrate how the kinds of insights offered by the various GUR methods are known to differ, as well as to show the practical research considerations that go into planning game user testing.

In order to compare the effectiveness of the different GUR methodologies in contributing to better level design, we need to compare the quality of their actionable insights. This can be done in an indirect manner by comparing the quality of levels that the different GUR methods have been employed to modify. There are many competing theoretical perspectives about what constitutes a 'good' game, and there is no single measure or standard by which the quality of level design is measured. This study uses player experience as a determinant of level quality, and this section of the document provides some theoretical background to this choice.

2.1 GAME USER RESEARCH (GUR)

Game user research (GUR) is a relatively recent field of research, which draws upon theories and methodologies from Human Computer Interaction (HCI) and Experimental Psychology to study digital games (El-Nasr et al., 2012). 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.

Insofar as a video game is a software application, it can be tested using traditional usability techniques, but specifically designed heuristics are needed due to the fact that the goals of a game are different than that of productivity software (Pagulayan, Keeker, Wixon, Romero, & Fuller, 2003). To determine the quality of a game, however, it is not enough to measure traditional usability outcomes such as task completion times or errors. It is also necessary to measure the subjective player evaluation of the game. This is where GUR comes in to complement traditional QA by investigating player behavior and the corresponding attitude and motivation.

Nacke et al. (2009, 1) make a similar distinction when they differentiate between playability and player experience research. They argue that "playability methods evaluate games to improve design, whereas player experience methods evaluate players to improve gaming".

GUR requires that existing tools and techniques in HCI developed for virtual environments or interactive systems be extended and adapted to the unique requirements of games. 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 (El-Nasr et al., 2012).

2.1.1 Classic User Research Methods

Classic methods of evaluating the user experience of game players as part of the game development process include objective reports through observation and subjective reports through questionnaires, interviews, and think-aloud protocols (Mirza-babaei, Long, Foley, & McAllister, 2011).

2.1.1.1 Interviews and Questionnaires

Self-reports through interviews and questionnaires are the most common way of assessing the player experience. Since answering questions constitutes a disruption of gameplay, interviews and questionnaires are used to obtain subjective information about the user after the playtest is completed. Usually impressions and opinions obtained this way from players will be hard to relate to any specific moment of gameplay or any single design element. The post-factum nature of this information also means that it will reflect the experience as a whole and will be subject to memory effects and post-rationalization. Important details may be forgotten or neglected because of this (Mandryk & Atkins, 2007). To mitigate these effects, the memory of players is often prompted with a video recording of the play session (Gow, Cairns, Colton, Miller, & Baumgarten, 2010; Mirza-babaei, Nacke, Gregory, Collins, & Fitzpatrick, 2013).

The benefits of interviews and questionnaires include being convenient, relatively quick and easy to administer, and generalizable in their findings (Lewis-Evans, 2012a). In addition, while players might be unreliable in reporting their own behavior, self-report techniques do provide a good understanding of player attitudes (Pagulayan et al., 2003).

Questionnaires can provide quantitative results that lend themselves to statistical analysis, especially in cases where a consistent comparison is desirable. For optimum effect, this requires sample sizes large enough for statistical validity. By contrast, interviews offer rich qualitative data about gameplay, and offer the opportunity for follow-up questions and unexpected directions that questionnaires do not (Lewis-Evans, 2012a). A feature of interviews that is useful for the development process is that they offer good quotes or sound bites from players that can be used to communicate key messages to the development team concerning the outcomes of the test (Lewis-Evans, 2012a). The drawback in using interviews is that they can be time-consuming and hard to generalize beyond the individual experience.

2.1.1.2 OBSERVATION

Observation is a method that involves watching the user play the game and making note of their in-game behavior, as well as their body language and facial expressions. Observation sessions have the advantage of being easy and inexpensive to organize. They provide rich, objective data about how players actually interact with the game (as opposed to how they think, or say, they do).

On the downside, observation sessions can be difficult to analyze. For best results, observation sessions require an observation schedule, an experienced observer and software that supports video recording and coding. Without a video recording, the observer can miss important details or neglect the context in which events occur. Interpreting the data gathered can be time-consuming and runs a high risk of bias (Mirza-babaei et al., 2011).

2.1.2 METRICS

Game metrics are instrumentation data about user behavior and user-game interaction obtained from the game software itself (Drachen & Canossa, 2009). Any action that the player takes within the game can potentially be recorded, from low-level data such as button presses to in-game interaction data of the player character. The great advantage is that this data can be collected in large volumes, at the desired degree of temporal resolution, and is directly mapped to the specific sites of interaction in the game (Tychsen, 2008). Metrics scale well with the addition of player data, since the results are typically aggregated amongst all players or predefined player groups. Collecting the data is non-obtrusive to the player, and can be aggregated to reveal patterns or trends.

The process starts with acquiring raw metric data, which is stored and later processed according to various factors of interest. Variables of relevance could, for example, be percentage of map explored or number of deaths before a checkpoint. Metrics can then be used to draw 'heatmaps', which plot player data over a level map with changes in hue. This approach allows designers to visually connect statistic measures with locations in the game. Expressing data in heatmaps is therefore especially useful in regards to level design.

There is no standard procedure for metrics output or analysis, in large part due to the differing demands of individual games and the fact that many metrics systems are industry secrets. In general, transforming raw metrics data into useful reports involves filtering and structuring the data, analyzing, and then visualizing. This process can undergo iterations to improve precision, presentation or customization.

The limitations of metrics are that they show nothing of the internal state of the user, in either cognitive or emotional aspects. Also, a minimum amount of player data is often required before the interpretation of aggregated data is statistically relevant. The nature of metric data does not allow players to express their preference for options that are not already present in the game. Furthermore, metrics cannot account for factors of influence outside of the game software itself, such as social or contextual. The resulting data can also be overwhelming in quantity, time-consuming to interpret, and without specific points of interest can be hard to derive actionable insights (Lewis-Evans, 2012b).

2.1.3 BIOMETRIC METHODS

The potential of biometric (also referred to as psychophysiological) testing for GUR has led to the method quickly being recognized as integral in the future of the field (Mirza-babaei et al., 2011). As the study of emotional engagement has become popular in video game research in the last decade, many studies have used physiological sensors as evaluation tools. Kivikangas et al. offer a detailed review of the current state of biometric methods in game research (Kivikangas et al., 2010). For the purposes of this study, the basic theory behind the method will be outlined before focusing on the practical implications.

In the field of games, commonly used biometric measures to evaluate emotional experience are facial electromyography (EMG), cardiovascular measures such as interbeat intervals, and skin conductance. Electroencephalography (EEG) has relevant applications but is rarely used because of its relatively complex analysis procedure. Combinations of these measures have been used by researchers to create an emotional profile of players during gameplay (Mandryk & Atkins, 2007; for example Ravaja, Saari, Salminen, Laarni, & Kallinen, 2006).

Two different approaches to biometric analysis have emerged in terms of the temporal dimension – tonic and phasic analysis (Nacke, 2011). Phasic analysis studies physiological responses at specific points in time, usually by comparing the second of onset of a game event with the following seven seconds. Tonic analysis studies variations over a time span, by normalizing and averaging the data over a period of gameplay and comparing average values of physiological activation within it.

The use of physiological signals to study psychological phenomena was established in the field of psychology. The mapping of emotions to the two dimensions of valence and arousal was elaborated in numerous studies since its introduction in 1985 (Watson & Tellegen). This insight was used in subsequent studies that employed electromyography (EMG) to explore the physiological responses of facial muscles. Subtle increases in activity in certain facial muscles was found to correlate to particular emotions. In particular, eye brow muscle activity (corrugator supercilii) was found to increase when a person is in a negative mood (Larsen, Norris, & Cacioppo, 2003), while cheek muscle activity (zygomaticus major) was found to increase when a person is in a positive mood. The mapping of physiological data to discrete emotions is not straightforward, but it is possible to classify emotions during gameplay in this two dimension model of valence and arousal with the help of surface electrodes on the muscles of the brow and cheek (Mandryk, Atkins, & Inkpen, 2006).

For an accurate assessment of arousal, however, an additional measure of the player's electrodermal activity (EDA) is necessary. This can be done through galvanic skin response (GSR) sensors positioned on the palms of the hands. GSR is a measure of the electrical conductance of the skin. The sweat glands in the palms change skin resistance based on psychological stimulation (Mandryk et al., 2006) in a manner accurately correlated to arousal (Ravaja et al., 2006). The terms EDA and GSR are sometimes used interchangeably, but a distinction exists based on whether skin conductance is analyzed at a conductance level or in terms of discrete conductance responses (Kivikangas et al., 2010). Electrodermal responses have a delay of one to four seconds, which needs to be taken into account, but in general EDA is a sturdier and less ambiguous measurement than EMG or heart activity.

Heart rate is also indicative of emotional arousal and stress, but because so many biological processes influence heart rate it can be hard to isolate the particular cause for changes (Nacke, 2009). Cardiac activity is nevertheless among most widely used biometric measures and has been successfully deployed in game studies (Kivikangas et al., 2010). Cardiovascular measures include heart rate, interbeat intervals, heart rate variability, blood pressure and blood volume pressure (BVP).

The measures captured from biometric testing are continuous and involuntary, which offers access to players' true psychological states during every moment of gameplay (Mirza-babaei et al., 2011). This circumvents the difficulty self-report methods face in relying on players to be aware of their internal states, to remember them and to verbalize them. This direct, almost unmediated, access to the internal experience of the player is the main benefit that biometric testing has to offer, even if it can be hard to tell **what** exactly that experience is. It cannot be said, however, that biometric testing measures the player experience itself – it only deals with the part of it that has a recognized physiological expression.

Biometric testing has the benefit of removing a lot of the subjectivity from the data collection process, although the results still need to undergo interpretation by the researchers. The involuntary nature of the measurements means that the data is not contaminated by interpretation, social desirability, miscommunication or participant answering style (Kivikangas et al., 2010).

Another strength of the biometric method is the high precision and sensitivity of the data collected. For instance, the emotion of players during gameplay can be evaluated by observing their facial expressions. The advantage of using EMG for the same purpose would be a temporal precision measured in milliseconds, and the detection of covert muscle activity that is invisible to the human observer (Kivikangas et al., 2010). Furthermore, the process is automatic and free from observer bias. The real-time, automatic nature of biometric data collection also means that the playing session of participants does not need to be disrupted with questions.

The method is not without considerable challenges, however. Physiological processes typically do not have a one-to-one relationship with psychological phenomena (Kivikangas et al., 2010). A single physiological response can be linked to number of psychological causes, and to different structures such as attention or emotion (Cacioppo, Tassinary, & Berntson, 2007). It can therefore be difficult for researchers to understand exactly to what facet of the player experience the quantitative data obtained from the play session points. This is compounded by the fact that the biometric measurements will reflect not only the player's response to in-game stimuli, but indiscriminately also their response to stimuli external to the game, such as conversation, disruption, or even anticipation (Gow et al., 2010).

Kivikangas et al. (2010) summarize the complexity of successfully performing biometric research regardless:

> The practical challenge is to identify research questions that can be answered even when the game is complex and the psychological processes numerous, and then to create an experimental design with the proper and necessary controls so that no confounding variables will affect the results. (Kivikangas et al., 2010, 1)

Note: The APA referencing style of page numbers has been modified here in response to supervisor preferences.

The flipside to the high resolution of the data that biometric sensors provide is that noise is a significant concern. For instance, facial EMG measurements are sensitive to both technical noise (i.e – electrode contact on the skin) and noise from unrelated activities, such as speaking (Kivikangas et al., 2010). In order to get meaningful results about the game despite the presence of noise and the complex psychological causation, either the game used to test must be very simple, the studied phenomenon of interest must be very strong or the sample size needs to be rather large (Kivikangas et al., 2010). Among other things, this means that biometric testing is at this moment less suited to handle games that involve subtle gameplay (Gualeni, Janssen, & Calvi, 2012), as opposed to the visceral gameplay elicited by action or horror games.

From a practical perspective, it can be said that a disadvantage of using biometric methods is the significant time and money the testing setup requires. The sensor equipment tends to be expensive and requires careful maintenance. Experimental procedures and data analysis require specially trained researchers. Furthermore, using sensors on the experiment participants will restrict the use of some body movement. This means that games that have particular interface or control requirements (such as using both hands, or using motion controls), may be incompatible with the use of some types of sensors.

2.1.4 COMBINING METHODS

The methods outlined in the previous sections are rarely used on their own. It is both industry and research practice to use several methods in combination in such a way that the strengths of one method offset the shortcomings of another. For example, metrics show the behavior of a player (what they did), but not motivation (why) or emotion (how they felt). Therefore, metrics can be used to supplement interview and observation data by tying the reported player experience to specific game design features (Tychsen, 2008). Similarly, to fully understand biometric results, it is necessary to correlate the physiological responses to the subjective experience report of the player (Nacke, 2009). To illustrate the point with an example: this correlation helps researchers to distinguish whether stress that a player experiences is welcomed (the player is looking for challenge), or unpleasant (the player is looking to relax). Conversely, biometrics and observation can help prevent the memory bias of self-report (Mirza-babaei & McAllister, 2011). Self-reporting suffers from a 'serial positioning effect', where players are more likely to remember what happened at the beginning and the end of an experience than what happened in the middle.

Recent studies in the field of GUR have highlighted the need for future research into a better understanding of the value of the different testing methodologies relative to each other. A 2011 study (Mirza-babaei et al.) compared the data obtained from traditional observation-based methods with that of biometric methods (using GSR only). Their 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 GSR 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 (Mirza-babaei et al., 2011).

Other studies have also hinted that different testing methods are suitable for different kinds of game development goals. It has been suggested that biometric testing is useful for adjusting level design and difficulty (Gualeni et al., 2012; Mirza-babaei et al., 2013), while classic methods seem to elicit more drastic gameplay and mechanics change recommendations (Mirza-babaei et al., 2013).

Mirza-babei et al. (2013) performed another study in the same direction, which is strongly in line with the aims of this thesis. 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 a 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” (Mirza-babaei et al., 2013, 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 the designers working with Classic UT or no UT. The designers using BioSt UT data approached the design task in a more player-centered way, while those using Classic UT focused more on the problematic design issues themselves.

Another key point in the paper is evaluating GUR methodologies from within the game development context, where a number of different professionals will be involved the process of testing the game, proposing design changes and implementing them (QA staff, designers, programmers). A challenge in this process is communication and persuasion between these different parties. Designers, for instance, can be distrustful of the results of GUR and may resist changes, especially if they feel it criticizes their work. Mirza-babei et al. (2013) suggest in the results of their study that due to the player-centric nature of biometric data, biometric user testing increases the trust of the designers in the reported issues and helps them communicate more constructively and persuasively with the rest of the team.

2.2 CONSIDERATIONS FOR GAME DEVELOPMENT

Understanding and improving the game experience of users is of particular interest to the commercial development of games (Leone, 2012). User-oriented testing has emerged as a means to gain competitive advantage in the increasingly saturated game market (Lameman et al., 2010). Key industry players have set up internal user testing departments, such as Microsoft’s Game User Research labs (Leone, 2012). Meanwhile, academia has amassed decades of research into game evaluation and user experience. There is obvious potential in collaboration between industry and academia, but the different priorities and constraints of each can be in conflict. Game companies are not interested in the understanding of games and their users for its own sake, but are looking for efficient formative evaluation and, ultimately, a return on investment (Lameman et al., 2010).

GUR is particularly well suited to serve as a middle ground between industry and academia, by providing information about the game experience of users in a way that can help professionals make better design decisions. On part of the researchers, developing tools that professionals can use and presenting research results in an accessible way are important prerequisites for this cooperation. BioSt is a technique for presenting user testing information that has been developed precisely with the needs of developers in mind (Mirza-babaei & McAllister, 2011). Cross-referencing the data from a mixed methods approach and presenting it in a visually informative way makes it easier for the designer to utilize research outcomes. From the point of view of game developers, actionable research results are not only those that are clear and simple to use, but also those that resonate equally well with the different professionals in a game development team. This is illustrated in an example in the previous sub-chapter of this document.

A primary consideration for game developers with regards to user testing is the cost and duration. Companies need game improvements to be as cheap and rapid as possible, and thus GUR needs to fit into the short iterations that characterize game production (Gualeni et al., 2012). Any delay between the implementation of a feature and the GUR feedback on that feature will make it less likely that the feedback is implemented (Medlock, Wixon, & Terrano, 2002). There needs to be a high cost-benefit ratio between the amount of actionable information a research method produces and the effort required to perform it. For example, hand-coding information or going through lengthy video capture will have only limited use in situations where a quick and effective testing process is required (Tychsen, 2008).

For developers the important thing about GUR is not the number of game problems it exposes. It is important that reported problems are plausible to the relevant decision-makers, that their relative importance to the success of the game is clear, and that they have specific proposed solutions (Medlock et al., 2002). For example. Microsoft's Research Labs are fitted with cameras that record players' facial expressions, but this data is not used. Senior user researcher Eric Schuh says, "We're not just here about collecting information to understand things. We're here about making the game better by having actionable data, and knowing someone smiled at four minutes and 38 seconds into a mission doesn't help us out a lot" (Leone, 2012).

2.3 MEASURING USER EXPERIENCE

User experience is a nebulous concept that has been the focus of much research and debate. Due to the large variability of gaming activity in terms of context, players, and gameplay, there is currently no unified model of user experience and no definitive way to test its quality (Tychsen, 2008). What exists instead is a huge variety of "vaguely defined experiential phenomena, such as immersion, flow, presence and engagement" (Nacke, 2009). A problem with these concepts that is relevant to GUR is that none of them provide a consistent framework for the experience of negative affect (which is reflected in biometric measures of stress and frustration) (Tychsen, 2008).

The absence of an established framework for user experience is problematic when it is required to compare the experience derived from two different games. In order to provide a consistent basis for comparison, game research studies in the past have used the game experience questionnaire (GEQ) (IJsselsteijn, De Kort, & Poels, 2007). The GEQ covers different components of the subjective game experience: immersion, tension, competence, flow, challenge, positive affect and negative affect. Each of these components consists of 5–6 questions (e.g., "I had to put a lot of effort into the game") rated on a five-point scale. The scale ranges from a 0 (not agreeing with the statement) to 4 (completely agreeing with the statement). The GEQ was applied in a number of experimental research projects that testified to the measure’s test-retest reliability and construct validity (IJsselsteijn, van den Hoogen, & Klimmt, 2008). The GEQ is not formally overall accepted as the way to self-report somebody’s game experience, but has become a benchmark due to its validation against psychophysiological measurements (Gualeni et al., 2012; Nacke et al., 2009).

3. Research Goals

In the pursuit of addressing the research question, this study seeks to fulfill two main research goals, which can be considered as the core motivation behind all efforts:

3.1 Bridging Academia and Game Development

As it is mentioned in the introduction and further illustrated within the theoretical framework, difficulties can emerge in reconciling academic findings in GUR with the production environment of game development. Such difficulties can take many shapes. One example is the strict time schedule of commercial game developments that can make it difficult to implement emerging GUR methodologies into the development process. Another example is the traditionally intuitive and heuristically supported work mentality of designers, which can make an analytic approach seem detached and at odds with the creative aspects of design.

This study seeks to bridge the efforts of GUR with the production-oriented side of game development for the purpose of improving the ability of game designers to make informed decisions. To this end, conclusions and discussions in this study evaluate not only academic implications but also implications that are relevant for design or production.

3.2 Neutral Evaluation of Individual GUR Methodologies

While it stands to reason that each and every academic effort should be conducted with a neutral perspective, it should be emphasized that this study does not argue for or against the use of any specific GUR methodology. Rather, we want to explore and dissect the impact of GUR methodologies, or more specifically the combination of methodologies, in a simulated game production environment.

4. RESEARCH FOCUS

As tempting as it might be to attempt a research with the goal of improving all kinds of games regardless of its genre, this study has to focus its efforts on a subsection of design challenges.

4.1 2D LEVEL DESIGN ORIENTED

This study focuses on the impact of GUR methodologies within the iterative process of designing 2D video game levels. Since GUR methodologies cannot easily be used in all parts of a development process, we put our focus on a production phase in which the core mechanics and rules are to a large extent determined and tested: level design.

To this author's knowledge, most if not all the current research around GUR focuses on general game design challenges, while research into level design is predominantly concerned with the possibilities of procedural level creation. We therefore consider the specific focus of this study a unique addition to the research field.

While it would be interesting to conduct this study on both 3D and 2D video games, this research focuses on 2D video games which typically tend to feature less complex level designs than those taking place in 3D space. This is easily explained by having one spatial axis less to take into account. Generally, video games played in 2D space also require less expertise in terms of control, which allowed for a broader selection of participants.

4.2 PLATFORM VIDEOGAME: SUPERTUX

As basis for our game research we chose *SuperTux,* a side-scrolling 2D platforming game developed by an 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 (i.e. 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 and blocks in mid-air that can be traversed by the player. 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.

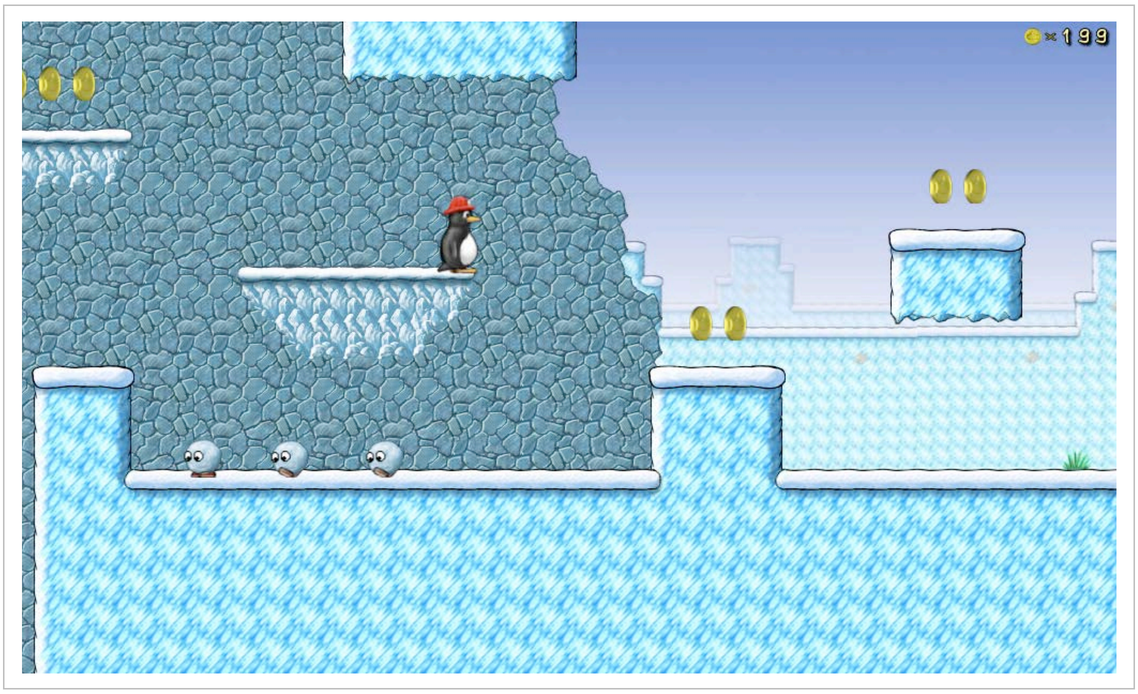


Fig.1: A screenshot of SuperTux showing Tux, the protagonist, in upgraded form (with red helmet), three enemies, and several bonus coins

While the concept of navigating a virtual character through a virtual space must have been an unusual concept to comprehend in the early beginning of video games, we argue that today most people have a basic comprehension of basic video game principles such as a moving camera or the control of a virtual avatar on screen. The fact that *SuperTux* can be played with a basic understanding of controlling a virtual character is one of the reasons for its use in this study.

Further reasons for the use of *SuperTux* are:

- The game comes with a level editor, allowing for easy modification of levels.
- Tile-based nature of the game makes modifications modular and therefore easier to modify discretely.
- The game is freely available and can be modified by anyone due to its open source roots. 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.
- Audiovisual assets in the game are sufficiently high in quality and more importantly consistent in terms of detail.

5. Research Design: Comparative Design With Mixed Methods

Quality Assurance (QA) is an iterative process in the development of video games (Medlock et al., 2002). Information that is acquired by applying GUR is used to improve the game during QA phases. Such improvements are subsequently subjected to yet another phase of applying GUR. This cycle (see *Figure 2*) of modification and analysis continues until the design goals are met, or more accurately, met closely enough in regards to other contributing factors such as the development budget.

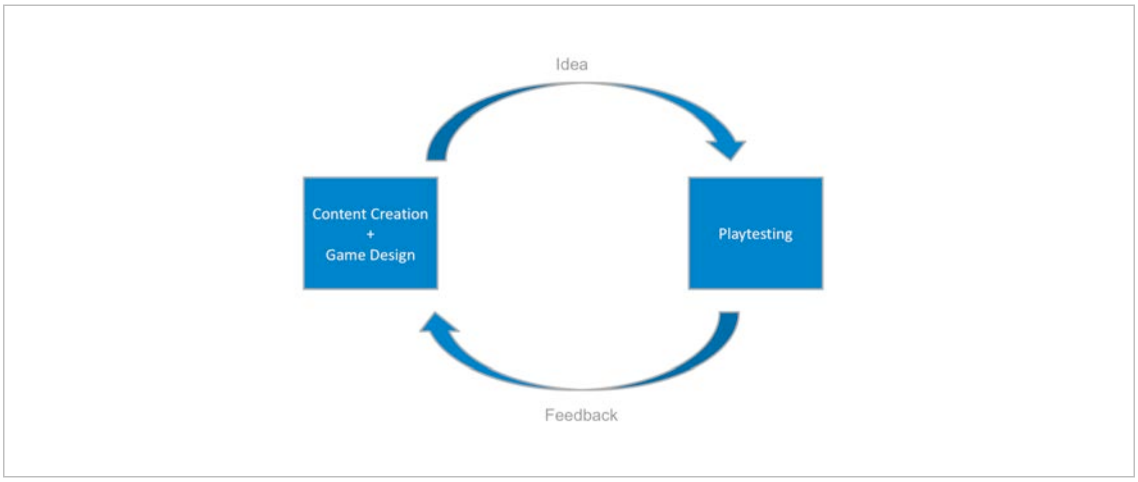


***Fig. 2:* Feedback loop at Valve – Presentation Slide from 'Valve's Approach to Playtesting: the Application of Empiricism' at GDC 2009 (Ambinder, 2009)**

While modifications are motivated by GUR, their implementation, in itself, is in need of evaluation. Regardless of which parameters are used to evaluate the success of a modification, a comparison needs to take place. By comparing the state of the game before introducing modifications and after, game designers ensure that changes are in fact leading to a differing outcome as well as to an outcome that follows the targeted design goals.

To emulate the comparative aspect of QA processes in the game industry, this research uses a **comparative design with mixed methods.**

Bryman (2008) describes comparative research design as:

> A research design that entails the comparison of two or more cases in order to illuminate existing theory or generate theoretical insights as a result of contrasting findings uncovered through the comparison. (Bryman, 2008, 692)

As will be illustrated in the following sub-chapters of this document, the comparative design of this study will entail comparing between level-sets that have been modified with a different combination of GUR methodologies. The first phase of the research involves creating level-sets. The second phase aims to gather GUR data by means of 'Testing Round 1'. The third phase is the evaluation and the processing of the gathered data. The next phase involves implementing changes using three different combinations of GUR methods into three different level-set versions. The final phase features 'Testing Round 2', and a comparison of the different level-sets. More explanation on each phase follows in the rest of the document. Refer to *Figure 3* for a visualization of the research phases.

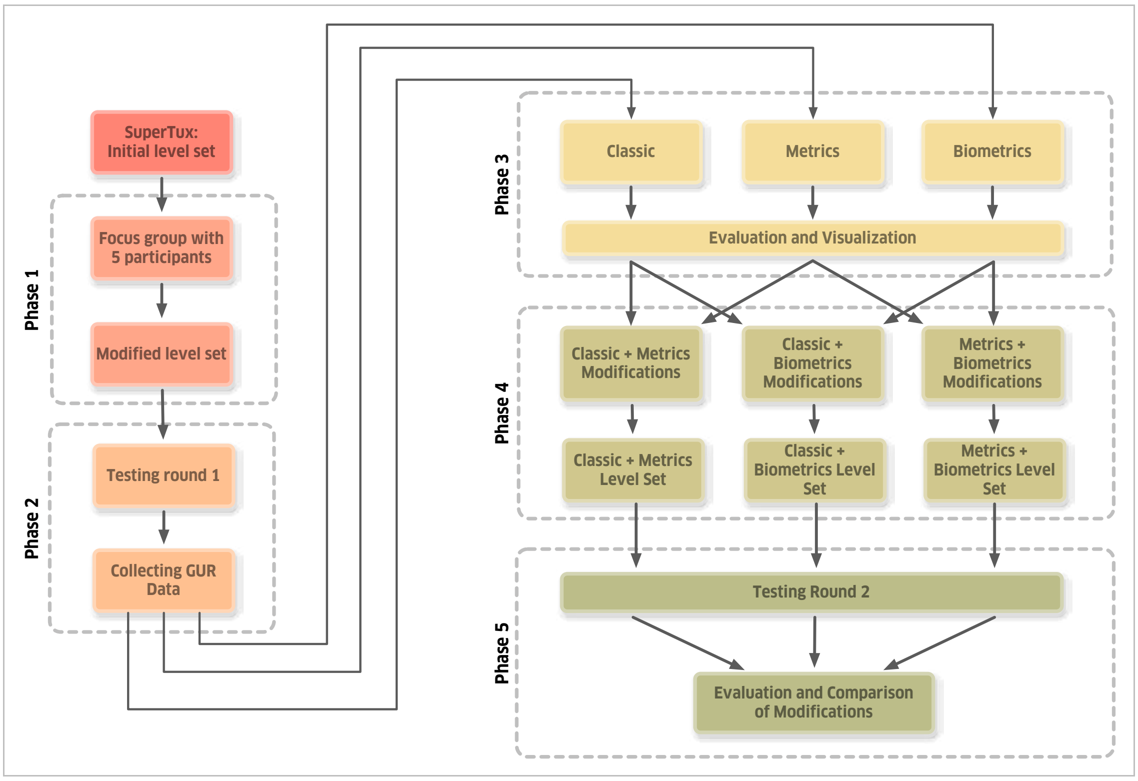

SuperTux: Initial level set
Phase 1
Focus group with 5 participants
Modified level set
Phase 2
Testing round 1
Collecting GUR Data
Phase 3
Classic
Metrics
Biometrics
Evaluation and Visualization
Phase 4
Classic + Metrics Modifications
Classic + Biometrics Modifications
Metrics + Biometrics Modifications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

Fig. 3: Flowchart of the main phases of this study

5.1 QUASI-EXPERIMENT

This study consisted of two experiments. The first one - this document refers to it as 'Testing Round 1' - was strictly used to collect data by means of different GUR methodologies for subsequent level design modification. In the second testing round, players were subjected to one of three level-sets (a sequence of three levels with an additional introduction level that precedes them) that were modified by the combination of two GUR methodologies.

While it would have been desirable to control all variables of the testing environment, the available research environment did not permit us to conform to the rigor required for a full experiment. For this reason the quasi-experimental research design was used for this research.

Singleton & Straits (2005) define quasi-experimental design as such:

> Legal, ethical, or practical considerations make it impossible to employ a true experimental design in some research situations. Frequently, random assignment of persons (or other units) is not possible. At other times, control or comparison groups cannot be incorporated into the design. [...] To deal with these problems, researchers have developed a number of **quasi-experimental designs,** so named because they take an experimental approach without having full experimental control. (Singleton & Straits, 2005, 206)

5.2 FOCUS GROUP

Since this study aims to simulate the design processes that would occur in a game production, we needed to prove that the initial level-set could be considered ready for QA from the perspective of game designers.

Focus groups are a suitable method for evaluating a subject matter in which the opinion of a professional field is of importance. While a series of interviews with game designers would have been a possible way to evaluate the quality of the initial level-set, a focus group is more likely to induce a discussion about what should be modified before attempting more granular and time-consuming tests on a wider player population. Having the game designers discuss (and defend) their points of view with each other allowed a distinction to be made between opinions that were idiosyncratic and what could be considered an expert consensus.

The open-ended nature of a focus group allowed attention to be brought to issues or problems that may not have been noticed by the researcher. A focus group consisting of several game designers has the additional benefit of reducing uncontested subjectivity.

5.3 Game User Research (GUR) Methodologies

This chapter focuses on the individual methodologies that were included in the study for the purpose of comparing their potential impact in the iterative process of designing game levels.

Three GUR methodologies were used in this study:

- Classic (Interview, Observation)
- Metrics
- Biometrics

5.3.1 Classic

This type of GUR methodology can be seen as the ‘traditional’ approach to QA in the game industry (see *2.1.1 Classic User Research Methods*) – in the sense that it is the most common and it requires the least amount of additional technology.

Classic GUR methodology is one of the three methodologies used for comparison between GUR methodologies. In the study we asked players after each level to answer a set of predefined questions including the possibility to freely add further comments on the level or the game in general. In addition to the questions, two researchers observed the players during their time in the game. Observations were then noted for each level and participant for later review.

In the context of this study, it was found to be hard for players to keep their comments confined to level design issues, as for the most part they do not recognize the difference between level design and general game design. It is then up to the interviewer to refocus the discussion on level design issues.

5.3.2 Metrics

In the context of QA in the game industry, metrics refers to the practice of tracking game states and player actions during play sessions - also referred to as ‘logging’. Implementing metrics for the purpose of QA usually involves additional technological efforts since logging data needs to be created, exported and interpreted. For more details see chapter *2.1.2 Metrics*.

In this study, player actions and game states were logged during game sessions and expressed in aggregated statistics and heatmaps (for an example, see *Figure 4*).

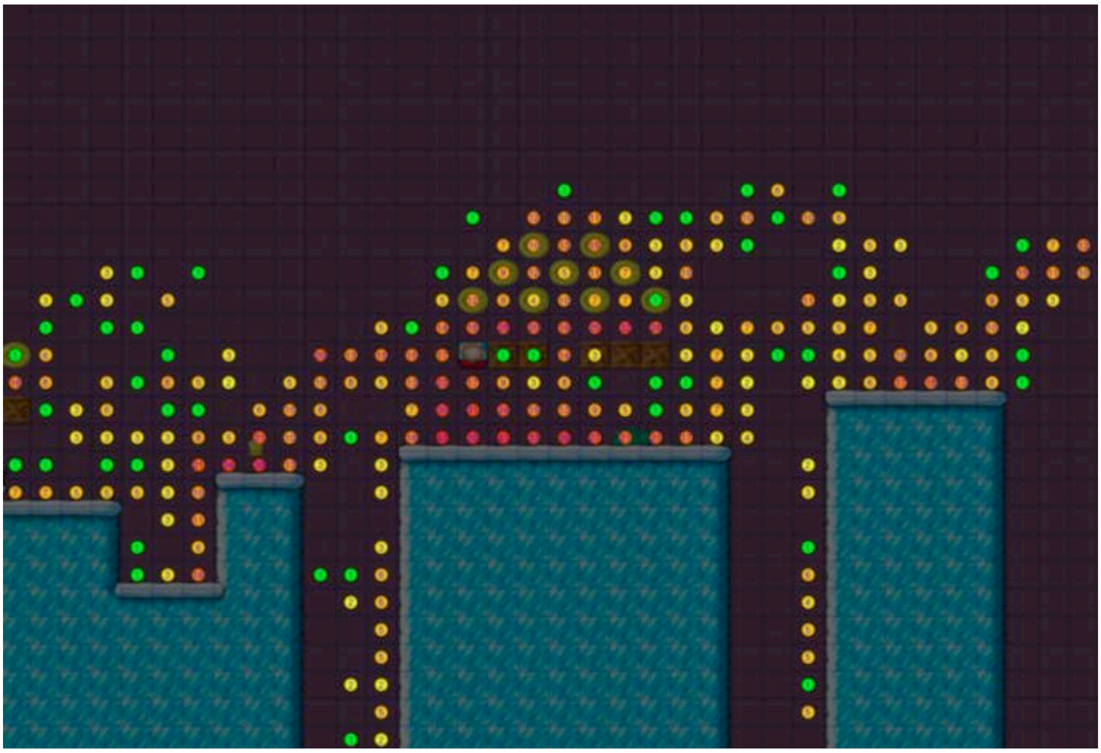


***Fig. 4:* Partial heatmap depicting player presence in level 2**

5.3.3 Biometrics

Within the field of GUR, biometrics is arguably one of the latest methodologies. Here biometrics refers to analysis of physiological signals (e.g. a player's heart rate) for the purpose of interpreting the inner state of a player while interacting with a game (see more in *2.1.3 Biometric Methods*).

Biometric testing was the third and last GUR methodology included in this study. During the game sessions of the first round the players were connected to sensors (heart rate, facial muscles and skin conductivity).

5.4 GAME EXPERIENCE QUESTIONNAIRE (GEQ)

With a comparative research design set up for this study, a measure was required to provide a basis on which comparisons can be drawn. If the stance is taken that players must appreciate their experience for a game to be considered ‘good’, it follows that the quality of a game level should be linked to the experience of the player during his or her time in the game.

Within the game industry a common approach is A/B testing - the practice of having one test subject test two versions of the same basic situation, usually with one particular difference that could change the user behavior. Since we would be testing three different level sets in the end, we could not expose the test subjects to all sets, and thereby use an approach similar to the A/B testing method. Testers would most likely be overwhelmed by the length of the test session, which would in turn influence the results. Another reason was that the changes between some of the levels were very subtle and would most likely not be noticed by the players. We therefore decided to use an approach that could give us a result that could be used to compare between players which have not played the same levels.

For this purpose we use the Game Experience Questionnaire (GEQ) as measurement for a comparison between modified level-sets. First presented at the ‘Fun and Games workshop’ (IJsselsteijn et al., 2007), the questionnaire has been used in several publications within the academic field (Gualeni et al., 2012; Nacke, 2009). While this makes the questionnaire a sufficiently solid basis for comparison, it has to be noted that there are few, if any, alternatives for the quantification of game user experience at the time of conducting this study.

In this study we used two of the GEQ modules: The ‘**Core’ module** and the ‘**Postgame’ module.** Both modules were presented to the players - also referred to as participants - after having completed a level-set. While the GEQ also offers an ‘In-Game’ module, we chose to not add further question modules into the study. We considered the amount of time that players spend in the game too short to justify a longer questionnaire at the end.

With the modules used in the study, the following aspects were scored:

Core Module:

- Competence
- Sensory and Imaginative Immersion
- Flow
- Tension / Annoyance
- Challenge
- Negative Affect
- Positive Affect

Postgame Module:

- Positive Experience
- Negative Experience
- Tiredness
- Returning to Reality

While the use of the GEQ in this study is focused on the comparison of level-sets, it was also used in the phase in which participants play through the initial, unmodified, level-set for the purpose of data collection. This allowed comparison of the GEQ ratings between individual players and the creation of player groups based on the GEQ rating they give. This information was also used when subsequent modification efforts were made to determine which changes were needed to motivate higher GEQ ratings, which presumably correlate to a better game user experience, among play groups that gave low ratings in the initial round.

6. PHASE 1: PREPARATIONS FOR SIMULATING A QA ITERATION

In this first phase of the research an initial level-set needed to be designed which was then used in the next phase to collect GUR data. The initial level-set consisted of four levels in total (see *Figure 5*). The first level in the set was a 'tutorial' level, which introduces players to the controls and mechanics of the game. The remaining three levels were considered to be the targets of QA efforts in this study.

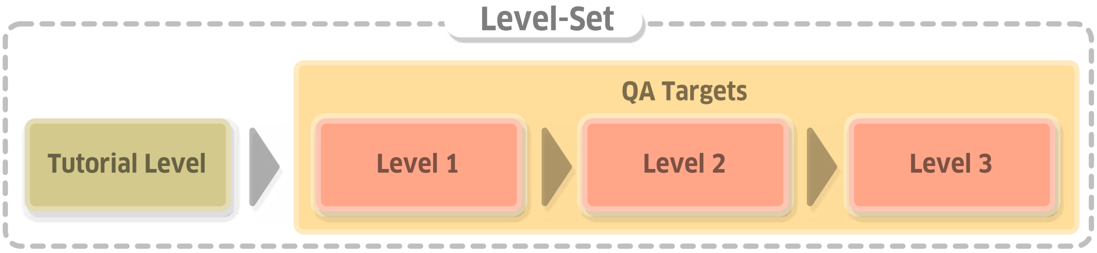


Fig. 5: Flowchart illustrating the structure of the level-set

The individual levels of a level-set were designed to be played in sequence as each level offers more challenges. The sequential connection between levels was further illustrated by the presence of a so-called 'overworld map' (see *Figure 6*) which directed the player automatically from one level to the next.

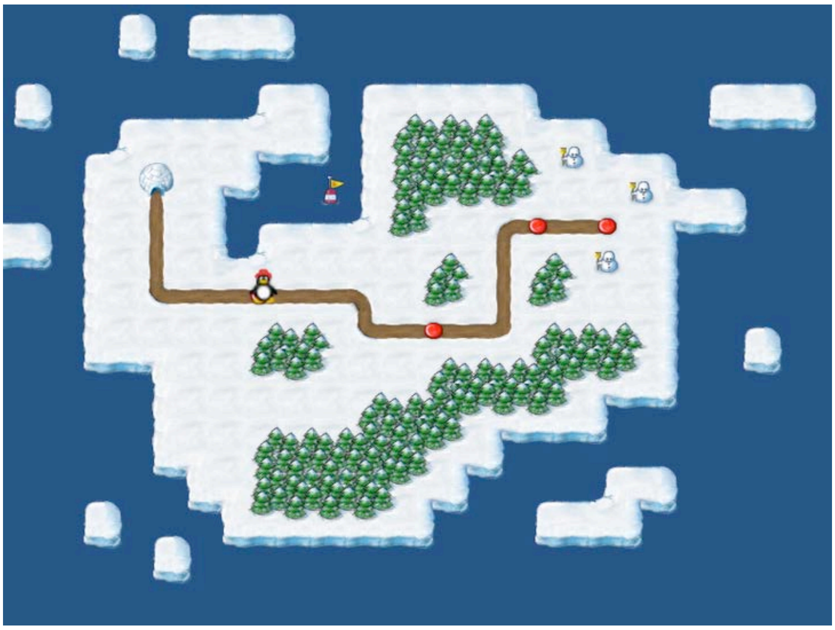


Fig. 6: Customized world map connecting the individual levels of a level-set

6.1 Level Design of the Initial Level-Set

SuperTux comes with a wide selection of levels, all of which are created by contributors in the open source community. While the open source nature of the game is advantageous in a lot of aspects of the research, it also means that levels tend to vary in quality, length and difficulty. For the purpose of the research however all the levels in a level-set had to be consistent in terms of quality and only increase the difficulty in a sequential manner.

As a starting point, the first three levels in SuperTux were used in sequence for the initial level-set. A fourth introduction level was added to the beginning of the level-set and was designed to contain explanatory signs for controls and game mechanics.

The three main levels were then modified in order to comply with heuristic design guidelines, as well as guidelines established by the development community of *SuperTux* (SuperTux Development Community, n.d.). Further modifications were taken to unify or structure some varying parameters, such as the spatial length of a level or the amount of available pickups. A full list of all modifications, including descriptions regarding the motivation behind changes, can be found in the appendix of this document (refer to *C. Focus Group Modification Table* in the appendix).

6.2 Initial QA - A Focus Group with Designers

To ensure a sufficiently high quality standard of the initial level-set in terms of level design, a team of game design students playtested 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. Given the implementation of the feedback, we argue that the initial level-set could be considered a viable representation of a professional level design in a 2D platform game at the point where we applied GUR methodologies to identify further possibilities for improvement.

6.2.1 Focus Group Design and Environment

The focus group was conducted in the faculty premises of the International Game Architecture and Design course at the NHTV Breda. In preparation to the meeting a rough guideline of questions was prepared to focus the discussion on the aspects of level design.

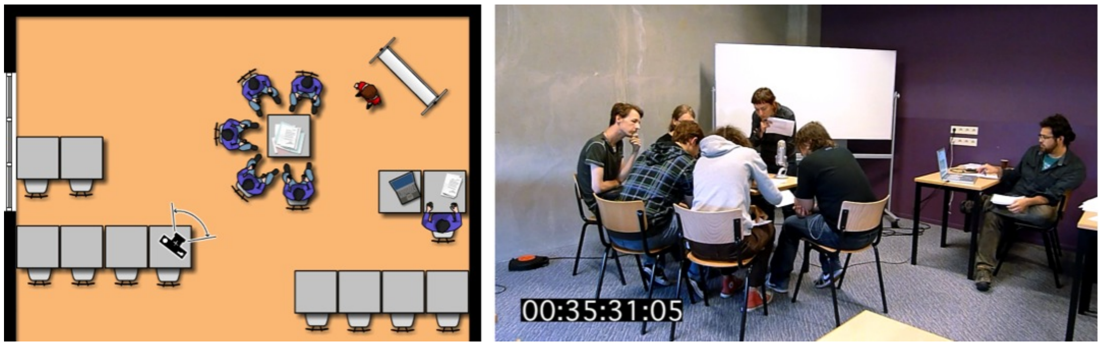


Fig.7: Left - Floor plan of the focus group setup. Right - Still frame of the focus group recording

Five focus group participants were assembled around a table with a microphone in the middle and illustration material of the levels on the table (see *Figure 7*). The moderator stood in front of the table and directed the discussion. A researcher was present on the side to protocol what was said.

6.2.2 Sampling: Focus Group Participant Selection

For the purpose of selecting the focus group participants we relied on the sampling method called purposive sampling (Singleton & Straits, 2005). The participants of the focus group were selected from among the students of the Design and Production course at the International Game Architecture and Design department at the NHTV Breda. We used academic prowess and professionalism as the determining factor in selecting participants. This was judged by the recommendations of two design lecturers and industry distinctions they received (such as game festival awards).

We decided to invite five students as compromise between allowing a discussion with a broad range of perspectives and keeping the group small enough to ensure that everyone was able to voice their opinion.

6.2.3 Focus Group Procedures

Before the focus group was conducted, we formulated a guide sheet to help guide the discussion (included in *B. Focus Group Guide Sheet* in the appendix). The guide sheet had some example questions to follow which focused on the quality and playability of the levels.

The focus group took place in the school building of the International Game Architecture and Design course of the NHTV in Breda. We prepared one for the larger classrooms to accommodate the focus group. Before we started we placed tables and chairs for the participants to place their laptops. One table was prepared with chairs for all participants to facilitate the group discussion. On this table we placed the microphone and prepared visuals of the level to aid the discussion. The camera was set up to record this table.

When the participants arrived we showed them to their tables where they could set up their laptops. We then provided them the game via USB stick.

Before the participants started the game we gave them a short introduction on the research and what was going to happen, as well as a short explanation about the game. The participants then played the initial level set twice.

Afterwards the participants were asked to the discussion table and we started the camera recording. The moderator introduced the first question of the guide sheet and the discussion began.

In total the discussion lasted 1 hour. Once the time was up the focus group concluded and the participants were thanked for their cooperation. Afterwards the audio was transcribed and audio and video files were archived.

6.2.4 Focus Group Results

With the transcription of the focus group, a list of suggested modifications was formulized and subsequently executed in the level design of the initial level-set. A table that documents the changes that were encouraged by the focus group can be found in the appendix (*C. Focus Group Modification Table*). The table lists the modifications next to the time-code of the focus group recording.

The following matrix table lists the level parameters as they are after the implementation of focus group feedback – changes are bolded, with previous values in parenthesis:

	Level 1	Level 2	Level 3
Star Blocks	1	1	1
Tux Doll Blocks	**0** (1)	**0** (1)	**0** (1)
Tux Upgrade Blocks	**4** (3)	**4** (3)	**4** (3)
Secret Areas	1	1	2
Level Gaps	4	8	**10** (13)

Total Enemies	**18** (15)	**23** (22)	**26** (31)
Enemy Type I	**15** (12)	15	**15** (18)
Enemy Type II	2	**6** (5)	**8** (10)
Enemy Type III	1	2	3
Enemy Groups (2 in a row)	**3** (2)	**4** (3)	**5** (3)
Enemy Groups (3 in a row)	0	1	**3** (4)

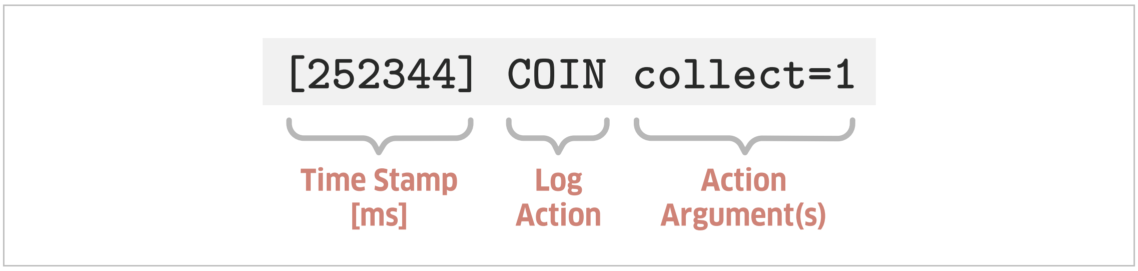
Note: As reaction to the perceived danger of 'Icicle' obstacles in the focus group, a pair of two is counted as one type I enemy. Furthermore, icicles are not considered when counting enemy groups.

6.3 ADDING LOGGING FUNCTIONALITY TO THE GAME

With the implementation of feedback from the focus group, the level-sets are, from a level design standpoint, ready to be play-tested by randomized participants. Before such playtests could begin, *SuperTux* needed to be modified so that player events in the game are protocolled in a log file.

SuperTux is developed by an open-source community, which allowed us to access the code-base for the purpose of adding logging functionality. Unfortunately, the open-source nature of the game also means that the game code can be inconsistent in terms of quality or clarity. Ideally, the functionality of logging game states and events should be implemented during the development of a game. For the purpose of this study however, creating a game from scratch was not a feasible choice.

After some considerations, we devised the following logging functionality:

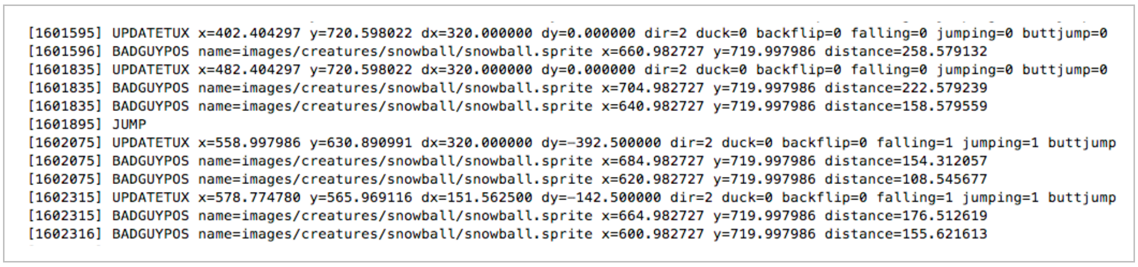


```
[252344] COIN collect=1
```

Fig.8: Example of a logging line, depicting the collection of a coin by the player at a specific time.

Each logging line is time stamped, which is simply the declaration of time that has elapsed since the initiation of the logging activity. For an example of the log lines, see *Figure 9*. In our study the timer is started once a player enters the tutorial level and runs until the end of the last level.

The second argument in the log line declares the logged action in capital letters. For most actions a set of arguments follow the log action. These arguments describe the action in detail and often contain 'state switches', which keep track of whether an action has been performed (marked by a '1') or not (marked by a '0').



```
[1601595] UPDATETUX x=402.404297 y=720.598022 dx=320.000000 dy=0.000000 dir=2 duck=0 backflip=0 falling=0 jumping=0 buttjump=0
[1601596] BADGUYPOS name=images/creatures/snowball/snowball.sprite x=660.982727 y=719.997986 distance=258.579132
[1601835] UPDATETUX x=482.404297 y=720.598022 dx=320.000000 dy=0.000000 dir=2 duck=0 backflip=0 falling=0 jumping=0 buttjump=0
[1601835] BADGUYPOS name=images/creatures/snowball/snowball.sprite x=704.982727 y=719.997986 distance=222.579239
[1601835] BADGUYPOS name=images/creatures/snowball/snowball.sprite x=640.982727 y=719.997986 distance=158.579559
[1601895] JUMP
[1602075] UPDATETUX x=558.997986 y=630.890991 dx=320.000000 dy=-392.500000 dir=2 duck=0 backflip=0 falling=1 jumping=1 buttjump
[1602075] BADGUYPOS name=images/creatures/snowball/snowball.sprite x=684.982727 y=719.997986 distance=154.312057
[1602075] BADGUYPOS name=images/creatures/snowball/snowball.sprite x=620.982727 y=719.997986 distance=108.545677
[1602315] UPDATETUX x=578.774780 y=565.969116 dx=151.562500 dy=-142.500000 dir=2 duck=0 backflip=0 falling=1 jumping=1 buttjump
[1602315] BADGUYPOS name=images/creatures/snowball/snowball.sprite x=664.982727 y=719.997986 distance=176.512619
[1602316] BADGUYPOS name=images/creatures/snowball/snowball.sprite x=600.982727 y=719.997986 distance=155.621613
```

***Fig.9:* Example of several log lines from an actual participant log file**

A detailed table, describing the actions that were logged with the corresponding action arguments, can be found in the appendix (*D. Log Action and Arguments Table*).

6.4 FORMULATING A QUESTIONNAIRE FOR PLAYER INTERVIEWS

This section describes how the questionnaire for general feedback and classic GUR methodology was created. The questionnaire can be found in the appendix (*I. Player Interview Form Used in Testing Round 1*).

At the beginning of the questionnaire, we included the questions regarding general player feedback. These questions were based on a 5 point Likert scale and asked to player after how fun, long/short and difficult the level was. We decided on these terms because of how easy they are to understand. These first three questions were used in all three GUR methodology combinations as a basic reference point (refer to chapter *8. Phase 3: Data Evaluation and Visualization* for more information).

The parts of the questionnaire used for collecting classic GUR data featured open questions. This allowed participants to point out specifics that they felt were worth mentioning. The questions concerned confusion, frustration, enjoyment and surprise. While answering these questions the participants were allowed to look at the visual aid (refer to *H. Level Sheets for Player Interviews in Testing Round 1* in the appendix) to help them recall the level. The last question gives the participants the opportunity to mention anything that did not fit into the previous categories. Finally the researcher had the option to ask the participant about anything that was noticed during observation when applicable.

The questionnaire was designed with the intention of creating a balance between brevity and gathering a reasonable amount of data. Since the interview was repeated three times during the test it couldn't last too long or become tedious to the participant, yet we had to be sure that we collected enough data.

6.5 Conclusion and Discussion of Phase 1

The focus group organized in this phase of the study yielded many actionable results that were used to improve the game. The fact that all participants in the group had experience with the challenges of level design meant that very little time was spent on discussing aspects of the game that were secondary to our research, such as visual fidelity or general game mechanics.

We believe that the experience of the participants was a key factor in the explicitness of the results and would have not been achieved if the focus group would have been conducted with less specialized participants. It can be difficult for people to distinguish what constitutes as level design and what not – a problem that is to a big extent circumvented by involving participants that work as game and level designers.

On the other hand it should also be noted that a focus group could lead to results that are biased towards practices that are favored by more the outspoken professionals in the group. Another potential influence could be introduced if participants are familiar with the professional work of each other and neglect their own opinion out of respect or insecurity in front of peers with more experience.

The second major aspect in this phase was the implementation of logging functionality, which turned out to be a time intensive endeavor. Arguably, many game developers track metrics in their games routinely already. In small development teams the integration of logging functionality can, however, be hard to justify in terms of required programming time. While we had to add logging functionality to a finished game, it is conceivably more time-efficient to include support for logging alongside the development of a game.

At this point it should be noted that the usage of biometric GUR methodologies also requires logging functionality in the game code.

7. Phase 2: Testing Round 1 - Collecting GUR Data

This chapter describes the first round of testing including the set up of the testing environment, the necessary equipment, details on the participants and the research procedure.

7.1 Testing Design and Environment for Round 1

We set up the environment for the first round of testing in the biometrics lab at NHTV University of Applied Sciences in Breda. Testing took place on workdays between 9 AM and 6 PM. Because of the measuring of biometric data, audio and video recordings, and game data logging, the required equipment for the first testing round was as followed:

- Three desktop computers – one for the participant to play the game on, one for monitoring biometric data and one for data collection
- Three monitors – of which two were connected to the game computer, one for the participant and one for video recording, and the third to the biometrics computer.
- Two keyboards
- Biometric sensors and sensor preparation materials
  (described in chapter *7.1.1 Biometric Sensors*)
- Two webcams
- One microphone
- Information sheet, consent form and debrief sheet
- Visual reference cards of the three game levels
  (refer to *H. Level Sheets for Player Interviews in Testing Round 1* in the appendix)
- Interview sheets
  (refer to *I. Player Interview Form Used in Testing Round 1* in the appendix)

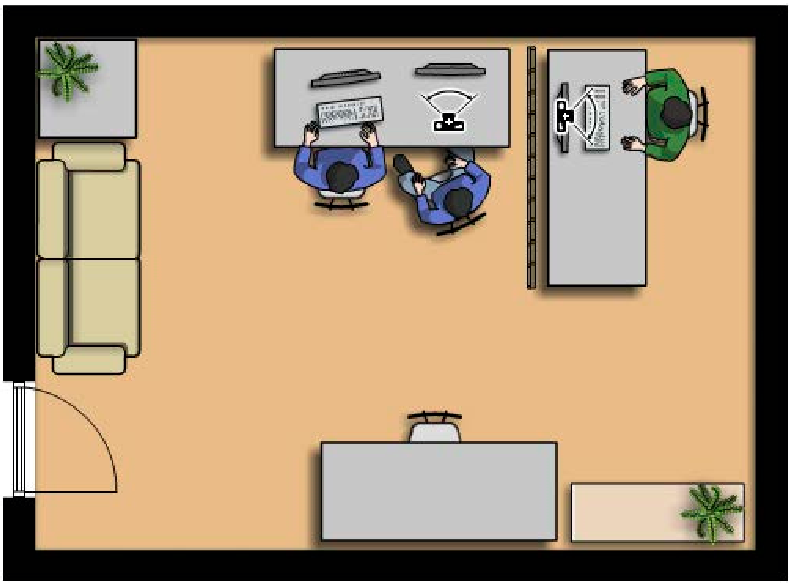


Fig. 10: Floor plan of test session setup

As can be seen in *Figure 10*, the room was set up into two areas using a room divider. The participant was seated behind the divider. Here we placed one of the monitors, which showed the game as well as a keyboard for the participant to play the game with. Additionally there was a webcam, which recorded the participant's face as well as a microphone for audio recordings.

To not give the participant the feeling he or she was being watched, the researchers were on the other side of the room divider during game play. The second monitor attached to the game computer was placed here, showing the game screen as well as the output of the webcam, allowing the researchers to monitor the game and the participant. This monitor was once again recorded using the second webcam, giving us a single video file per participant in which we could easily see facial responses and to which game actions they correlated (see *Figure 11*).

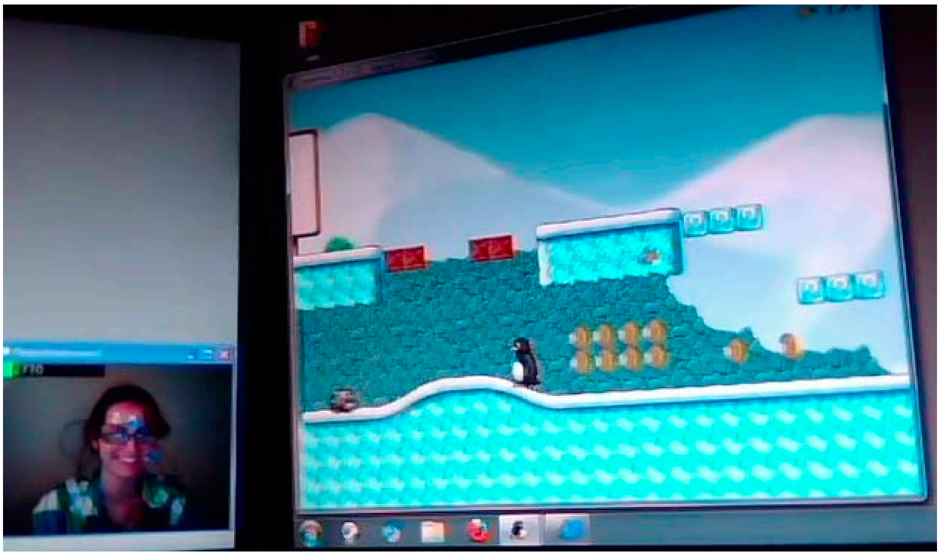


Fig. 11: Still from second webcam footage, showing both the game and the participant in a single video file

The second computer and its monitor were also placed on this side of the room divider, allowing the researcher to monitor the biometric readings. The sensor readings and data logs were sent to a third computer, which was placed in a neighboring room.

7.1.1 Biometric Sensors

For testing we used several biometric sensors to monitor the participants during play. For this test we used sensors that recorded facial movement, heart rate and skin conductivity.

Fig. 12: EMG Sensor

Facial Electromyography (EMG) - Detects activity in the 'Corrugator Supercilii' muscle group, associated with frowning, and the 'Zygomaticus' muscle group, associated with smiling. It is used to measure emotional valence (positive/negative emotions)

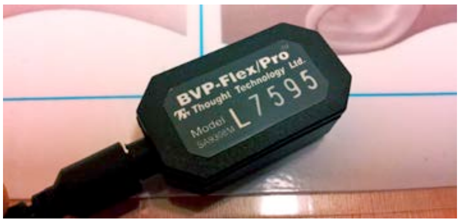


Fig. 13: BVP Sensor

Blood Volume Pressure (BVP) - BVP measures blood volume pressure, which we used to infer the heart rate. This informed us about the player's stress, fear and excitement.

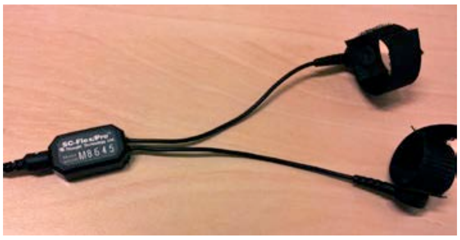


Fig. 14: GSR Sensor

Galvanic Skin Response (GSR) - GSR measures skin conductivity and is analyzed to in relation with other physiological data to measure excitement, fear, engagement and arousal.

7.2 Sampling: Participant Selection for Testing Round 1

For this round of testing we used convenience sampling to select the participants (Singleton & Straits, 2005).

We decided to not include students from the International Game Architecture and Design program of NHTV Breda University of Applied Sciences. While we wanted knowledge of game and level design in the focus group, in this round we did not want the participants to know too much about the inner workings of a game. A deeper understanding of game design would make the participants focus too much on the level design, instead of just experiencing the game.

Since the study focuses on casual gaming, there were not too many restrictions besides excluding 'hardcore gamers'. The final group of participants had ages ranging from 18 to 57 years, with a median age of 25. Of the 23 participants, of whom 20 provided all the required data and were therefore valid for the research, 12 were male.

7.3 Testing Procedures for Round 1

The experiment was split up into three phases: the briefing, the test and the debriefing phase. A minimum of two researchers was present at any given test session. Both researchers welcomed the participant to the test environment and introduced themselves. Then one of them described the experiment to the participant by use of the information sheet. Once the researcher was sure the participant understood exactly what the experiment entailed and the participant still agreed to participate the researcher had him or her fill out and sign the consent form.

The participant was asked to play three introductory levels, which required the understanding of the core mechanics of the game in order to finish. These levels included all game elements (enemies, items, obstacles) that the participant would find in the test levels. While the participant played these levels the researchers prepared the computers and equipment for the test. After finishing the introductory levels the participant was asked to take a seat behind the room divider. The participant's chair was chosen to limit movement, i.e. not a swivel chair. To prepare for the sensors the participant was asked to use skin preparation gel on the areas of the face which would have sensors attached to them. The researcher applied the facial and finger sensors. Finally the wires of all sensors were taped to the participant’s clothing to limit their interference during the test session. While one of the researchers prepped the participant, the other researcher prepared the game and launched the program *Biograph Infinity*. This is a computer program designed to visualize the sensor data recorded by the sensors. Participants were asked to frown and smile to test if all sensors were placed correctly. If not, the sensors were repositioned until the data showed up correctly in the program.

With all the sensors properly attached the researchers took their places behind the room divider. If everything was in working order the participant was asked to start the game and play the first two levels, the tutorial level and the first official test level. The participant was asked to stop playing after the first test level, as there would be a short interview. During the test session the participant was not supposed to talk and the researchers did not answer any questions. Once the participant finished the first test level, one of the researchers went behind the divider and held a short interview with him or her. During the interview the researcher used a printed overview of the level the participant just played to help him or her in pointing out certain areas they did or did not enjoy. The participant was asked to rate the level on a scale from one to ten in difficulty, length and fun. He or she was also asked after what they enjoyed, surprised them, annoyed them, etc. With these questions the participant could look at the printed reference card of the level and give extra comments on specifics.

After the first interview the researcher went back behind the divider, leaving the participant behind the screen. The participant was then asked to continue with the next level. The same steps and interview were repeated until the participant finished all levels. Once the third interview was conducted the participant was allowed to remove the sensors with assistance of the researcher.

After the sensors were removed the researcher opened the GEQ, which the participant was then asked to answer. To conclude the test the researchers went through the debriefing sheet with the participant, which he or she got a copy from to take home. After the participant left, the researchers cleaned the sensors and prepared the room for the next test session. The information sheet, consent form, questionnaires, visual aids and debrief sheet are included in the appendix.

7.4 CONCLUSION AND DISCUSSION OF PHASE 2

This second phase of the study gave us first hand experience with the different GUR methodologies, which allowed us to derive some initial conclusions regarding the usability of the individual GUR methodologies.

In terms of biometric data collection, we witnessed that it can be very hard for participants to avoid making occasional comments or vocal sounds while playing the game. This brought the researchers into the interesting conundrum that reminding players to not talk would in itself present an interruption in their game experience. For the analysis of the facial sensors however, this meant that such instances would have to be taken out as noise during the data evaluation phase.

This can make the use of biometric GUR methodologies in a development environment more time consuming. Given the fact that testing players with biometric sensors takes already more time than without, it can be argued that implementing this methodology can seem an expensive addition into established QA processes of a game developer. It is therefore crucial for the viability of biometric GUR methodologies to develop a framework that reduces the amount of time it takes to measure and evaluate data.

In regards to classic GUR methodologies, we believe that the approach of using level sheets during interviews with players proved successful for gathering data that is relevant for level design. With their use, participants were able to not only recall events in a level but also tie them to specific locations – an important aspect for potential improvements in the level design. In retrospect, the case could be made that dividing a level into more sections would have provided a finer granularity of interview data.

With regards to the observation aspect of classic GUR, we encountered that with a high amount of player observations, the observers (one of which is the author of this document and the level designer in this study) were left with stronger impressions of particularly lengthy test sessions. Since the designer of a level might have specific intentions regarding how long a level is ‘supposed’ to take, players that differ from that intention can leave a bigger impression. An observing designer might then be tempted to accommodate players based on their performance instead of their preference. We believe that this potential bias should be kept in consideration when designers act as observers during QA sessions.

8. PHASE 3: DATA EVALUATION AND VISUALIZATION

In the third phase of the study, we analyzed and processed the GUR data that we acquired from test sessions in phase 2. Since our research question focuses on combinations of GUR methodologies, the data from each combination was treated separately. By evaluating the gathered data from each combination without using data from other methodologies we wanted to reduce potential bias. Such a bias could be introduced due to belief or disbelief in the value of one or more of the GUR methodologies in this study or simply due to the fact that the researchers of this study were able to observe players as part of the testing procedure.

While this approach of actively ignoring information from different GUR methodologies can seem counter-intuitive to designers, it is a necessary step to ensure that subsequent modifications in the level-set are not influenced by personal beliefs or preferences.

During the interpretation of the data sets as well as during the modification phase the designer could refer to the general feedback data (fun, length and difficulty) at all times. The reason for this is that we deemed it necessary to have some basic feedback in order to properly contextualize the more objective GUR methodologies. With the basic data it was possible to make judgment on whether a measure was desirable or not.

8.1 A CASE FOR VISUALIZATION OF DATA FOR DESIGNERS

Before describing the individual data sets, this section seeks to highlight the importance of visualizing data from GUR methodologies for the use by designers. We argue that level design is a highly visual process of which the goal is to establish an evenness of challenge and rhythm. In this process the designer needs to keep track of several variables that are given in the range of architectural possibilities set by the game, like enemy range, game items, etc. GUR methodologies tend to express results in data, i.e. numbers, tables and percentages. Given the nature of level design, this type of data can be confusing and hard to put into context. Proper visualization of the data is therefore crucial. Ideally the data is presented in a way that it can be connected to the level in terms of location, making it clear for the designer what the data points to.

8.2 Classic GUR Data

As introduced in *2.1.1 Classic User Research Methods*, there are several classic GUR methodologies that can be used in the QA phase of game development. In this research, we focused on open interview questions that have been asked after each of the three levels in the level-set (excluding the tutorial) as well as observation notes that have been taken by observing the game screen during the testing session. Both the feedback of the players and the observer notes have been hand-written on the questionnaire sheets that are used for each level.

After the testing sessions, all questionnaire sheets have been collected and transcribed into a spreadsheet.

8.2.1 Interpretation of Classic GUR Data

Once all the data was digitalized into a spreadsheet, the feedback was filtered for information that was considered to be irrelevant for the specific goals of this research.

Examples of filtered feedback include:

- Players comparing the game to Nintendo's *Super Mario* (1985) game series
- Comments regarding the visual quality of the game
- Requests for additional game mechanics or items
- Recounts of the players' performance or intentions if the researchers were not able to identify information that could be used for the purpose of level design

The filtered data was then divided into different general themes in correlation to the topics of the interview questions. The themes were 'Confusing Instances', 'Frustrating Instances', 'Enjoyable Instances' and 'Surprising Instances'.

Apart of these themes, the questionnaire also gave players the possibility to add additional remarks. Likewise, the researchers were able to add further information based on their observations during the test session. Any information that came from such additional sources was categorized under one or more of the aforementioned themes.

Early in the evaluation process it became clear that many players mentioned similar instances. The next step was therefore to identify the underlying situation that players described in the interviews and count them across all players as well as per section of a level. We marked situations as 'keep' if players mentioned elements in the game that they enjoyed and marked them as 'change' if players had complaints about elements.

In some instances, players felt ambivalent about elements they encountered in a level. In such cases, we marked both a counter for 'keep' and a counter for 'change' to highlight the fact that players felt strongly enough to mention the instance in the first place.

After having prepared the data in the way described we were able to create counts for 'themed' instances that were noteworthy to players as well as a list that ranked those instances between desired (signified by a high 'keep' count) and undesired (signified by a high 'change' count).

8.2.2 Visualization of Classic GUR Data

To visualize the gathered data, we decided to convert counts of 'themed' instances (refer to the previous sub-chapter) into graphs that plot the values over the four sections of a level.

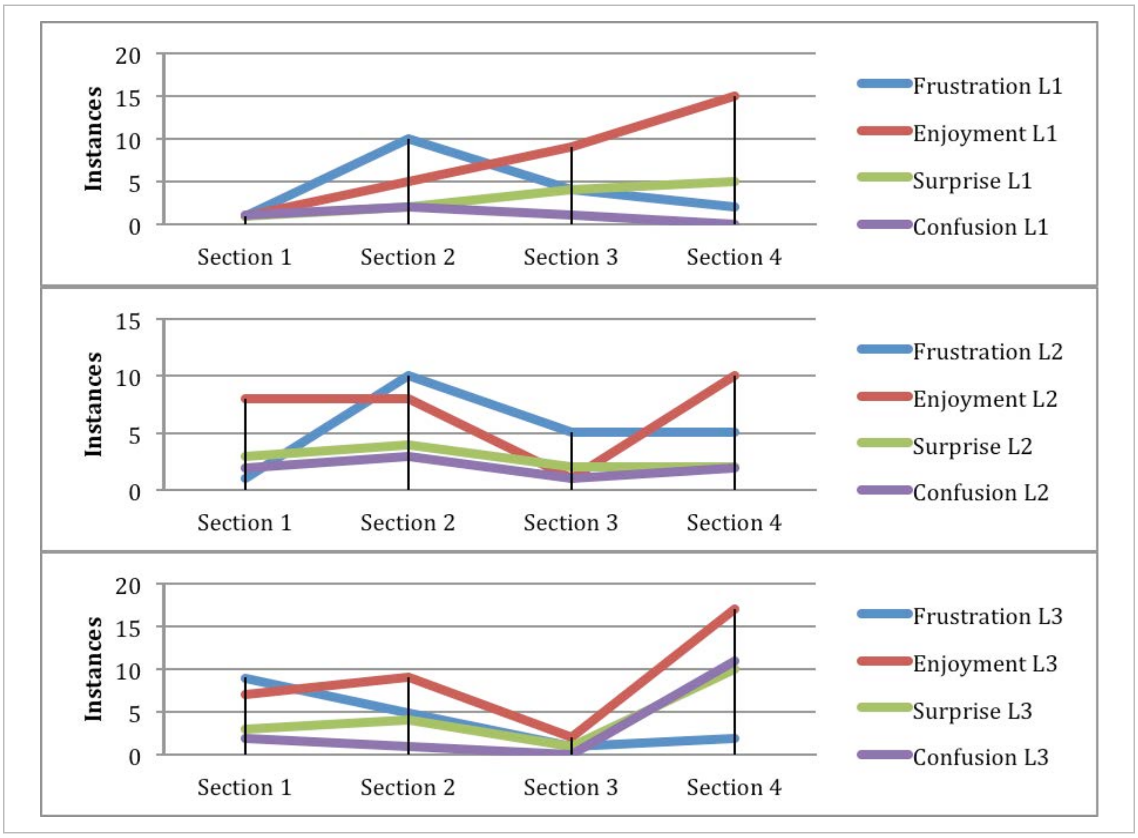
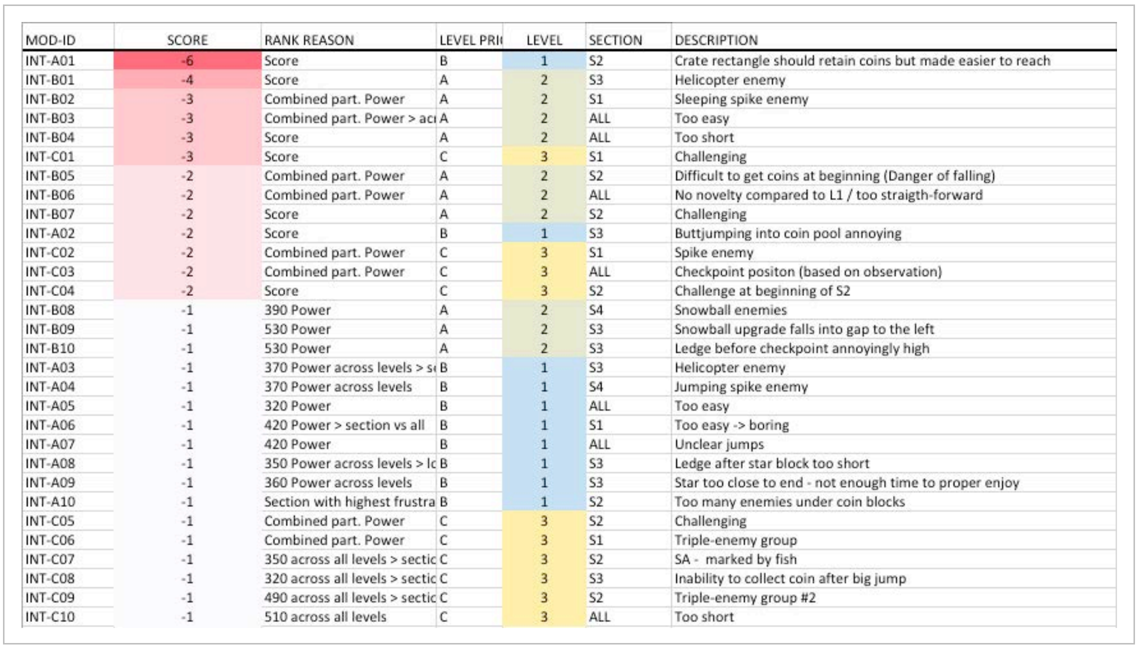


Fig. 15: Graphs charting instances of frustration, enjoyment, surprise and confusion for each level

8.2.3 Results of Classic GUR Data

The result of the classic GUR data collection is a set of graphs (as shown in the previous sub-chapter) as well as a ranked list of specific elements or instances that should be improved.



MOD-ID	SCORE	RANK REASON	LEVEL PRI(	LEVEL	SECTION	DESCRIPTION
INT-A01	-6	Score	B	1	S2	Crate rectangle should retain coins but made easier to reach
INT-B01	-4	Score	A	2	S3	Helicopter enemy
INT-B02	-3	Combined part. Power	A	2	S1	Sleeping spike enemy
INT-B03	-3	Combined part. Power > acı	A	2	ALL	Too easy
INT-B04	-3	Score	A	2	ALL	Too short
INT-C01	-3	Score	C	3	S1	Challenging
INT-B05	-2	Combined part. Power	A	2	S2	Difficult to get coins at beginning (Danger of falling)
INT-B06	-2	Combined part. Power	A	2	ALL	No novelty compared to L1 / too straigth-forward
INT-B07	-2	Score	A	2	S2	Challenging
INT-A02	-2	Score	B	1	S3	Buttjumping into coin pool annoying
INT-C02	-2	Combined part. Power	C	3	S1	Spike enemy
INT-C03	-2	Combined part. Power	C	3	ALL	Checkpoint positon (based on observation)
INT-C04	-2	Score	C	3	S2	Challenge at beginning of S2
INT-B08	-1	390 Power	A	2	S4	Snowball enemies
INT-B09	-1	530 Power	A	2	S3	Snowball upgrade falls into gap to the left
INT-B10	-1	530 Power	A	2	S3	Ledge before checkpoint annoyingly high
INT-A03	-1	370 Power across levels > s(	B	1	S3	Helicopter enemy
INT-A04	-1	370 Power across levels	B	1	S4	Jumping spike enemy
INT-A05	-1	320 Power	B	1	ALL	Too easy
INT-A06	-1	420 Power > section vs all	B	1	S1	Too easy -> boring
INT-A07	-1	420 Power	B	1	ALL	Unclear jumps
INT-A08	-1	350 Power across levels > lc	B	1	S3	Ledge after star block too short
INT-A09	-1	360 Power across levels	B	1	S3	Star too close to end - not enough time to proper enjoy
INT-A10	-1	Section with highest frustra	B	1	S2	Too many enemies under coin blocks
INT-C05	-1	Combined part. Power	C	3	S2	Challenging
INT-C06	-1	Combined part. Power	C	3	S1	Triple-enemy group
INT-C07	-1	350 across all levels > sectic	C	3	S2	SA - marked by fish
INT-C08	-1	320 across all levels > sectic	C	3	S3	Inability to collect coin after big jump
INT-C09	-1	490 across all levels > sectic	C	3	S2	Triple-enemy group #2
INT-C10	-1	510 across all levels	C	3	ALL	Too short

Fig. 16: Table with ranked suggestions for improvement

The list is ranked by a score, which increases for each 'change' request and decreases for each 'keep' request. Since several elements share the same score, we decided to include a list of additional factors that determine the ranking within items that share the same score:

- Rank by score
- Participants with low ratings of fun in a given level
- Change was not contested by a keep request
- Issues in sections had priority over issues with unclear locations

8.3 Metric GUR Data

For the evaluation of metric GUR data we developed several small scripts (written in the programming language *Python*). In general terms, we programmed these scripts to iterate over all the logging lines and perform certain counts or calculations to create 'derived measures'. Such derived measures were aggregated across all participants to build a mean for levels and level sections.

8.3.1 Interpretation of Metric GUR Data

In an effort to interpret the data gathered from metric GUR methodologies, a list of derived measures was created. These derived measures were calculations on the data with the purpose of contextualizing the information. As an example, many measures were likely to increase the more time a player spent in a level. By taking the level completion time into account, it was possible to compare player metrics regardless of how long it took them to complete a level. A full list with all derived measures can be found in the appendix (see *J. List of Derived Measures from Metric GUR Data* in the appendix)

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 many actionable correlations.

8.3.2 Visualization of Metric GUR Data

The visualization of metric GUR data was a challenging step in the evaluation phase. With a high amount data at our disposal, the challenge was to set a focus on what would be important to visualize and which metrics would benefit from this treatment.

As introduced earlier, within GUR, metrics data is often presented in the form of heatmaps. The correlation of color hues to specific locations in a game level make this form of visualization ideal for displaying data that is relevant for the level design of a game.

For our study, we decided to express the following measures as heatmaps:

- **Direction change** instances
- **Continuous movement** instances
- **Enemy kill** instances

- **Tux death** instances
- **Frustration death** instances
- **Jump** instances
- **Presence** of the player for each 'Update Tux' log action

The heatmaps that are produced with these measures are aggregated across all users using a logarithmic function both on the individual level as well as on the aggregation of all users. An example:

> *Example:* *Player A passes a given tile 1 times – a hue value for 1 is used on this tile to show a presence heatmap. If the player passes a tile 10 times, a hue value of 2 is used on the presence heatmap. Hundred passes then correlate to a hue value of 10 and so on. Likewise, the same logic is used to aggregate presences across players.*

The hue values on the heatmaps are therefore weighted in a way that repeated occurrences of a measure on a single tile influence the representing color hue less and less. In simple terms, we made the decision that lower values should be given a higher range of possible color hues than higher values.

Below is a list of heatmaps that were created in this study. Full format representations of these can be found in the appendix of this document.

- Presence map
- Change direction map
- Continuous movement map
- Tux death map
- Frustration death map
- Jump map
- Enemy kills map
- Instances of facing left map

Most of the derived measures we acquired with the help of self-written scripts are visualized in the form of graphs. Whenever possible, we used metrics data from level sections to put the data in context with locations in the level. Some derived measures were however only available for the level as a whole.

8.3.3 Results of Metric GUR Data

The result of the metric GUR data collection is a set of heatmaps and graphs that are featured in the previous sub-chapter. These are also the main results of metric data evaluation.

In addition to data visualizations, a table has been created that aggregates derived measures from all participants by showing the mean of each derived measure. A full list of tables has been included in *K. Metric Data Results* in the appendix.

During the creation of derived measures we encountered several situations in which the granularity of tracking states in the game could have been finer to provide a better resolution for tracking player input. We chose to log the position and state of the player avatar four times per second. The player is however able to trigger some actions more often than that. These actions are unfortunately lost in our logs. In retrospect the logging frequency of the 'Update Tux' log action should have set to a higher value.

8.4 Biometric GUR Data

For the evaluation of biometric GUR data, each level was divided into 12 equal sections, which were labeled alphabetically. This number was chosen in an effort to keep sections large enough to contain enough biometric data, while providing enough localized information.

Originally biometric measures provided us with 2000 data samples per second. Since it would have been impossible to filter the valid data from this manually, we programmed scripts to analyze the data and remove the noise.

In the process of removing noise from the biometric data, we encountered a technical problem that forced us to remove the recorded measures of skin conductivity. While this was an unfortunate setback in the study, we decided to move forward with the remaining biometric data. The evaluation of biometric GUR data is therefore based on the responses of facial muscles and the heart rate of participants.

8.4.1 Interpretation of Biometric GUR Data

Biometrics data gave use measures for BVP (blood volume pressure), ZYM (activity of the zygomaticus major muscle) and CS (activity of the corrugator supercilii muscle). BVP was interpreted in a different way from ZYM and CS.

BVP was measured at the beginning of a section and at the end. The mean of these two measures was taken as the measure for a section. The measures of all participants were then taken together to build a mean for each section across all participants.

Both ZYM and CS were interpreted in the number of peaks per section. To qualify as a peak the muscle activity had to provide a value over a certain threshold. As with BVP, the mean of all peaks over all participants was taken to provide the value per section.

8.4.2 Visualization of Biometric GUR Data

For the visualization of biometric data, the individual biometric measures were expressed in graphs and presented next to the corresponding level sections. This was done both for the equally sized 12 level sections as well as for specific regions of interest.

For a full sized version of the graphs, refer to *M. Biometric Result Graph*.

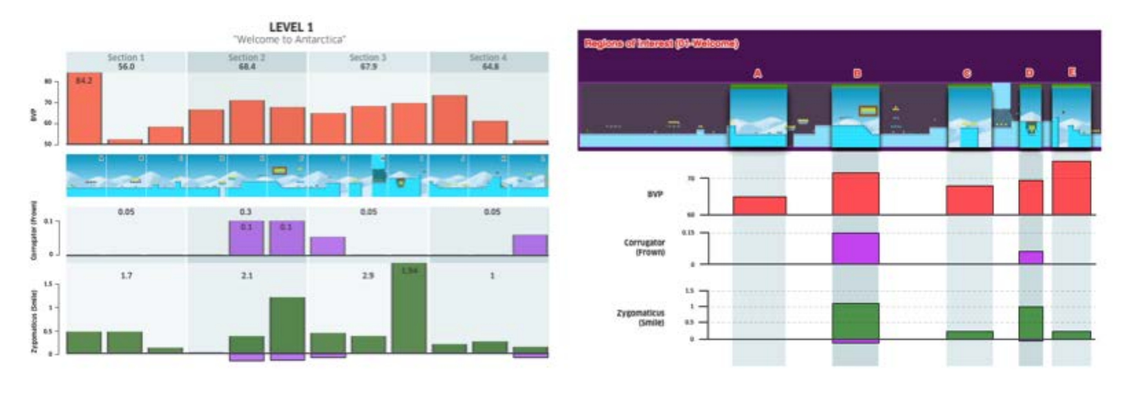


Fig. 17: *Left - Biometric data superimposed over map sections of level 1. Right - Biometric data for regions of interest (ROI) in level 1*

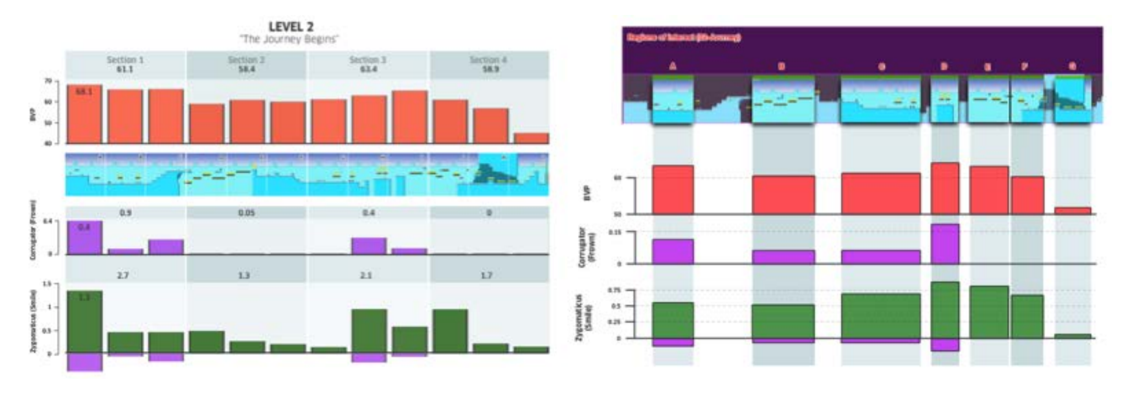


Fig. 18: Left - Biometric data superimposed over map sections of level 2. Right - Biometric data for regions of interest (ROI) in level 2

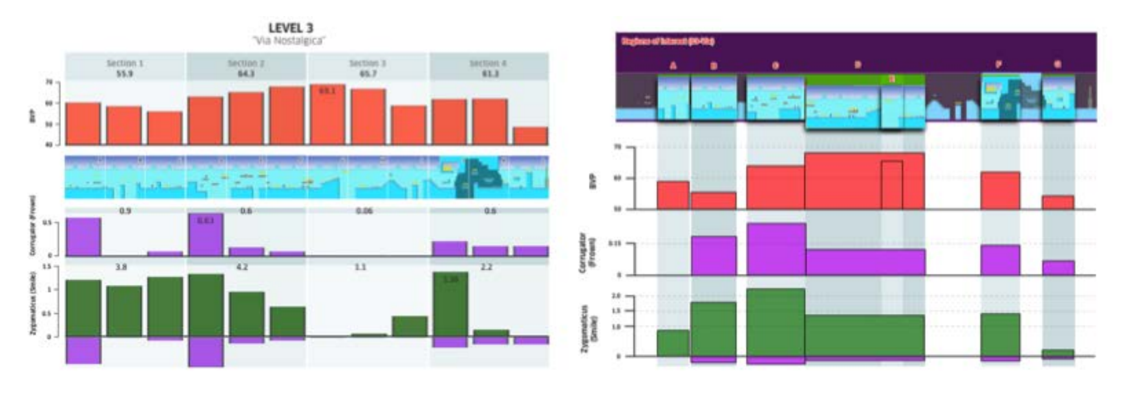


Fig. 19: Left - Biometric data superimposed over map sections of level 3. Right - Biometric data for regions of interest (ROI) in level 3

8.4.3 Results of Biometric GUR Data

The results of the biometrics data were graphs, depicted in the previous sub-chapter, which give a very clear overview of the data across the level.

We argue that these graphs give a unique insight into the flow of the level. By looking at the pattern of the values, it is easy to see if a regular flow in the readings has been established. In the graph of level 3 this regular pattern, in the form of a wave, seems to be most apparent. Since level 3 received the highest rating in terms of fun we argue that this wave pattern correlates to variation, which is considered favorable in level design. Modifications based on biometrics will therefore probably try to focus on sections breaking this pattern in the other levels.

8.5 Conclusion and Discussion of Phase 3

In this third phase or our study, we were able to not only evaluate data that was gathered during testing sessions, but also to assess the processes that can be used to lead to actionable results in terms of level design. As mentioned in chapter *8.1 A Case for Visualization of Data for Designers*, we argue that the visualization of data is crucial for communicating GUR results to designers. The methods we applied to visualize data are by no means complete, but have been applied to the best knowledge of the designer and researchers in this study. Such knowledge has been acquired predominantly through heuristic means and has been therefore described in such a way that allows the reader to follow our line of reasoning.

The addition of general player feedback into the evaluation of each methodology made it possible to put objective measures into context with basic feedback from the players. In retrospect, however, the use of a five point Likert scale resulted in a poor granularity of feedback, as participants tended to avoid the highest and lowest values of the scale. This left only three rating options, one of which was the midpoint. We believe a seven-point scale would have been a better choice. It could have provided more detailed feedback while keeping the amount of options readily comprehensible for participants.

Evaluating classic GUR data was a process that involved primarily manual work in form of transcription and interpretation of what has been said. We then systematically categorized the feedback, which made the process of evaluating classic GUR data very similar to that of evaluating metric GUR data. This similarity could perhaps be a necessity when it comes to processing feedback from a large group of participants.

Another noteworthy aspect of classic GUR data is the fact that different participants often mentioned the same situations in a level, but with contradicting feedback. It is not surprising that some participants might find, for example, a challenging situation too hard to overcome while other participants welcome the amount of challenge. It can however be hard to pinpoint which feedback to follow, given that many designers strive to provide an enjoyable game experience for a wide range of players. In addition to this, we also had players that commented negatively on the high amount of challenge while reporting a moderate to high frequency in playing video games in their personal lives. These players illustrate that it can be hard to determine which feedback to follow in cases of contradiction.

The evaluation of metric GUR data did not present such contradictions, but proved to be challenging in a different way, as it was the most time consuming one of the three GUR methodologies explored in this study. In order to interpret metrics data, we had to develop scripts that produced heat maps and derived measures - a time intensive endeavor. Even with this preparation of data, the subsequent analysis requires the designer to have clear goals to determine which metrics are desirable.

The evaluation of biometric GUR data requires similar insight from the designer. In this case, however, visualizing data was relatively simple as the biometric measures already corresponded to locations in the level. By superimposing these measures over the level as seen in *Figure 17, 18 and 19*, designers are able to get an insight into how their work is perceived on an emotional level. While we found it difficult to use biometric GUR data for very specific information, we see a potential for the analysis of level-spanning parameters, such as flow and the general rhythm of action and relaxation parts.

9. Phase 4: Level-Set Modifications

This chapter describes the changes that were made to the initial level set. We looked at the three possible combinations of two GUR methodologies:

- Classic and Metrics
- Classic and Biometrics
- Metrics and Biometrics

Each combination led to six changes to the level set each, giving us the three level sets for the last testing round. To remind the reader - the level designer had the general feedback data regarding fun, length and difficulty available for all modifications.

9.1 Modifications based on Classic and Metric GUR Data

The changes listed below were made based on the data we gathered from the Classic and Metric methodologies.

9.1.1 Classic & Metric Modification 1 (Level 1 - Section 2)

From the interview data we looked at the most frequent complaint made among the participants. This (INT-A01) related to the box with coins, suspended in the air in section 2 of level 1. The comments were that it was annoying and too hard to reach. These coincide with the reports of frustration and confusion, which both spike in section 2. Based on these observations we looked for data in the metrics of section 2 that would support change.

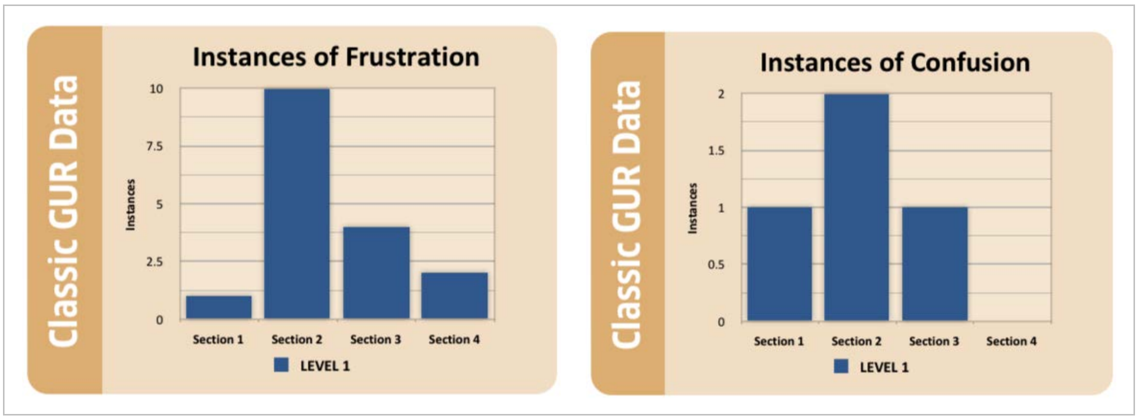


Fig. 20: Interview data on frustration and confusion

The metrics of this area show that this section has the shortest amount of time between jump instances in the whole level. This means a lot of jumping happens in this section. Additionally the amount of direction changes is the highest in this section. This suggests that the flow of the game is interrupted as participants go back and forth a lot.

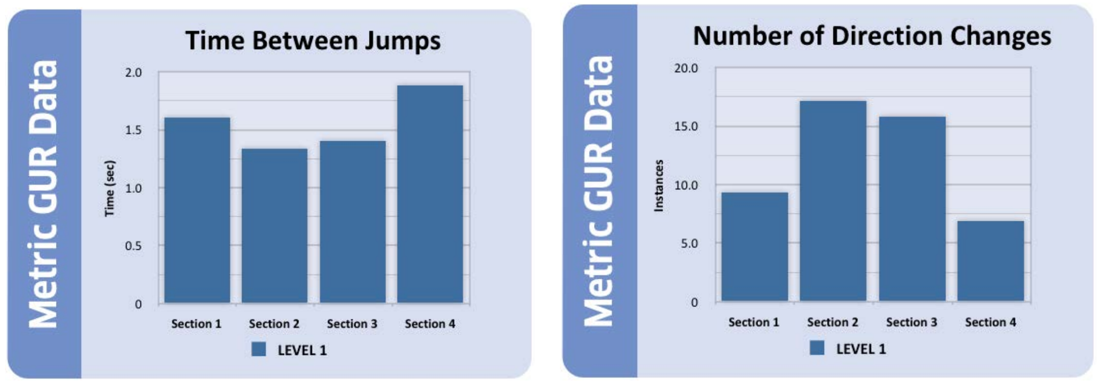


Fig. 21: Metrics data showing low time between jumps and high number of direction changes

When looking at the heatmaps of the level it becomes clear the player actions are not spread evenly across the level. Especially in section 2 the heatmap seems to suggest uneven activity. In changing this section it was our goal to make this section more balanced.

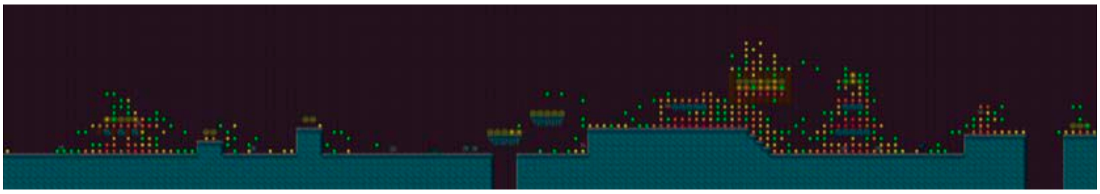


Fig. 22: Heatmap showing the amount of direction changes

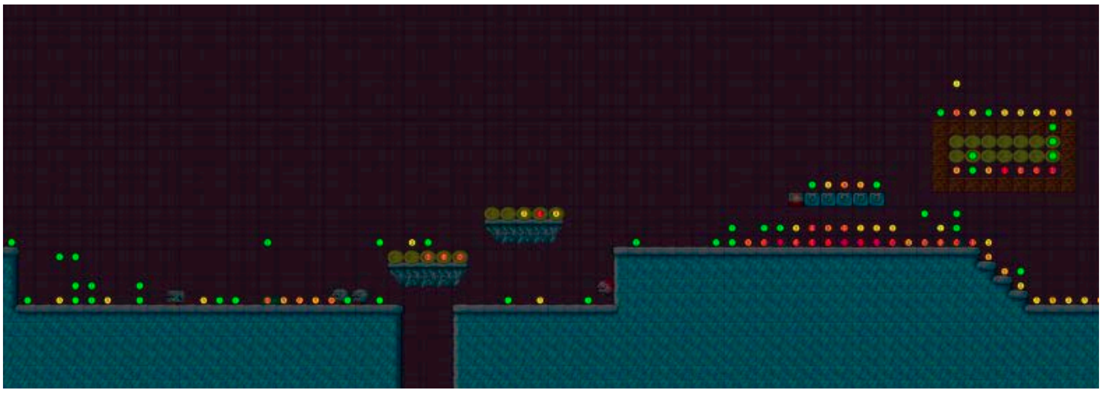


Fig. 23: Heatmap showing the amount of jumps from certain positions

Finally metrics showed us this section has the lowest percentage of time spent running, further suggesting that players are halted and the flow of gameplay is interrupted.

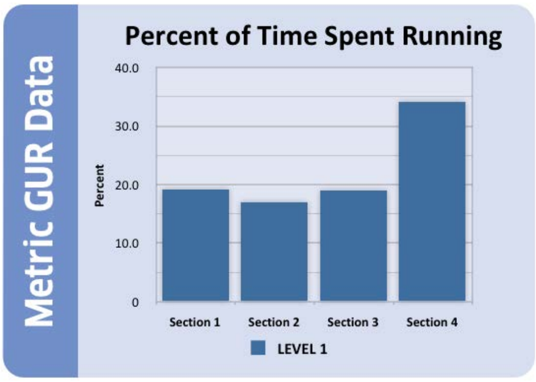


Fig. 24: Metrics data showing the percent of time spent running per section

Taking all the information we saw together, it can be suggested that a lot of jumping and direction changing (and therefore interruption of running) happens around the box of coins, which is therefore a serious flow breaker. The fact that many participants found the box annoying shows that they spent time trying to reach it (as supported by the metrics) but did not in the end, leaving them unsatisfied. When looking at the area preceding the box, the heatmap suggests a lack of challenge. Since there is a low presence, it would appear players went through this area very quickly.

The first change we decided on was to move the enemy. When looking closer at the heatmap is shows only a few players actually reached the area where the enemy was, making her somewhat superfluous. We therefore decided to move her to an area where the player would definitely go.

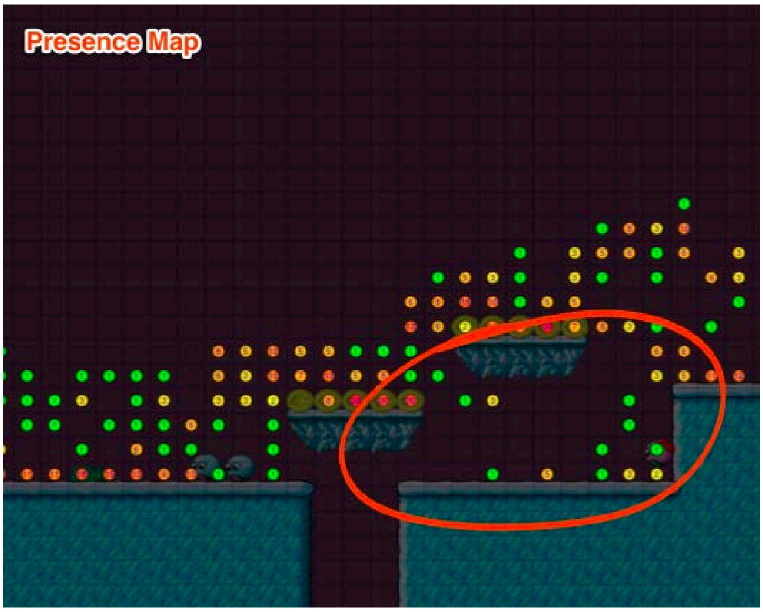


Fig. 25: Heatmap showing low player presence where enemy is

The next change was to move the platforms. The heatmap shows that almost all players make it across the platforms, taking away any challenge the gap could provide. With moving the platforms we aimed to spread out the amount of jumping in the section, as well as adding a bit more challenge by making the gap more open.

The final big change was in the box of coins, which we made more accessible by removing some of the boxes and moving the whole box down a row of tiles.

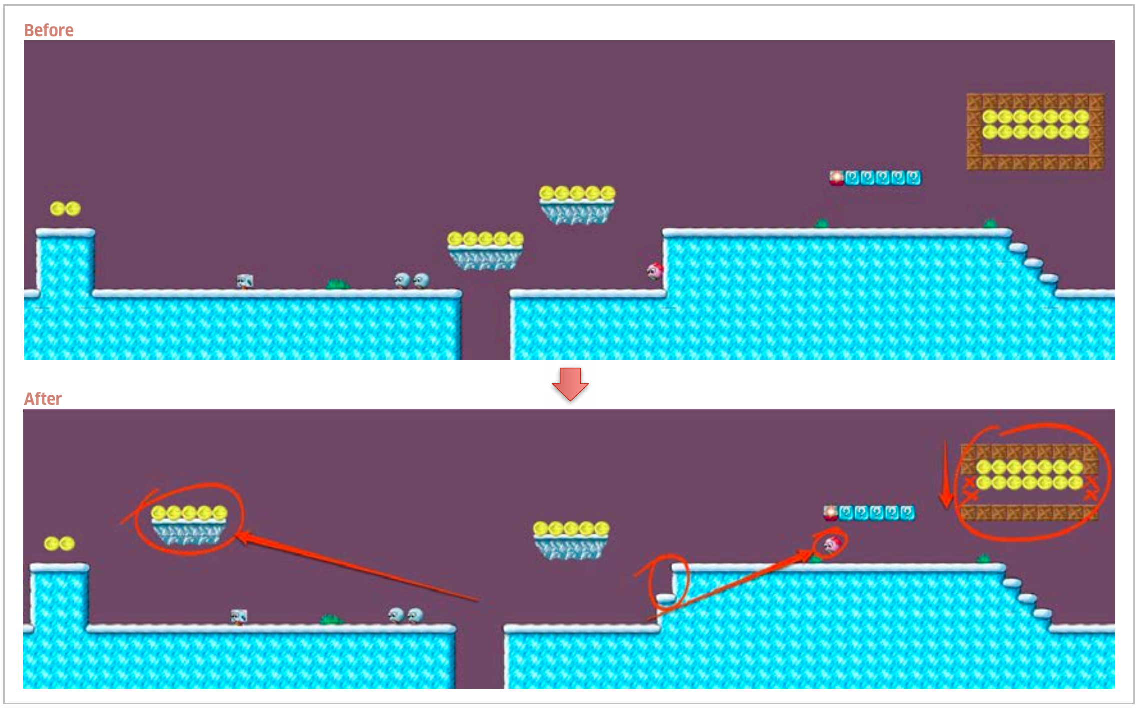


Fig. 26: Changes in Level 1 – Section 1

9.1.2 Classic & Metric Modification 2 (Level 2 - Section 1)

Based on metrics we found that less than half (only 43%) of the players found the secret area in level 2. This is the lowest find rate of a secret area across all three levels. From interview data we learned however that players highly enjoyed finding secret areas. Mention was also made of the fact that players liked knowing a secret area was nearby, with the added challenge of figuring out how to get there.

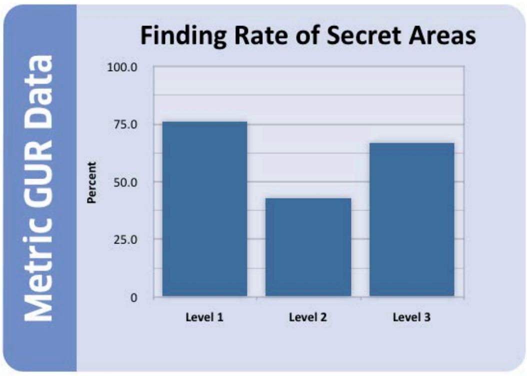


Fig. 27: Metrics data showing finding rate of secret areas across all levels

We therefore decided to make the secret area more accessible and obvious. The challenge in this was in making it still a surprise for the player to find the area, but also make it harder to miss. We decided to add another entrance on top of the area, which seemingly does not lead anywhere. The idea is the player needs to go out of their way to destroy the boxes, without an immediately visible pay-off, and then falls through into the secret area.

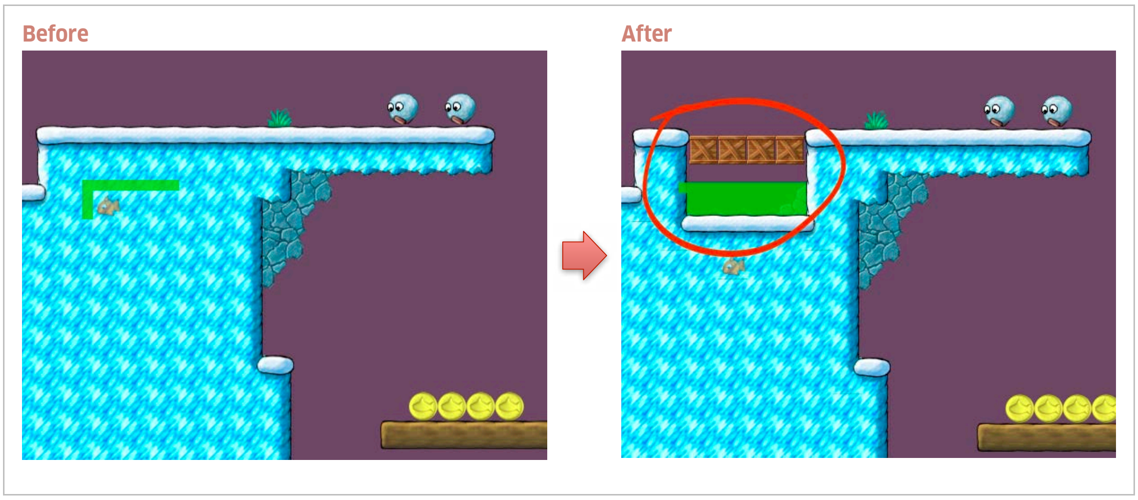


Fig. 28: Changes to secret area in Level 2 – Section 1

9.1.3 Classic & Metric Modification 3 (Level 2 - Section 2)

From interview data we noticed that section 2 of level 2 had the highest amount of reported frustration. It also featured the highest amount of surprise, although it should be noted that the amount of reported surprise was low overall compared to the other levels. Players suggested in the interviews (INT-B07) that the area was too challenging, while also mentioning that coins in the beginning of that area were too hard to reach (INT-B05). In spite of this, some players also mentioned the area was not challenging enough (INT-B03).

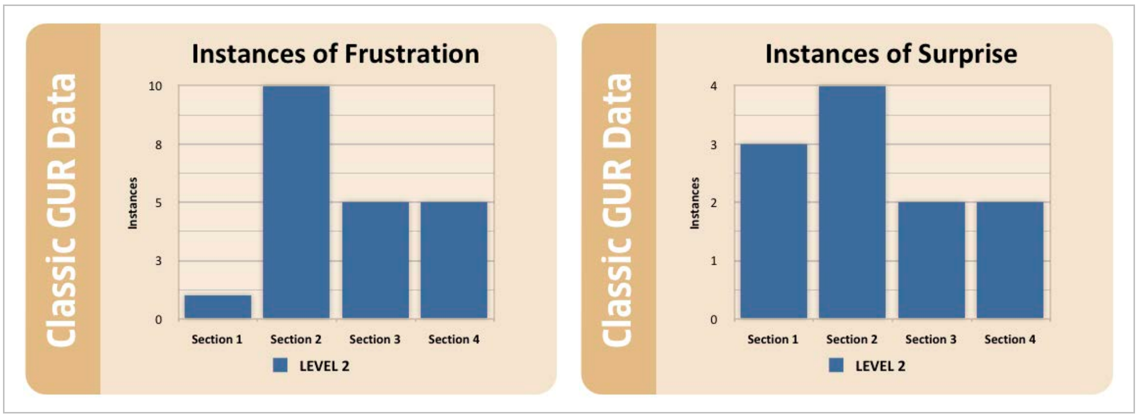


Fig. 29: Instances of frustration and surprise for level 2

When looking at the death heatmaps of the section we saw an increased amount of deaths in the section, they were however not concentrated on one single spot. When looking further into the player deaths in this section we saw that most of the deaths in section 2 were because of falling down.

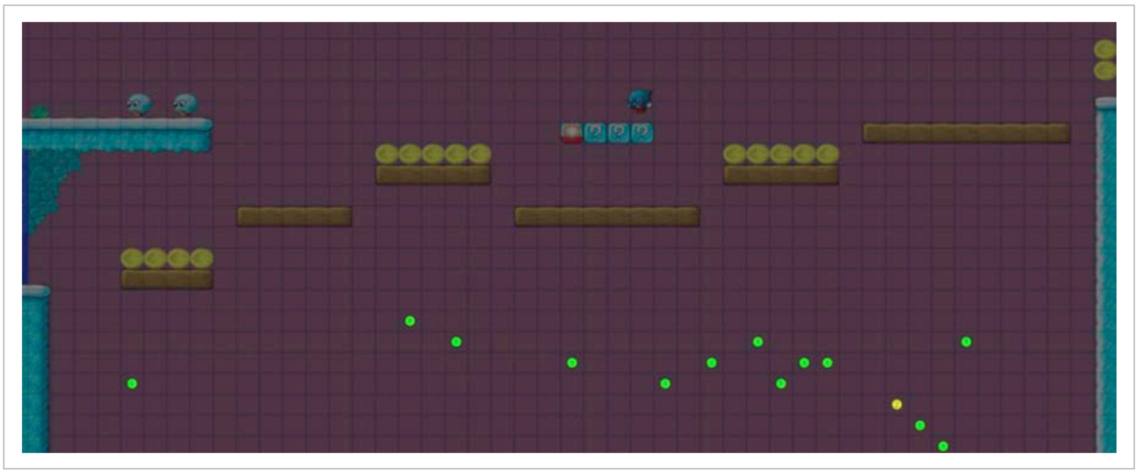


Fig. 30: Heatmap showing scattered deaths in section 2

Furthermore, metrics showed that the time between direction changes in section two was the lowest across the level with an average of 2.88 seconds. We argue that for this section this showed that players were required to modify their jumps a lot and use correction movements to get the coins. It follows that the section generally has a poor flow.

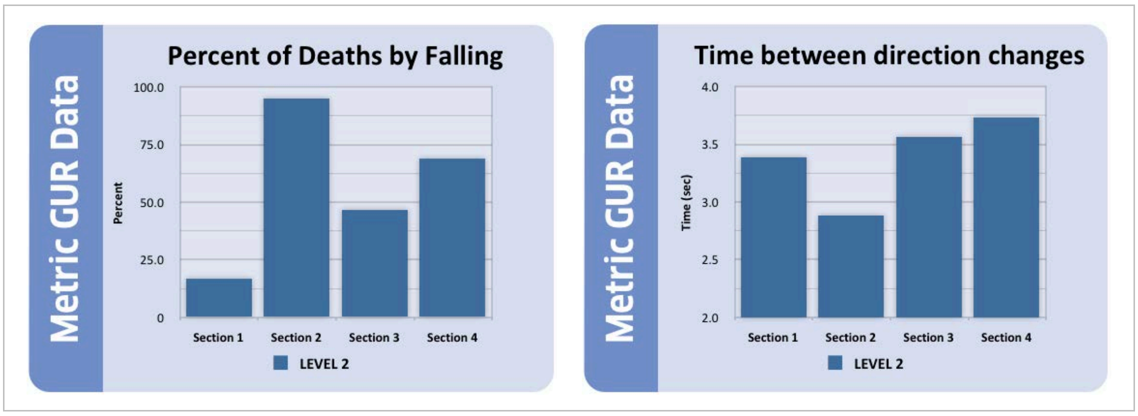


Fig. 31: Metrics data on deaths by falling and time between direction changes

Looking at the percent of time that was spent walking as opposed to running further supports this theory.

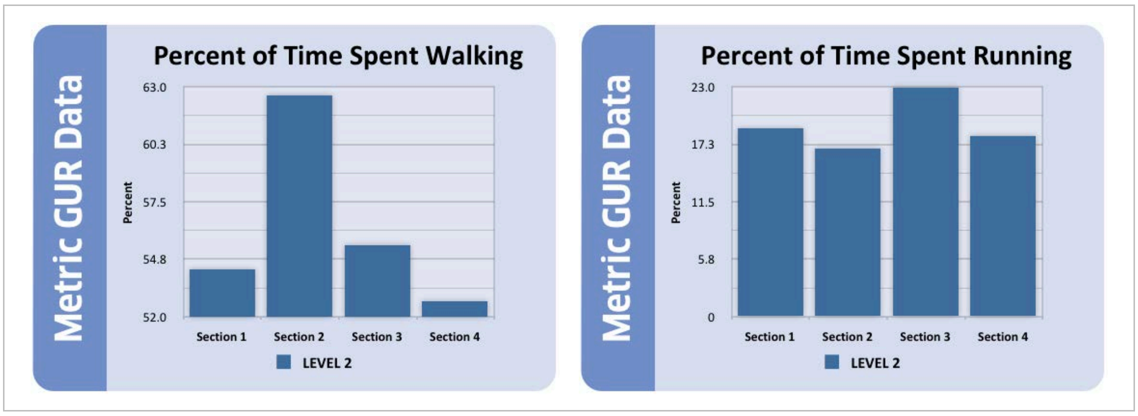


Fig. 32: Metrics data on the percent of playtime spent walking and running

By combining the data, we argue that the reason for frustration in this section stemmed from a lack of flow and frequent deaths from trying to collect all the coins. Direction changes showed that players had to adjust their position a lot yet they still fell down often. The fact that the deaths were spread out over the area also showed the challenge was not very focused, possibly adding to the frustration.

To improve the section we decided to space the platforms further apart and thereby present the player with a clearer challenge, while simultaneously reducing the amount of gaps. Coins were spaced further away from the edges of the platforms to decrease the risk of falling when trying to get all coins. Some coins were moved atop the question mark platform to create additional challenge for those who seek to collect all coins.

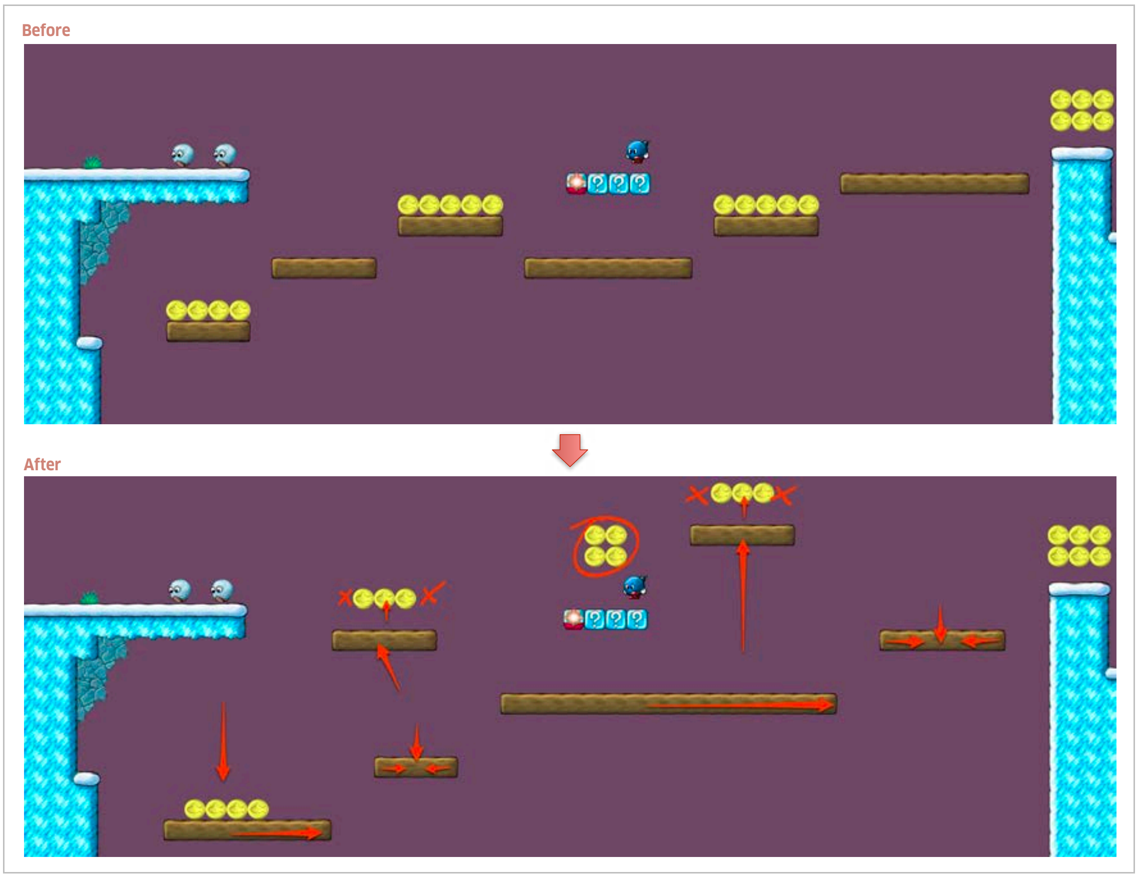


Fig. 33: Changes in Level 2 – Section 2

9.1.4 Classic & Metric Modification 4 (Level 2 - Section 3)

Based on interview data, we knew that there was a low amount of enjoyment in section 3 of level 2. Additionally, we noticed a relatively high amount of frustration in this section, although not the highest within the level.

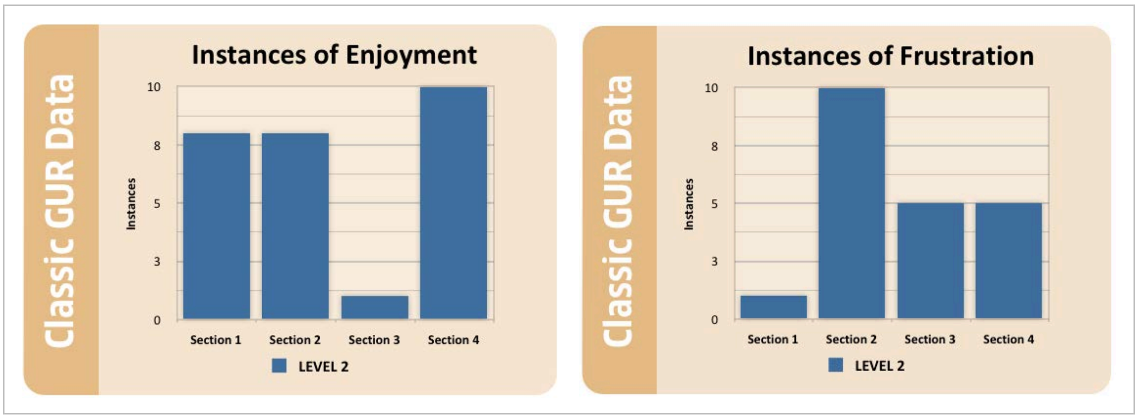


Fig. 34: Instance of enjoyment and frustration across Level 2

Based on observation, we know that a lot of players lost the upgrade snowball because it moved to the left and fell down the gap. Players trying to still get it tended to fall down the gap and die. Based on the metric heatmap, we see that the gap next to the platform is responsible for most deaths at a single spot, while not being too big or challenging on its own.

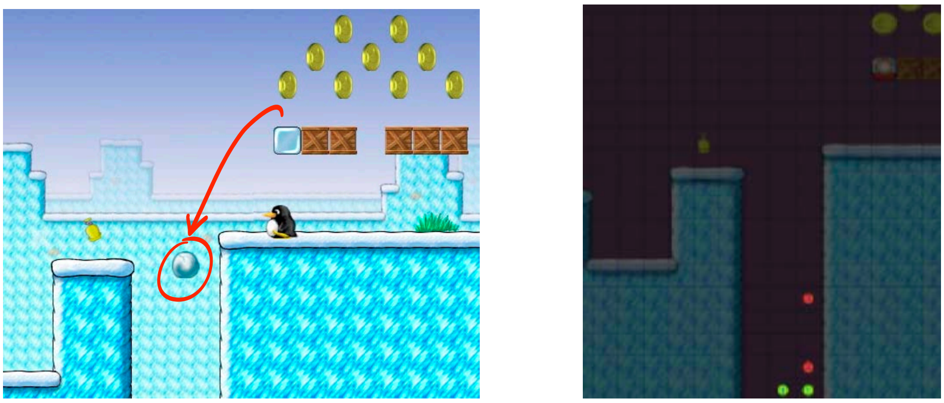


Fig. 35: Left - Game screenshot showing power up falling down. Right - heatmap of deaths in pit

To solve this problem we switched the hole on the left with the bottomless gap. By doing this the power-up can always be obtained without causing the player to fall to his or her death, but the challenge should stay very similar.

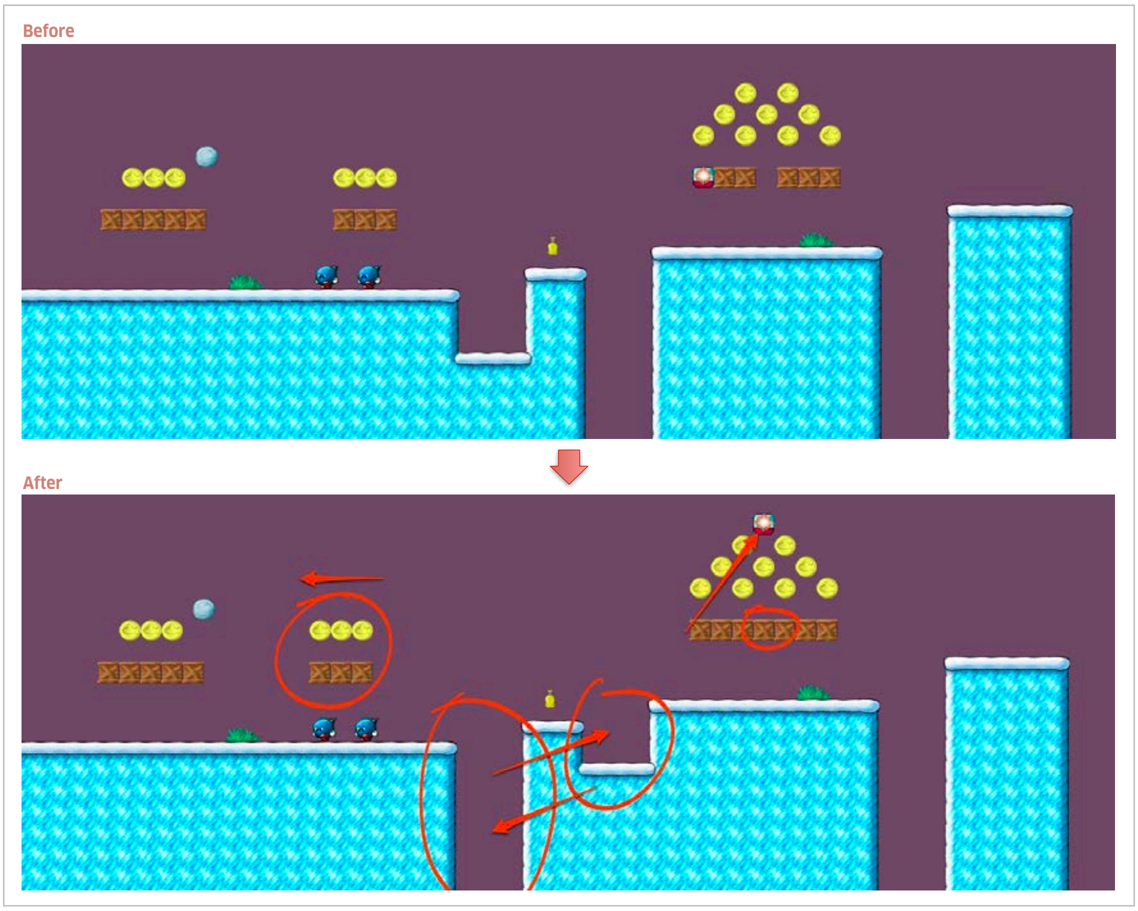


Fig. 36: Changes in Level 2 – Section 3

9.1.5 Classic & Metric Modification 5 (Level 2 - Section 3)

From the metrics, we saw a disproportionately high amount of deaths occurring at the flying snowball enemy at the end of section 3. This was further illustrated by the death heatmap of the section. Furthermore, the interview data (INT-B01) showed the participants often specifically mentioned the enemy as a negative factor.

In order to keep the amount of variables consistent, we decided to keep an enemy at this position, but replace it with an enemy that generated a more positive response from players: the ice-cube enemy.

To address the low amount of enjoyment in the section – and the low amount of fun in this level – we made it so that killing the ice-cube enemy should knock him against the next platform and either kill off an enemy, or destroy the question mark blocks.

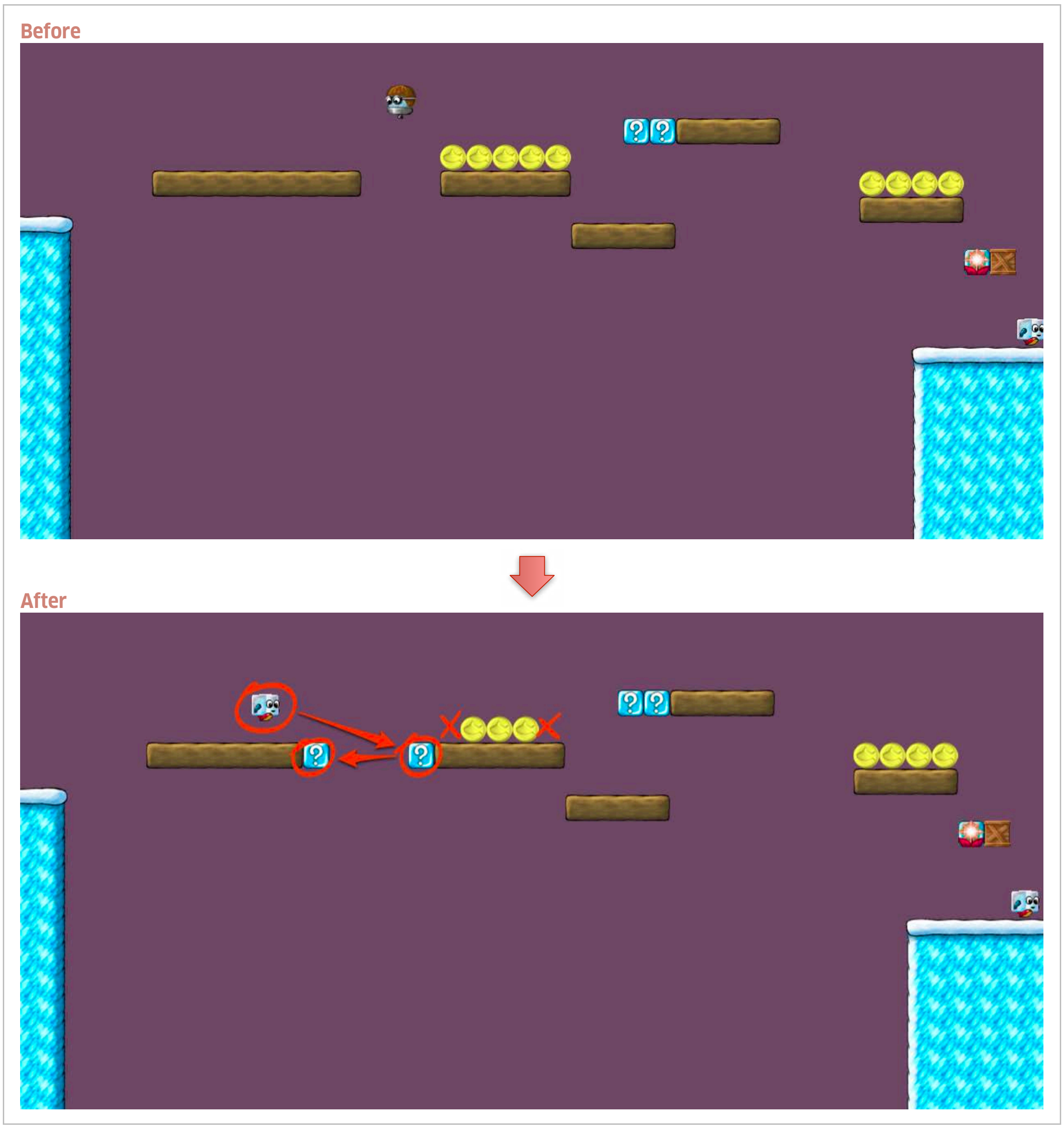


Fig. 37: Changes in Level 2 – Section 3

9.1.6 Classic & Metric Modification 6 (Level 3 - Section 2/3)

Observations of the test sessions showed that the checkpoint position in level 3 presented a challenge to obtain, especially to lesser skilled players (INT-C03). Generally, both section 1 and section 2 in the level were mentioned as challenging sections during interviews. While the challenge has been generally liked across players (as shown by a high rating in terms of fun of the level), we decided it would be a welcome change to distribute the challenge better over the level.

Metrics data supported the observation due to the fact that 76% of the deaths in the level occurred before the checkpoint. The metrics for this level count a high amount of deaths to begin with, giving weight to the conclusion that the checkpoint should be moved to a point earlier in the level.

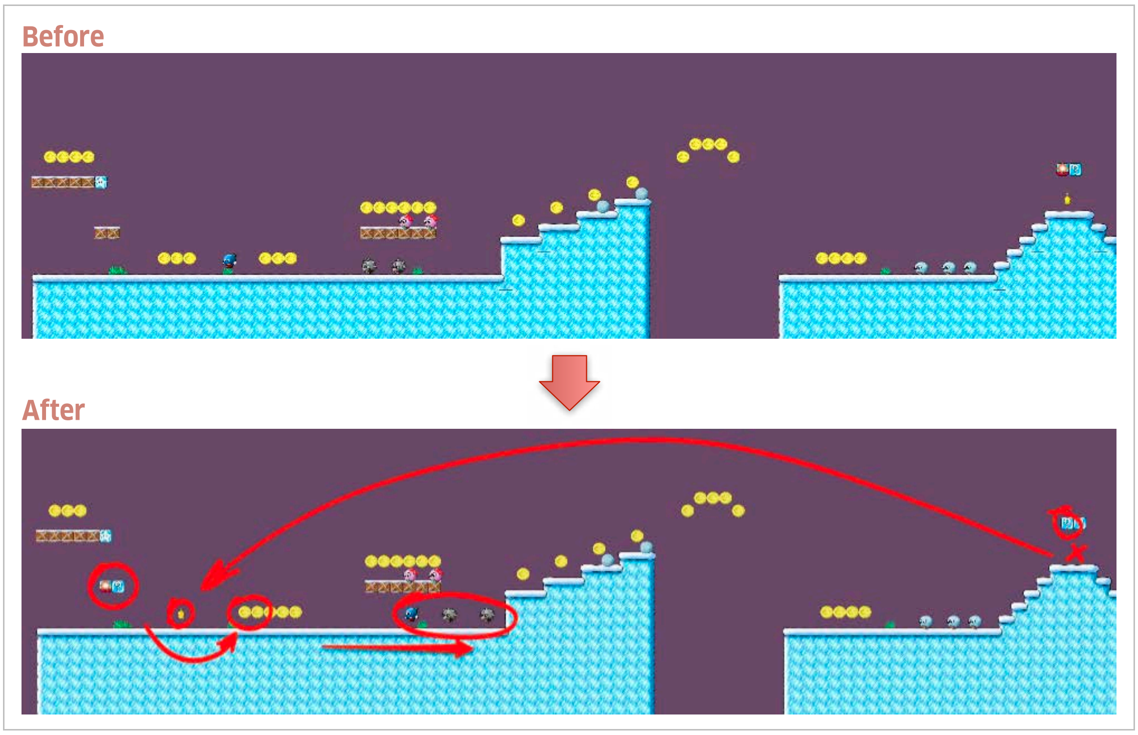


Fig. 38: Changes in Level 3 – Sections 2/3

9.2 Modifications based on Classic and Biometric GUR Data

The following modifications were based on the correlations between classic and biometric GUR data. Given the size of the biometric visualizations, the biometrics data that is referred to in the following sub-sections can be found in the appendix.

9.2.1 Classic & Biometric Modification 1 (Level 1 - Section 1)

Looking at the biometrics data of level 1 showed that there were few sections in the level that were very remarkable. Section 1 showed the lowest average heart rate, despite a peak at the very beginning, which we argue was from the excitement and anticipation of starting the experiment rather than from the game. It also had no recorded frowns and the second lowest amount of smiles. From this we inferred that the amount of involvement of the participants was very low and the section was therefore not very interesting or challenging.

Interview data showed that section 1 was rated consistently low across all measures (frustration, enjoyment, surprise and confusion). We argue that this in combination with the biometrics suggests that the players were very uninvolved in section 1 and it did not stand out in any way.

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 coins, yet there is no increased risk of death, which would not be desirable at the beginning of the first level.

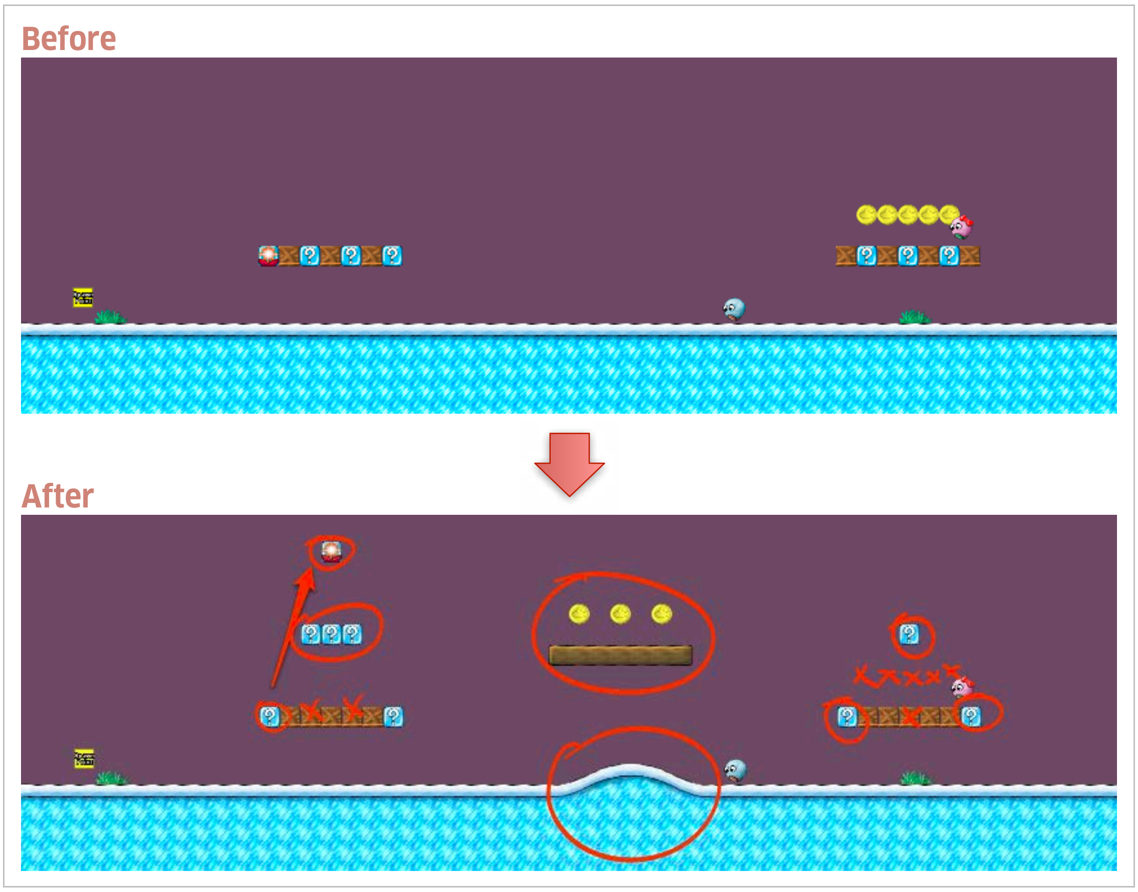


Fig. 39: Changes in Level 1 – Section 1

9.2.2 Classic & Biometric Modification 2 (Level 1 - Section 1/2)

The biometrics data from level 2 showed us that there were no smiles or frowns in section D, the first sub area of section 2. This contrasted with the rest of section 2 which did generate an emotional response from the participants. Section D also had the lowest BVP rate of the section, with a value of 66.6.

Again, interview showed us that players found level 1 too easy and boring. This was also shown by the low amount of feedback given on the level.

As a solution we firstly attempted to make the section appear more interesting visually by adding platforms and greater height differences.

Secondly we gave the impression of greater challenge. From observation we know most players did not have problems with the jumps across the two ice platforms. We therefore decided to make the gap below the platforms bigger. While the jump essentially remains the same, we believe the jump will feel more challenging without the safety net.

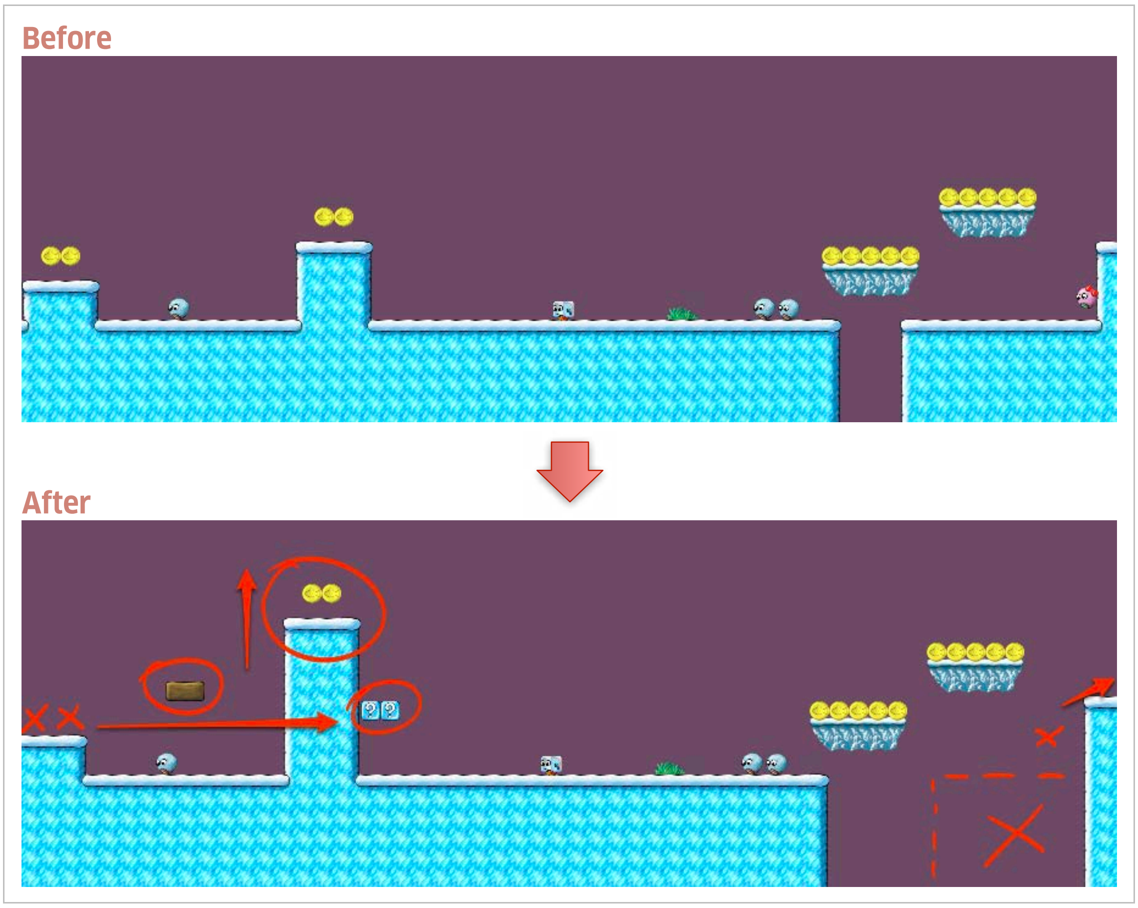


Fig. 40: Changes in Level 1 – Sections 1/2

9.2.3 Classic & Biometric Modification 3 (Level 2 - Section 2)

From interview data it showed that section 2 of level 2 accounted for the most frustrating moments in the level (see *Figure 29*). Biometrics confirmed this by showing the lowest amount of smiles as well as a relatively low amount of frowns. Although the frowns seem at odds with the amount of frustration reported, we argue that frowns are not always a sign of negativity. Rather we interpreted it in a way that the section lacked challenging or interesting aspects.

However, interview data indicated that section 2 was described as challenging. Another issue mentioned was that the coins in the beginning of the section were hard to pick up. Overall the biometric data did not seem to support this since, additionally to the low amount of frowns, the BVP mean in section 2 was the lowest of all sections in level 2.

Given the contradiction of the data we theorized that rather than actually being very difficult, the challenge in section 2 is perceived to be unfair. Instead of failing while trying to overcome a clear challenge, deaths are caused by accidental falling when trying to collect the coins (INT-B05).

Based on this idea we decided to move the coins from the edges of the platform to make them easier to collect, reducing the risk of falling and therefore the amount of unfair deaths.

Since the interview told us players did not desire an increase in challenge in this section, we had to think of a different way to provoke more of an emotional or physical response from the players. We decided to rather than increase the actual difficulty we wanted to create an increased perception of difficulty. This was done by spacing the platforms differently, utilizing more vertical space.

Another decision was to change one of the platforms to a solid surface, breaking up the original long sequence of wooden platforms and therefore focusing the challenge.

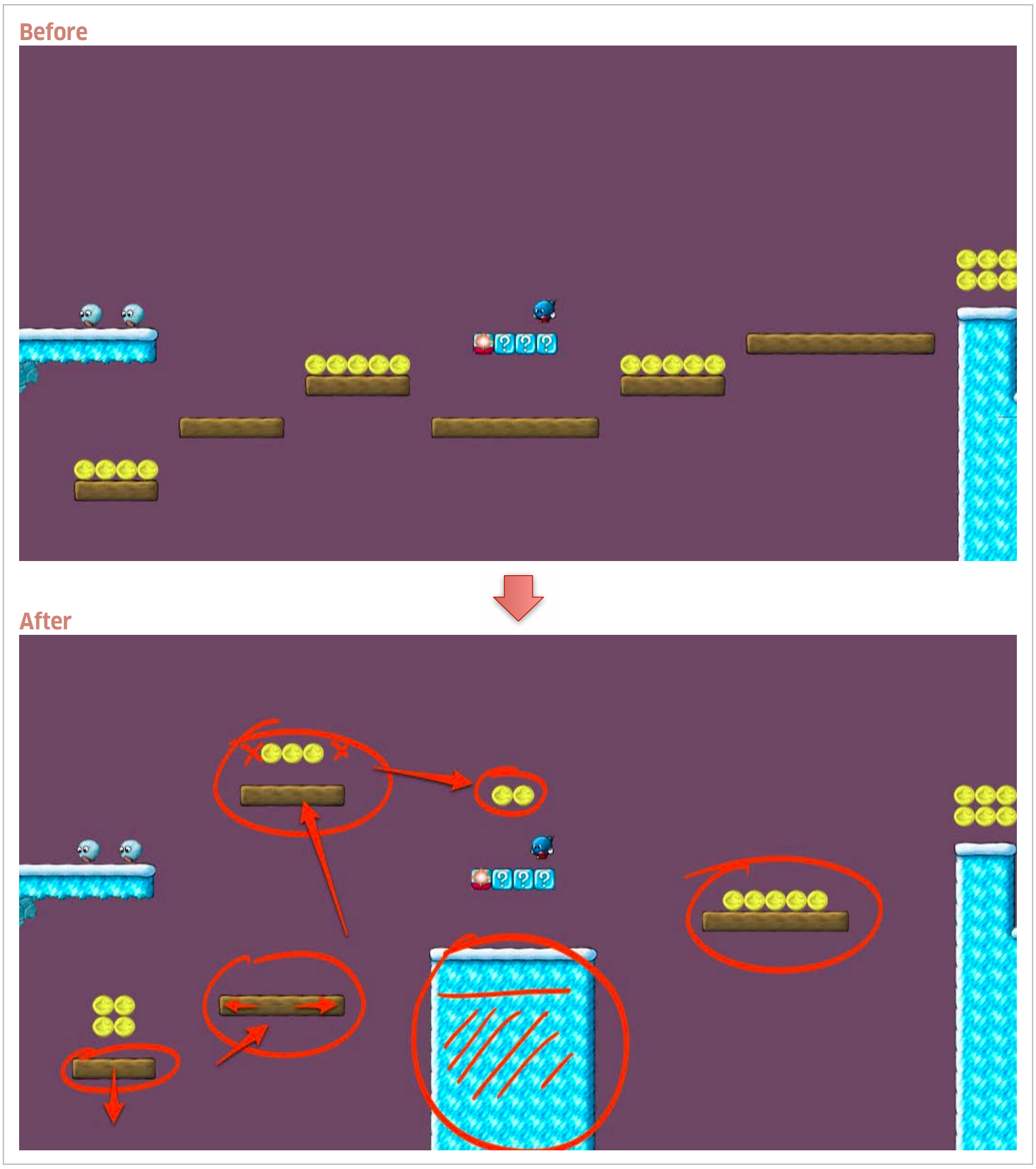


Fig. 41: Changes in Level 2 – Section 2

9.2.4 Classic & Biometric Modification 4 (Level 2 - Section 2)

Interview data suggested level 2 needed more challenge, felt short and was not novel enough compared to level 1. Biometrics seemed to support this, especially with low emotional feedback at the end of section 2 and the beginning of section 3.

Since there is very little happening in terms of vertical space in these sections we decided to add an extra element. This provides the player with an extra option on how to approach the obstacles and get more use out of the star power.

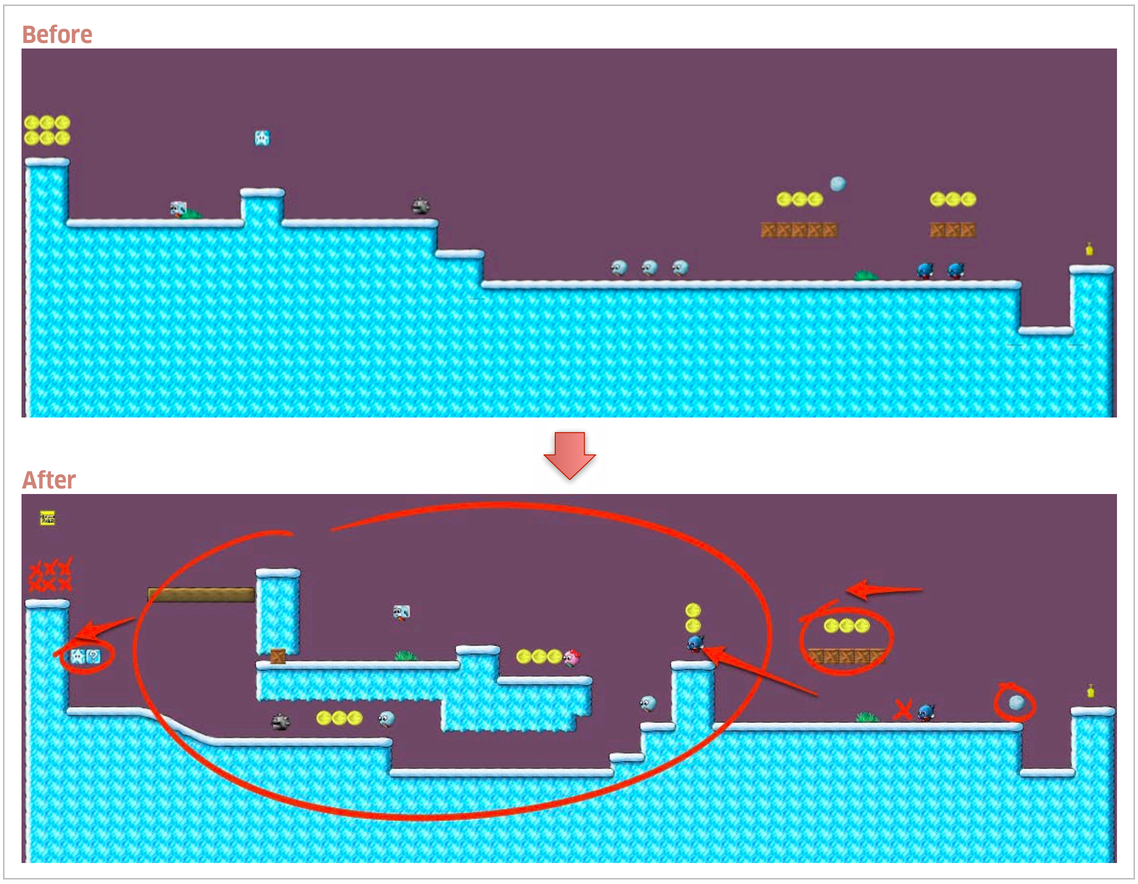


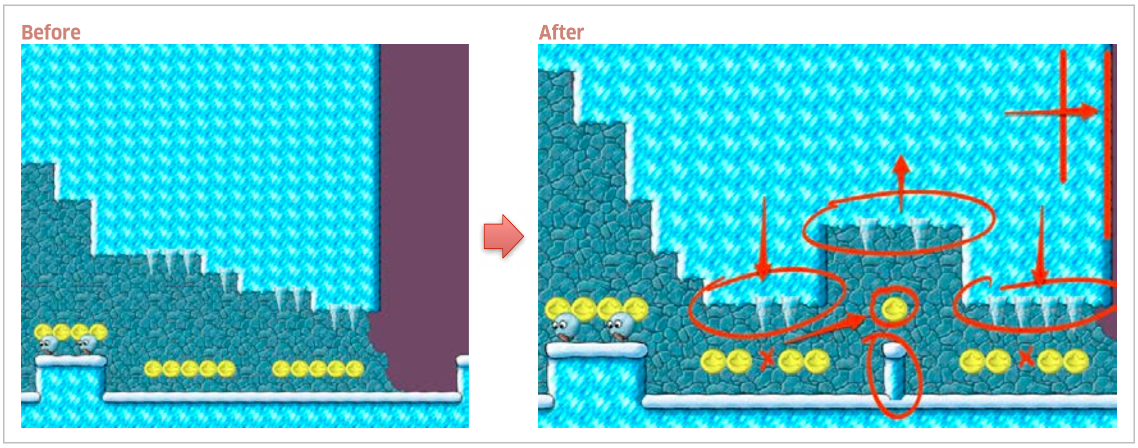
Fig. 42: Changes in Level 2 – Section 2

9.2.5 Classic & Biometric Modification 5 (Level 2 - Section 4)

Biometrics in section 4 of level 2 showed that the BVP dropped gradually over time. There is however an icicle chain in this section designed to provide a challenge and exciting point for the players. Biometrics did not reflect this at all, suggesting the danger was simply too low and/or short.

Interview data mentioned the icicle chain as a positive point, yet in general the level was perceived as too easy and too short.

We decided therefore to change the section to add more of a challenge. We did not want to add more threat, so the amount of icicles and enemies stayed the same. However, the space was designed in a way to bring the icicles closer to the player, therefore increasing the challenge and sense of danger.



<u>Fig. 43:</u> Changes to icicle chain in – Section 4

9.2.6 Classic & Biometric Modification 6 (Level 2 - Section 4)

As addressed in the previous modification, section 4 of level 2 lacked physical response from players, as we could see from the biometrics data. The BVP drops gradually, despite this area being designed as challenging. Interview data backed this up with reports of the level being too short and easy.

Next to changing the icicle chain we decided to change the very end of the level. In the original level it was just a set of straight jumps, not requiring much skill from the player. The way we changed it required the players to jump more precisely. The most difficult jump is at the top, which is an extra challenge for players wanting to collect all the coins.

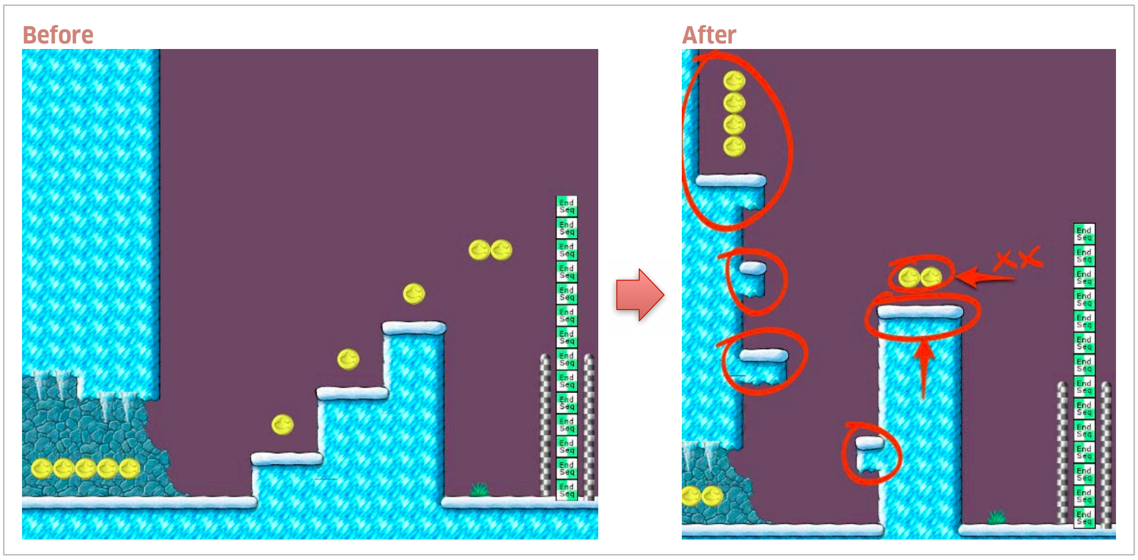


Fig. 44: Changes to jumping challenge Level 2 – Section 4

9.3 Modifications based on Biometric and Metric GUR Data

The following sub-chapters describe the modifications based on the data collected with biometrics and metrics.

9.3.1 Biometric & Metric Modification 1 (Level 1 - Section 1/2)

Biometrics data directed us to section 1 of level 1. This section had the lowest BVP mean in the level, despite the initial peak, most likely caused due to the tension of starting the experiment. Furthermore the drop in BVP was the sharpest here compared to the other sections of the level. The amount of smiles in this section was also very low. Most likely the only reason it was not the lowest amount was because participants generally smiled at the beginning of each level.

Metrics for this section showed the second lowest amount of jumps, the second shortest completion time (despite the player starting the section from standing still) and the second lowest count in direction changes (see *Figure 21*). Section 4 was the lowest in all these measurements, but here players obtained the star power. Wanting to get the most use out of it could explain the straightforward behavior. No such behavior was designed into the first section.

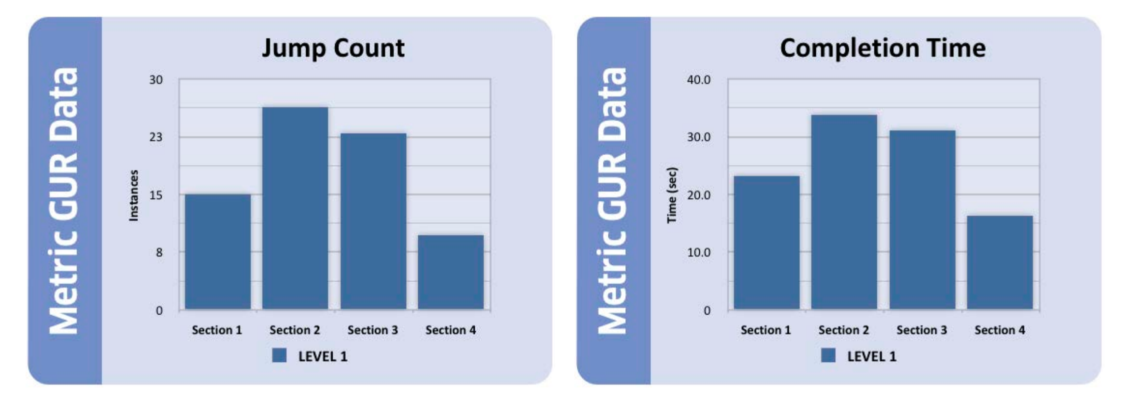


Fig. 45: Metrics data on completion time and jump count in Level 1

We decided to utilize the vertical space more. Some coins were moved up, requiring the player to jump more if they wanted to obtain them. Some elements were also moved closer to the beginning, to break up the long stretch of flat surface that started out the level.

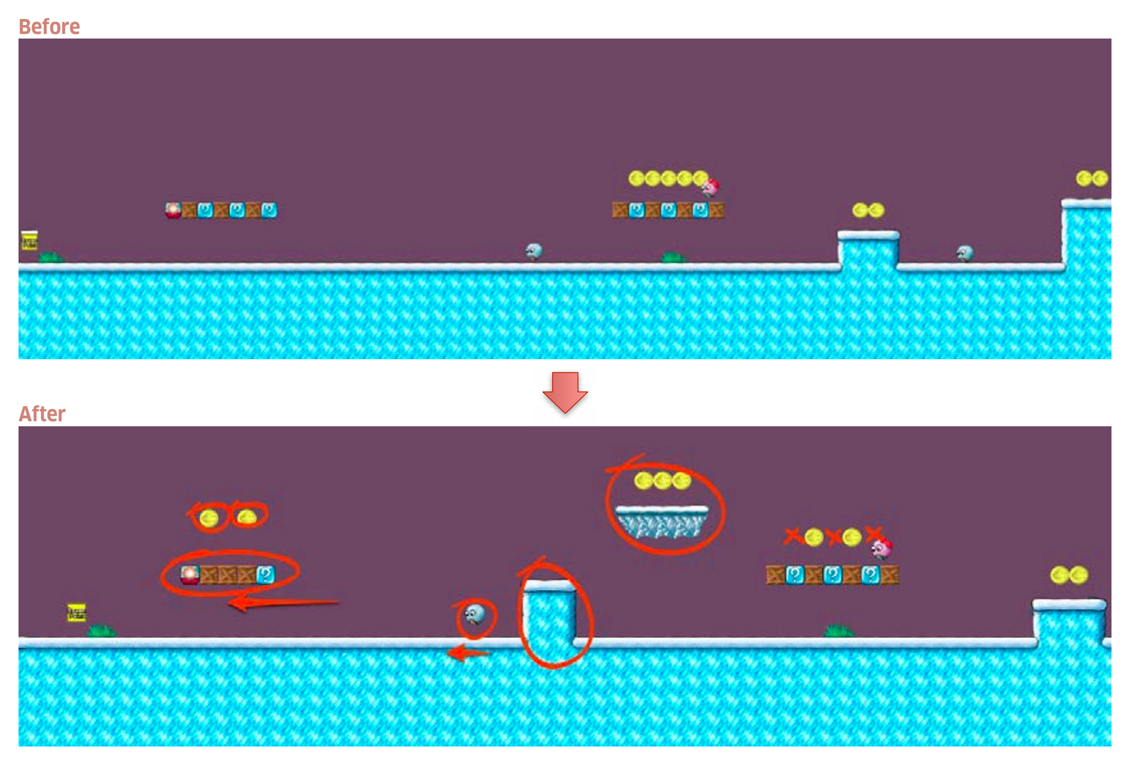


Fig. 46: Changes in Level 1 – Section 1/2

9.3.2 Biometric & Metric Modification 2 (Level 1 - Section 4)

Based on metrics we knew that section 4 of level 1 had the shortest completion time and the lowest amount of jumps and direction changes. At the same time biometrics revealed minimal emotional engagement in this section, showing minimal frowns and the lowest amount of smiles.

We believed this lack of emotional response came from not being able to utilize the star power much. Therefore we wanted to increase the amount of challenge, which in turn could increase the perceived amount of time the players could make use of the star power. We did not try to achieve this by adding more danger, but rather adding obstacles, making the player do more than just run in a straight line like it was the case in the initial level.

Generally, anything that adds to the perceived length of the level is supported by the results of the survey at the end, as well as the addition of elements that could potentially raise the amount of fun.

We added a large surface, giving the players two paths to choose from and distributed the coins between the paths. In order to get all the coins the player would need to backtrack a little. It should be very possible to get all coins while being in star power, but it will ask for a bit more skill from the player.

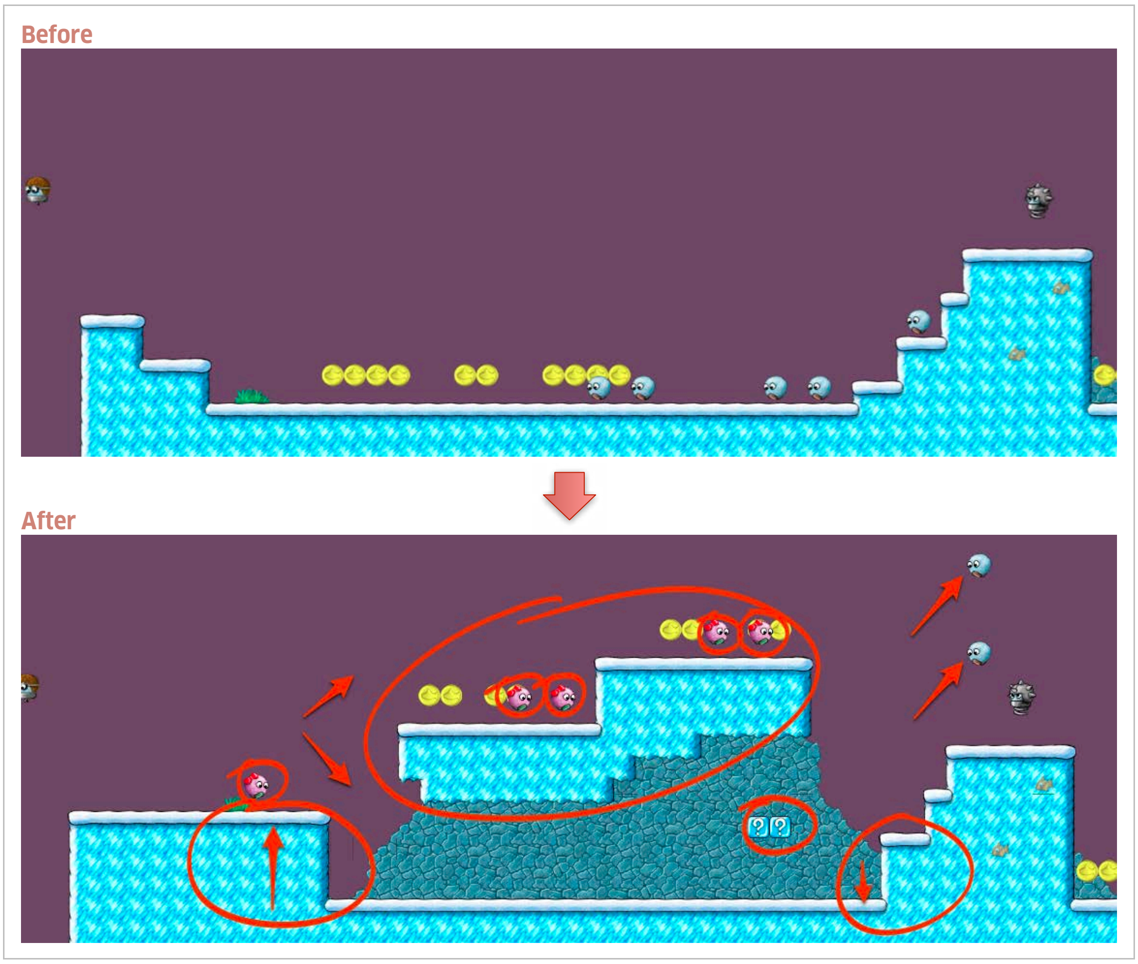


Fig. 47: Changes in Level 1 – Section 4

9.3.3 Biometric & Metric Modification 3 (Level 2 - Section 1)

Participants rated level 2 as the least fun level of the three. Biometrics seemed to confirm this with the lowest smile-to-frown ratio and with the lowest overall BVP mean. Generally, level 2 seemed the least exciting level in the set. In section 1 and 2 we could see a dip in smiles as well as frowns that picked up towards the second half of section 3.

From metrics data we knew that only 43% found the secret area in level 2, also by far the lowest amount across all levels (see *Figure 27*).

From a design standpoint, we would have expected to see an emotional response from players finding the secret area. The subsection in which the secret area is placed has a low amount of smile peaks. We therefore wanted to change this section to give players more of an idea about whether a secret area could be close.

Since the data of this combination pointed to the same issue as the interview and metrics combination did we made the same change here as we did in chapter *9.1.2. Classic & Metric Modification 2 (Level 2 – Section 1)*.

9.3.4 Biometric & Metric Modification 4 (Level 2 - Section 2)

Based on general ratings level 2 had the lowest rating for fun of all levels. It also showed the lowest smile-to-frown ratio as well as the lowest BVP mean. Based on this we concluded level 2 was the least exciting level of the three. In section 2 and 3 we could see a dip in both smiles and frowns that only picked up towards the second half of section 3. This suggested that especially the beginning of section 2 should be more interesting and/or challenging.

General feedback told us that level 2 was perceived not just as the shortest level, but even to be too short. Therefore our efforts always had to look at the possibility to extend the playtime. Secondly, as mentioned in chapter *8.4.3 Results of Biometric GUR Data*, when looking at the biometrics data over the whole level we would like to see a regular wave pattern within the readings, suggesting a natural flow of more challenging or exciting parts and recovery moments. At this moment there were breaks in the flow in section 2 and towards the end of the level.

Looking at the death heatmap provided by metrics, we could see the deaths in section 2 were spread out over the section. At the same time, section 2 had the second most deaths in the level, after section 3. While we have seen that deaths are often accompanied by smiles, this was not the case here. This was especially interesting since while most deaths in this section were because of falling, correlation metrics show that players had more fun with death by falling than dying because of an enemy. We theorized that this is because the enemy is an external factor, while falling is in complete control of the player. Based on this we would have expected to see more smiles in this area.

We believed this lack of emotional response as well as the scattered death occurrences pointed to a general lack of focused challenge. We observed that deaths that were accompanied by smiles were usually because of a clear obstacle the player failed to overcome. The fact that there was no localized point for the deaths as well as no real emotional response suggested that deaths might be perceived as unfair or accidental rather than because of player error.

Based on this we decided the challenge in this section should be more focused. We tried to make the jumps clearer by adjusting the position and size of the platforms. We also added the additional challenge for players wanting to collect all the coins to reach the upper platform as well as changing some coins to coin blocks in a trickier position.

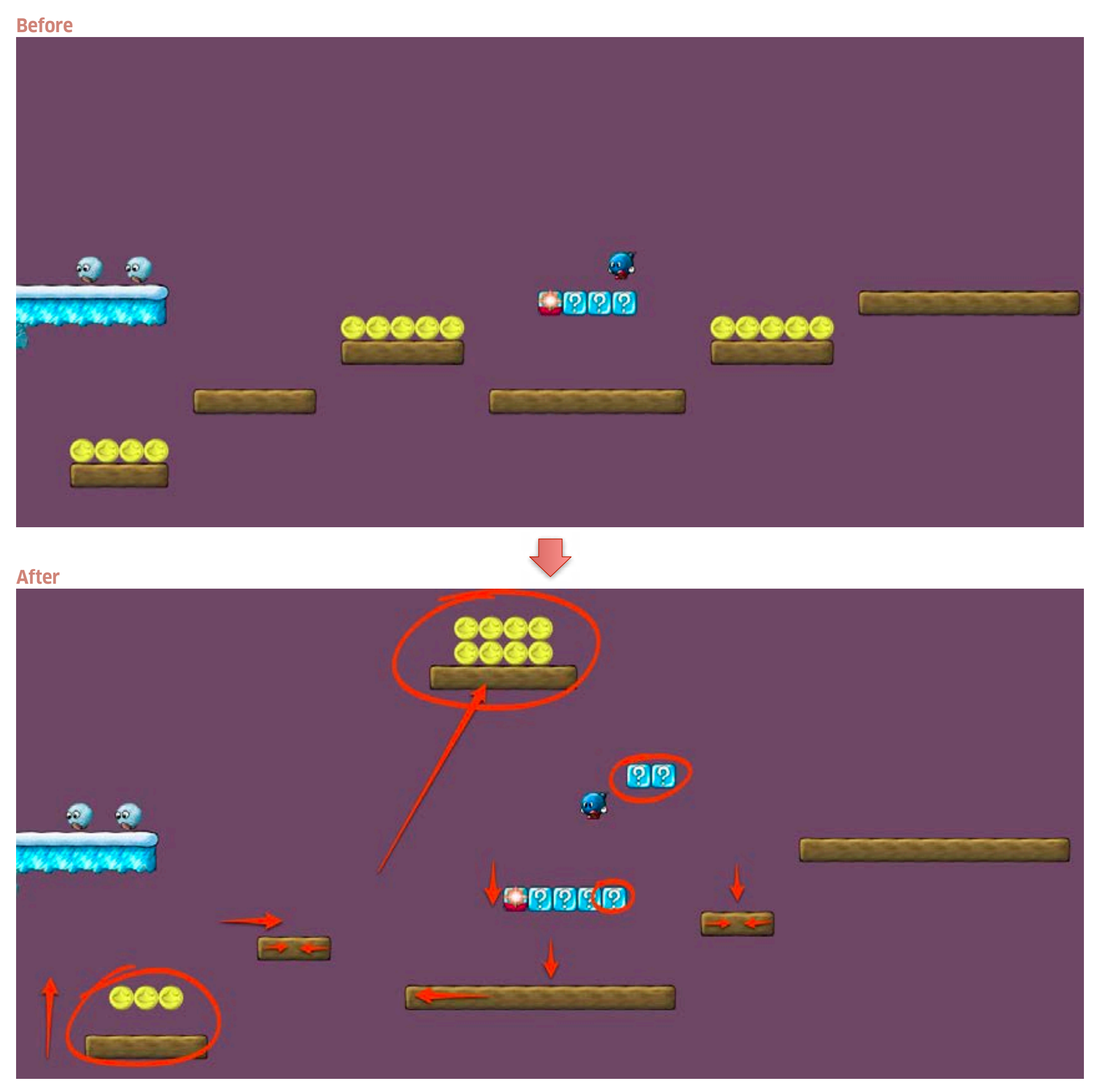


Fig. 48: Changes in Level 2 – Section 2

9.3.5 Biometric & Metric Modification 5 (Level 2 - Section 4)

Biometrics showed no frowning and a drop in smiles towards the end of level 2. While section 4 was designed to be challenging, the emotional feedback is not showing much evidence of this. As mentioned with the previous modification we wanted to aim for a wave pattern in the biometric readings. The ending was one section that broke the flow; the level seemed to end on a low note instead of with a satisfying bang.

Metrics showed us that the icicles only accounted for a single death among participants, while these were intended to be a challenge. They also showed that this section had by far the lowest amount of mode shifts of the level. This meant the icicles very rarely hit the players and therefore the challenge could be increased.

The data from this combination of GUR methodologies pointed to the same problem area as the biometrics and interview combination. Therefore the change we implemented was the same as described in *9.2.5 Classic & Biometric Modification 5 (Level 2 - section 4)*.

9.3.6 BIOMETRIC & METRIC MODIFICATION 6 (LEVEL 2 - SECTION 4)

As mentioned before biometrics showed no frowning and a drop in smiles towards the end of level 2. While section 4 was designed to be challenging, the emotional feedback did not show much evidence of this.

Metrics showed section 4 had the shortest completion time and the lowest amount of jumps of the whole level. Since the change described in the previous sub-chapter leaves less space until the finish, we wanted to use more vertical space in this section. Again, general ratings backed this up suggesting that level 2 was too short. By adding more jumps the amount of actions required to get all the coins increases and therefore also the perception of length and possibly fun.

Since this addressed the same issue as in interview and biometrics, we implemented the change from *9.2.6 Classic & Biometric Modification 6 (Level 2 - section 4)*.

9.4 CONCLUSION AND DISCUSSION OF PHASE 4

In this phase of our study we combined the data gathered from the testing sessions and looked for correlations that would justify changes to the game levels. In this phase we gained more clarity on the usefulness of each data set and how the different sets worked together.

From all three data sets, the ‘classic data’ was the most specific in guiding us to parts of the levels that needed changing. Since players tended to point out the aspects of the game which they enjoyed the most or the least, the data was very clear in suggesting where to take actions. Therefore it was easy for the designer to make changes based on this data set.

Metrics provided a massive amount of data. While this is generally useful, it made it hard to determine which information had meaning in regards to level design and therefore whether or not it should be followed as a basis for changes. Because some of the raw data output was hard to put back into context, it left a lot of interpretation to the designer. Additionally, we felt that dividing the level into more sections could have helped to pinpoint problematic areas.

Heatmaps however were extremely useful in suggesting localized changes and offered a high resolution of information, as each tile in the game correlated to a data point on the heatmap.

Biometrics provided interesting and sometimes surprising results. While the readings were informative, they were not very specific. This gives the designer a lot of freedom in terms of how to interpret the data but could also allow for misinterpretation to support existing assumptions. Formulating more focused questions could relieve this possibility.

From the standpoint of a designer the combination of classic and metric GUR data was the easiest to work with. It provided both the targeted comments from the players as well as clear data that provided more insight into the problem.

10. Phase 5: Testing Round 2 - Evaluation and Comparison of Level-Set Modifications based on GUR

This chapter details the necessary equipment, testing set up and environment, testing procedures and participant selection for round 2.

10.1 Testing Design and Environment for Round 2

For the second round of testing we tried to keep the testing environment consistent to the first round in terms of set up. Since in this round we did not need to collect the GUR methodology data, but only to get the results from the GEQ, we could be more flexible with the location of testing. While most tests still took place in the research lab, we also tested some people on locations more convenient to them. In these cases we made sure the participant was seated somewhere quiet and separate to avoid distraction. The necessary equipment for this round was:

- A laptop for the participant to play the game on (with charger)
- A headset
- Information sheet, consent form and debriefing sheet
- Internet connection (to upload GEQ results to server)

In this round of testing all participants had to wear a headset to keep the testing circumstances more consistent while being in different locations. The use of a headset limits external interference, therefore allowing the participant to focus more on the game.

The participant was seated somewhere quiet to play the game. The researcher was present at all times during the experiment; making sure the participant was not disturbed. While the researcher stayed close during the test, he or she was not positioned in a way that they could observe the progress of the player in the game.

10.2 Sampling: Participant Selection for Testing Round 2

For this round of testing we used convenience sampling (also described in chapter *7.2 Sampling: Participant Selection for Testing Round 1*).

To get comparable data we needed to get a roughly equal amount of people to play each level set. The size of the total participant group was 40 people, 22 of which were female. The age range for this group was between 15 and 27 years of age. Since all participants from round 1 within the higher age categories were either unable to finish the three levels or took a lot longer to complete the test we decided to not include them in this round of testing. We attempted to keep the average age across both groups similar in spite of this. The median age of the second testing group was 23 years.

The statistics for the individual groups were as followed:

- **Classic & Metric:** 14 participants of which 7 female with a median age of 22.5 years
- **Classic & Biometric:** 13 participants of which 8 female with a median age 22 years
- **Biometric & Metric:** 13 participants of which 7 female with a median age of 23 years

10.3 Testing Procedures for Round 2

The experiment started and ended with the same briefing and debriefing phases as in round 1. After informing the participant about the experiment and getting his or her consent, they were asked to play the game. Prior to the experiment the researcher chose which level set (Classic & Metric, Classic & Biometric, Biometric & Metric) the participant would play at random. After setting up the game for the participant the researcher took a position close to the computer, but not so close the participant would feel watched.

When the participant was done playing the researcher started up the GEQ, which the participant then filled out. Afterwards the participant was thanked for his or her cooperation and talked through the debriefing sheet, which they were allowed to take with them.

10.4 Conclusion and Discussion of Phase 5

Testing in the second round was shorter and less complicated as this time around there was no need for biometric sensors. This allowed us to test participants on location instead of only in the lab, making it easier to find participants.

This however also made the environment harder to control. While efforts were made to provide the participants with a quiet area in which they would not be disturbed during the test, it happened on occasion that interruptions were hard to avoid. Possibly the less formal testing environment also influenced how serious the participants regarded the experiment, as tests were sometimes shortly interrupted by the participant or on some occasions even aborted completely before the test concluded (these were not taken as valid data sets).

For the results of the study however we believe the distractions were not of such magnitude that they influenced the results to a significant extent.

11. Research Results

As all of the individual phases dealt with the respective results, this chapter focuses on the comparison of the methodology combinations by use of the GEQ scores.

While the research question of this study focuses primarily on the possibilities regarding the combination of GUR data, we were also interested in the impact of the varying modifications. As described in chapter *5.4 Game Experience Questionnaire*, the GEQ scores different aspects of a game experience, which depend on the GEQ modules that are used.

The following graph illustrates the results of the GEQ:

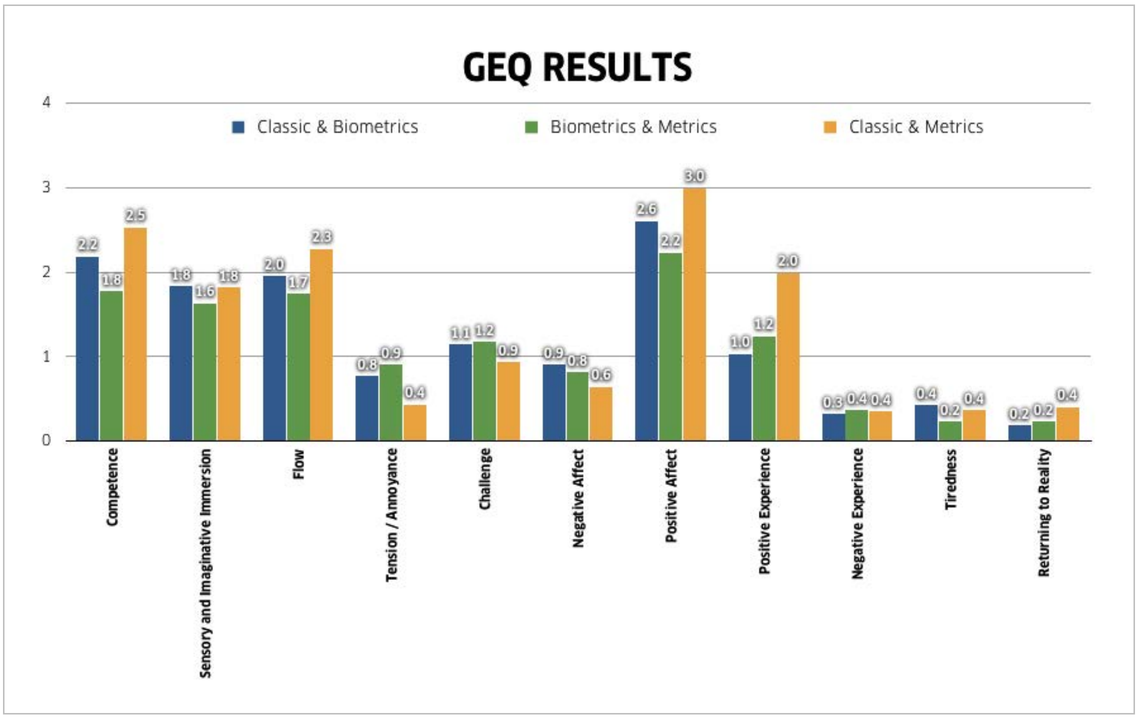


Fig. 49: Graph showing the GEQ scores of the individual methodology testing groups divided by the aspects that are scored in the GEQ

12. Research Limitations

While much effort went into reducing the amount of limitations in this research, it is also unreasonable to expect that they can be fully avoided. In the following list we describe which limitations of this study could have introduced deviations in the results:

Lead researcher as level designer and participant observer

The author of this document has acted as the lead researcher of this study, has performed changes on the initial level-set, acted as observer during test sessions and has modified the levels in accordance to conclusions during the evaluation phase. In doing so, it becomes a challenge to remain objective over the course of the research, especially during the evaluation of data. Prior experiences as game and level designer have given the researcher insights into common design practices, but 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. The focus group is one example, which has been used in the research to provide input from external parties when assessing the quality of the level design. Likewise, the researcher took care to stick to data-supported evidence when modifying the levels instead of acting on general possibilities of improving the level design.

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. It is inherently difficult to compare the quality of a design objectively. Consequently, the fact that a level designer has been part of this research meant that the changes in the levels were partly subjective.

It also should 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 (Mirza-babaei et al., 2013).

Test participants had to play using one hand only

Due to the involvement of biometric sensors during the first round of testing, namely sensors that were attached to one hand, participants had to play using the remaining hand. This meant that the controls of the game had to be changed so that all necessary buttons could be reached with one hand. In addition to this limitation, participants were also asked to not move the hand that had sensors attached to it. The point could be made that playing with both hands, as intended by the original design of the game, changes how difficult the game is perceived. This becomes a factor once the game is released to players that are free to choose how to play it and then experience the game differently.

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 different results. At the same time, it would have been interesting to see the individual influences of each methodology. Such considerations occurred during the course of the study but had to remain uninvestigated to keep scope of the study manageable.

GEQ might not be ideal for measuring actual quality of levels

Even though a lot of data was gathered through various means, we do not know whether changes in the level were considered qualitative improvements. The GEQ always evaluates the complete game experience, which, as it has been described before, is dependent on many factors outside of level design. Changes in level design might also simply be too subtle to significantly influence GEQ results.

13. RESEARCH CONCLUSION AND DISCUSSION

We began this study with the question of how combinations of GUR methodologies could be used to improve level design in a 2D platform game. In the process of investigating this question five phases were established which represent major milestones in this study. The following summary serves as a concise reminder of the conclusions we drew for each phase. For a more in-depth description, refer to the corresponding chapters in this document.

In the first phase we prepared the game for upcoming testing sessions. As such the individual GUR methodologies were not yet evaluated. A focus group conducted in this phase introduced several changes to the levels in order to ensure basic QA. Visual aids gave the group participants the possibility to clearly indicate locations in the levels. Having witnessed the benefits of such assistive visualization material, we decided to provide participants in testing sessions with similar material. In order to track metric and biometric data, logging functionality had to be added to the game. A time intensive process, which, in a real world scenario, should be part of the early development of game code.

The second phase of the research was a testing round with players, which was conducted to gather GUR data. Here we were able to get first hand experience with the usability of the individual methodologies. Interviews benefitted from the level sheets that were provided to players as visual references as the resulting data was easily correlated to specific locations in the level. In this phase, the collection of metric data took the least amount of time, as it fully took place in the background. In contrast, collecting biometric data took the most time. We also predicted that some time would need to be spent on removing unwanted measures, since participants often unwittingly commented their actions while playing. One of the most important goals for the future use of biometric GUR methodologies is therefore the simplification and automatization of the data collection.

Phase three was the evaluation and visualization of data. We argued that level design is a task that involves various variables that need to be brought into equilibrium by the designer. Processing the data in a way that suits the needs of designers is therefore highly important for the practical use of GUR data. Classic GUR data involved the most manual work, as interview and observation notes had to be transcribed and categorized. Metric data proved challenging due to the high amount of information that had to be processed and interpreted. Biometric data took the least amount of time to process in this phase and conveyed qualitative information that would be hard if not impossible to acquire with other methodologies.

The fourth phase of this study was the modification of levels in accordance to the evaluation of the gathered GUR data. In this phase, classic GUR data provided the most specific information. Both metric and biometric data provided interesting insights, on their own however we found it hard to formalize specific changes. Consequently, modifications that were based on metric and biometric GUR data took more liberties when it came to introducing changes in the level.

The fifth and last phase was a testing round in which the modified level-sets were played and subsequently scored by using the GEQ. This phase provided data that was used to compare how successful the respective methodology combinations were in terms of game experience.

Given the insight gained in the five phases of the research and the GEQ results of the individual methodologies, we feel that classic GUR methodologies are essential to the success of improving level design and should therefore always be involved. Especially in regards to level design we see great potential in devising assisting material that can be used in test sessions to guide the interview.

The combination of classic and metric methodologies puts both subjective and objective information into context. While classic data is of great use to uncover problems in a level, we found that metric data provides useful information regarding how to solve these problems. We feel that the biggest challenge for the use of metric data is the complexity 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.

As a final point of discussion, we would like to address the possibility of replicating the processes of this study in 3D 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 3D space. Heatmaps in 3D 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 3D games certainly raise the complexity compared to their use in 2D games, we believe that such challenges can be overcome.

We hope that future research will explore such possibilities!

14. Bibliography

Ambinder, M. (2009). GDC Vault - Valve's Approach to Playtesting: the Application of Empiricism. Presented at the GDC 2009, San Francisco.

Bryman, A. (2008). *Social Research Methods* (3rd ed.). Oxford University Press, USA.

Cacioppo, J. T., Tassinary, L. G., & Berntson, G. (2007). Handbook of Psychophysiology - John T. Cacioppo, Louis G. Tassinary, Gary Berntson - Google Books.

Drachen, A., & Canossa, A. (2009). Analyzing user behavior via gameplay metrics, 19–20.

El-Nasr, M. S., Desurvire, H., Nacke, L., Drachen, A., Calvi, L., Isbister, K., & Bernhaupt, R. (2012). Game user research. Presented at the CHI EA '12: Proceedings of the 2012 ACM annual conference extended abstracts on Human Factors in Computing Systems Extended Abstracts, ACM. doi:10.1145/2212776.2212694

Gow, J., Cairns, P., Colton, S., Miller, P., & Baumgarten, R. (2010). Capturing Player Experience with Post-Game Commentaries.

Gualeni, S., & Double Jungle SaS. (2011). Gua-Le-Ni; or, The Horrendous Parade [iPad].

Gualeni, S., Janssen, D., & Calvi, L. (2012). How psychophysiology can aid the design process of casual games: a tale of stress, facial muscles, and paper beasts. Presented at the FDG '12: Proceedings of the International Conference on the Foundations of Digital Games, ACM Request Permissions. doi:10.1145/2282338.2282369

IJsselsteijn, W., de Kort, Y., Poels, K., Jurgelionis, A., & Bellotti, F. (2007). Characterising and measuring user experiences in digital games. In *International Conference on Advances in Computer Entertainment Technology* (Vol. 2, p. 27).

IJsselsteijn, W., van den Hoogen, W., Klimmt, C., de Kort, Y., Lindley, C., Mathiak, K., ... & Vorderer, P. (2008). Measuring the experience of digital game enjoyment. In *Proceedings of Measuring Behavior* (pp. 88-89).

Kivikangas, J. M., Ekman, I., Chanel, G., Järvelä, S., Salminen, M., Cowley, B., Henttonen, P., et al. (2010). Review on psychophysiological methods in game research. *Proc. of 1st Nordic DiGRA*.

Lameman, B. A., El-Nasr, M. S., Drachen, A., Foster, W., Moura, D., & Aghabeigi, B. (2010). User studies: a strategy towards a successful industry-academic relationship. Presented at the Futureplay '10: Proceedings of the International Academic Conference on the Future of Game Design and Technology, ACM Request Permissions. doi:10.1145/1920778.1920798

Larsen, J. T., Norris, C. J., & Cacioppo, J. T. (2003). Effects of positive and negative affect on electromyographic activity over zygomaticus major and corrugator supercilii.

Psychophysiology, *40*(5), 776–785. doi:10.1111/1469-8986.00078

Leone, M. (2012, October 24). Data entry, risk management and tacos: Inside Halo 4's playtest labs | Polygon. *polygon.com*. Retrieved February 14, 2013, from http://www.polygon.com/2012/10/24/3538296/data-entry-risk-management-and-tacos-inside-halo-4s-playtest-labs

Lewis-Evans, B. (2012a, April 24). Gamasutra - Features - Finding Out What They Think: A Rough Primer To User Research, Part 1. *gamasutra.com*. Retrieved September 26, 2012, from http://www.gamasutra.com/view/feature/169069/finding_out_what_they_think_a_.php?print=1

Lewis-Evans, B. (2012b, May 15). Gamasutra - Features - Finding Out What They Think: A Rough Primer To User Research, Part 2. *gamasutra.com*. Retrieved September 26, 2012, from http://www.gamasutra.com/view/feature/170332/finding_out_what_they_think_a_.php?print=1

Mandryk, R. L., & Atkins, M. S. (2007). A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies. *International Journal of Human-Computer Studies*, *65*(4), 329–347. doi:10.1016/j.ijhcs.2006.11.011

Mandryk, R. L., Atkins, M. S., & Inkpen, K. M. (2006). A continuous and objective evaluation of emotional experience with interactive play environments. Presented at the CHI '06: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM Request Permissions. doi:10.1145/1124772.1124926

Medlock, M. C., Wixon, D., & Terrano, M. (2002). Using the RITE method to improve products: A definition and a case study. *Usability ...*.

Mirza-babaei, P., & McAllister, G. (2011). Biometric Storyboards: visualising meaningful gameplay events. Presented at the BBI Workshop CHI 2011.

Mirza-babaei, P., Long, S., Foley, E., & McAllister, G. (2011). Understanding the Contribution of Biometrics to Games User Research.

Mirza-babaei, P., Nacke, L. E., Gregory, J., Collins, N., & Fitzpatrick, G. (2013). How Does It Play Better? Exploring User Testing and Biometric Storyboards in Games User Research (pp. 1–10). Presented at the CHI 2013.

Nacke, L. E. (2009). Affective Ludology :Scientific Measurement of User Experience in Interactive Entertainment.

Nacke, L. E. (2011). Directions in Physiological Game Evaluation and Interaction. *In CHI 2011 BBI Workshop Proceedings*.

Nacke, L. E., Drachen, A., Kuikkaniemi, K., Niesenhaus, J., Korhonen, H. J., Hoogen, V. D. W., Poels, K., et al. (2009). Playability and player experience research.

Nintendo. (1985). Super Mario Bros. [NES].

Pagulayan, R. J., Keeker, K., Wixon, D., Romero, R. L., & Fuller, T. (2003). User-centered design in games. *The human-computer interaction handbook: fundamentals, evolving technologies and emerging applications*, 883–906.

Ravaja, N., Saari, T., Salminen, M., Laarni, J., & Kallinen, K. (2006). Phasic Emotional Reactions to Video Game Events: A Psychophysiological Investigation. *Media Psychology*, *8*(4), 343–367. doi:10.1207/s1532785xmep0804_2

Singleton, R., & Straits, B. C. (2005). *Approaches to Social Research* (4th ed.). Oxford University Press, USA.

SuperTux Development Community. (n.d.). Milestone 2 Design Document/Styleguide - SuperTux. *supertux.lethargik.org*. Retrieved March 8, 2013, from http://supertux.lethargik.org/wiki/Milestone_2_Design_Document/Styleguide

Tychsen, A. (2008). Crafting user experience via game metrics analysis, 20–22.

Watson, D., & Tellegen, A. (1985). Toward a consensual structure of mood. *Psychological Bulletin*, *98*(2), 219–235. doi:10.1037/0033-2909.98.2.219

Wesley, D., & Barczak, G. (2010). Innovation and Marketing in the Video Game Industry: Avoiding the Performance Trap.

15. APPENDICES

A. LIST OF MODIFICATIONS ON THE INITIAL LEVEL-SET

- **Cutting off or prolonging each of the levels to 300 tiles in width (9600 pixel)**
  Since the perceived length of a level is possibly connected to the difficulty of obstacles and enemies in the game, all levels shared the same physical length to investigate the correlation between difficulty and perceived length.
- **Unifying the amount of collectable coins to 100 coins per level**
  This modification brought the level design in *SuperTux* closer to early titles in the *Super Mario* series, which also featured 100 coins in each level. Choosing a consistent number might also help players with inferring at any time in the level how many coins can still be picked up.
- **Unifying amount and nature of pickup blocks**
  By unifying pickup blocks the difficulty of a level could be influenced with fewer parameters, such as amount of enemies or amount and width of gaps.
- **Ensuring a sequential increase in amount of enemies and gaps**
  This modification ensured that each level became more difficult in terms of obstacles that the player needed to face.
- **Ensuring a sequential increase in enemy difficulty and amount of groups**
  Aside of quantifying the amount of enemies, we also structured the difficulty of each individual enemy as different AI patterns result in varying degrees of difficulty. The same is true for enemies that appear in groups.
  To structure the difficulty of individual enemy types, we grouped enemies in three difficulty ranges:
  - Type I enemies are defeated by jumping on top of them and remain either stationary or only exhibit simple movement patterns like walking.
  - Type II enemies are only incapacitated by jumping on top of them or can exhibit complex movement patterns.
  - Type III enemies can only be eliminated with fireballs or by colliding with them during star mode.
- **Providing at least one hidden secret area in every level**
  In the study we wanted to give players the opportunity to discover a hidden area in every level. Hidden secrets can motivate explorative playing styles and increases the range of activities that players can engage in during their time in the game.

B. Focus Group Guide Sheet

- Introduce Focus Group mediator
- Introduce Researcher (Team): Marcello, Marta & Dirk Janssen
- Introduce research topic / goals - **Level Design and QA methodologies**
- You have been invited here today to take part in a focus group. A focus group is –

“Focus group discussions are a method of qualitative research that involves an open form of discussing a specific set of issues with a pre-determined group of people.”

- You have been invited in your capacity as game designers. We would like to hear both your general consensus on the topics brought up, as well as the variety your different individual perspectives
- Explain that it will be recorded : audio / video
- Explain how long it is expected to take: 1 hour (Is there anybody here that cannot be present for the whole hour?)
- Explain what happens to the data - Given your consent, it will be used as teaching material and for further research, will not be made public without your further consent; sign release forms?
- Let participants introduce themselves: for the sake of protocol please briefly introduce your name and your game design specialty
- Instructions to the participants

  Open, informal discussion about the levels you just played

  Mediator is just here to moderate, and the researcher may comment, but for the most part it’s a discussion between you

  Try to not speak over each other or interrupt each other.

  Try to ignore your own gaming literacy and your personal playing style. Take into account that the target group of the game is not hardcore players.

Topic guide
Quality Assurance: How well do the four presented levels meet industry standards in terms of quality and playability?

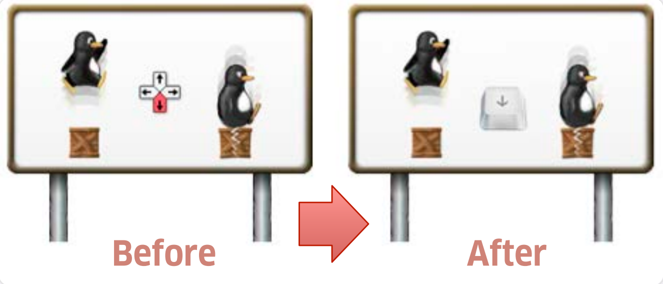
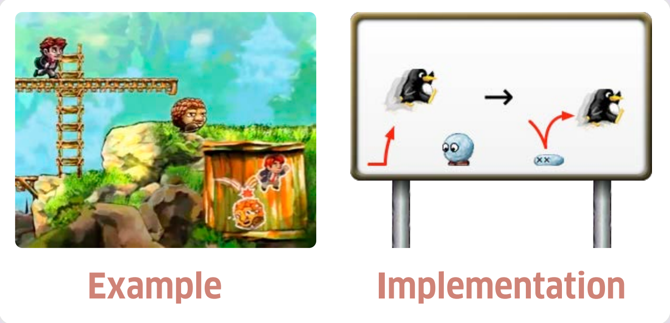
The levels need to follow a clear progression, the tools to complete them are provided in the tutorial, a minimum amount of game literacy is needed to complete all four levels.

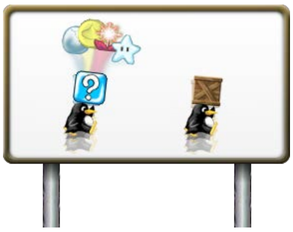
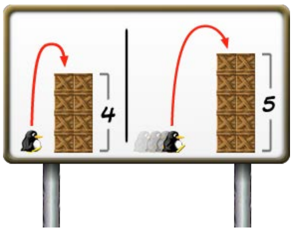
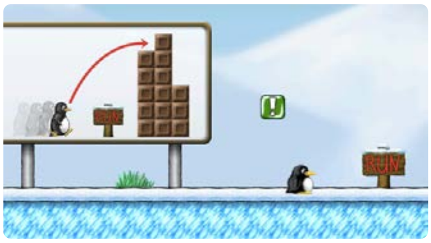
Playability :
Controls explained: To what extent did the demo level prepare the player to handle the following three levels?
Were the gameplay elements well introduced?
Duration: What did you think of the duration of the levels?

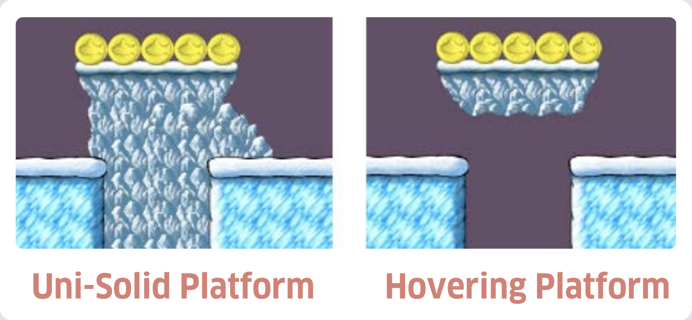
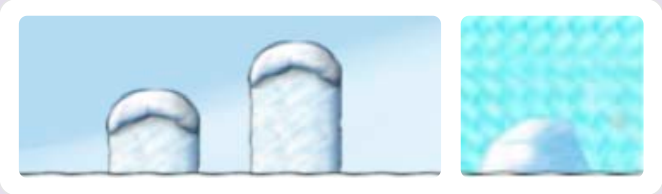
Quality:
Did you feel that the game elements were arranged purposefully?
Visual consistency: Did you find any visual inconsistencies that affected gameplay or level clarity?
Rhythm: What can you say regarding the rhythm of the levels? (in terms of obstacles and actions)
How did you find the Ramp in difficulty / challenge in between the 3 post-tutorial levels?
Target group: How suitable would you say the levels are for the demographic of players with a casual approach to gaming?
Player enjoyment: How can the enjoyment of the player be increased? What would you change to improve player enjoyment, given the target audience?

/Were enemies in the demo level well introduced?
/How challenging were enemies in each level?

C. Focus Group Modification Table

Time code(s)	Modifications
3:00, 31:32, 45:22	Removal of the explanation sign for the ability to perform a backflip in the tutorial level. It has been reasoned that this action creates confusion and increases complexity too quickly without a real benefit. As consequence, this also leads to the removal of coins that could previously only be reached by a backflip
4:30	Change in the graphical representation of the 'buttjump' explanation sign. The sign uses a D-Pad representation for buttons that need to be pressed while the typical player would play this game on a computer keyboard. 
6:00	Addition of an explanation sign that illustrates that the player avatar can eliminate enemies by jumping on top of them. An example for portraying such a mechanic can be found in the game *Braid* – developed by Number None, Inc. and released in 2008 on Xbox Live. The explanation sign is added to the tutorial level. 
6:00, 7:45, 8:25	The info box blocks that the game provides to relay written information to the player have been found to be distracting by the focus group. Instead they suggest replacing all info boxes with explanation signs. As a consequence the following signs are added to the tutorial level:

	Check Point Checkpoint Explanation  Bonus Block Explanation  Run Jump Explanation
6:30	Removal of crate blocks that can yield coins. Focus group participants have been confused by their workings and suggest to only use regular blocks that are destroyed by hitting them from below in big Tux form.
7:00, 14:13	Change in enemy placement to emphasize the difference between 'Snowball' and 'Mrs. Snowball' enemy types. 'Snowball' enemies keep moving into gaps while 'Mrs. Snowball' enemies turn around to not fall into a gap.
9:14, 36:30	Removal of the 'run' sign in the background of level 1. The sign was previously used next to the 'run jump explanation' sign (refer to time-code 6:00). The focus group participants suggested that the sign should be taken out to avoid redundant information. 
21:15, 22:00	Participants in the focus group mentioned that they found 'coin blocks' in close proximity to 'star pickup blocks' a hindrance to the feeling of flow in the game. For this reason, coin blocks that are close to 'star pickup blocks' have been replaced by floating coins.
23:00, 51:00	Addition of 'Tux upgrade blocks' in close proximity to the checkpoint of each level.
25:10, 51:50	More variations in height have been added in response to focus group feedback. Additionally, each level has been modified to include a ramp shortly before the end of a level in an effort to let players anticipate the level end.
25:40	Removal of decorative level elements that indicate the possibility of using them as platforms for the player avatar while not actually featuring the possibility to do so.
26:50	All checkpoint bells are slightly repositioned to make it easier for players to hit them. Previously, it was possible to miss a checkpoint bell simply due to the short height of the non-upgraded Tux avatar.
28:54,	The focus group identified a sharp spike in difficulty for the last section of level 3. The section has been modified in an effort to reduce the overall

51:00	difficulty at this point in the level.
32:15	Enemies in the last section of level 2 have been moved further to the right of the level to make more space. Participants of the focus group argued that their initial position created a ‘chokepoint’ – a location in a level that features a significantly increased amount of enemies or obstacles.
34:10	Redesign of ‘upgrade block’ locations to ensure that upgrades remain obtainable by players, regardless of the direction in which they move upon release from an ‘upgrade block’.
39:05	The focus group motivated the use of a visual element in close proximity to secret areas. Fish tiles that were previously used for decorative purposes only are now only used to visually hint at the presence of a secret area.
40:45	Platforms in the tutorial level and level 1 have been changed from a ‘uni-solid’ appearance to a hovering appearance. Participants of the focus group argued that players might consider ‘uni-solid’ platforms as elements that block the player avatar. They instead advise to use hovering platforms, which fulfill the same function in terms of level design.  Uni-Solid Platform Hovering Platform
52:30	Replacement of stone blocks with wooden crate blocks in the tutorial level since stone blocks never occurred again throughout the rest of the level-set.
54:28	Removal of all Tux doll blocks. These dolls can be picked up by the player to add 100 coins to the coin count. However, the pickup appears fast and needs to be picked up within a short amount of time before it disappears. Participants of the focus group have stated that they would prefer to not have this pickup at all in the game. All Tux doll blocks are replaced by coins.
58:20	Addition of a section in level 2 that features multiple icicles in a row in response to focus group feedback.
General Observation	The focus group seemed confused by graphical assets in the background of a level. As consequence these assets are removed from the levels in order to keep the level design clear. 

D. Log Action and Arguments Table

Note: '#' denotes a numeric value. Furthermore, many arguments are expressed as values on x- and y- coordinates. The point of origin of these coordinates is on the top left corner of the level. The x-coordinate increases horizontally from left to right while the y-coordinate increases vertically from top to bottom.

Log Action and Arguments	Description

LEVEL
state = start/restart/endsequence/finished
level = "Levelname"
pause = 0/1
	Log action concerning level wide events.

UPDATETUX
x = # (position on 'x' coordinate)
y = # (position on 'y' coordinate)
dx = # (horizontal velocity)
dy = # (vertical velocity)
dir = 1/2 (facing left/facing right)
duck = 0/1
backflip = 0/1
falling = 0/1
jumping = 0/1
buttjump = 0/1
grab = 0/1
climbing = 0/1
numkeys = # (keys pressed since last update)
idleframes = # (unchanged frames since last update)
totalframes = # (frames since last update)
elapsed = # (time since last update in seconds)
	This action logs events that are connected to the player avatar every **0.25 seconds**. Most arguments of this log action are state switches.

BADGUYPOS
name = path/to/enemyname.sprite
x = # (position on 'x' coordinate)
y = # (position on 'y' coordinate)
distance = # (pixel distance from Tux avatar)
	This action is logged every **0.25 seconds** if an enemy is visible on screen. Triggers an action whenever an enemy is visible on the screen. The enemy can be identified by the sprite name.

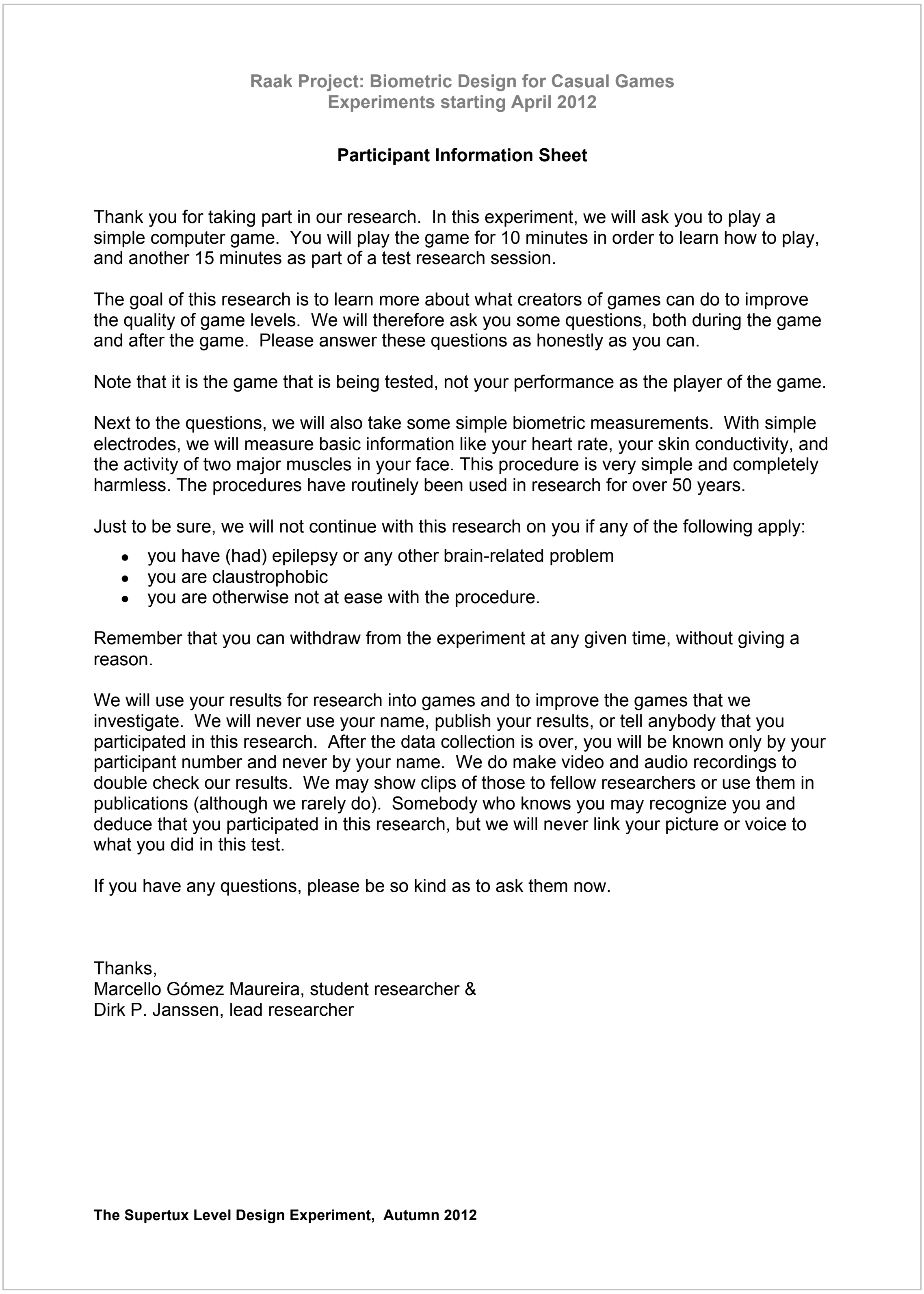
TRYKILLPLAYER enemy = path/to/enemyname.sprite	Triggers if the player comes in contact with an enemy and identifies the enemy based on the sprite name.

KILL
method = squish/fall
reason = fall/ignite
enemy = path/to/enemyname.sprite
	A 'kill' action is logged both for when the player defeats an enemy, as well as when the player is defeated.

CHECKPOINT
state = touched
sector = "levelsector"
pos.x = # (position on 'x' coordinate)
	Logs when the player touches a checkpoint.

pos.y = # (position on ‘y’ coordinate)	
PLAYERSTATUS invincible = start/stop	Marks the start and end of the invincibility period granted by the ‘star’ pickup.
PLAYERADJUST bonus = 0/1/2 (small/big/fire)	Triggers every time the player changes from one player state to another. The argument ‘bonus’ can be misleading, as a player might change from state 1 (big Tux) to state 0 (small Tux) due to enemy contact.
BONUSBLOCK content = grow/coin/fire/star	This action is logged for each ‘question-mark’ block that the player hits. The content of the block is logged as argument.
BROKEN block = brick	Triggers every time a ‘crate’ block is destroyed.
SECRETAREA totaldiscovered = #	Logs the discovery of a secret area. The location of the area can be inferred by the ‘Update Tux’ log action.
JUMP	This action is logged every time the player performs a jump. The time code of the log is used in combination with a preceding ‘Update Tux’ log action to infer position and time of the jump.
COIN collect = 1	Triggers a log line every time the player collects a coin. Even though an argument is passed on, it always returns the value ‘1’.
MRICEBLOCK state = normal/flat	Logs an action every time the player incapacitates a ‘Mr. Iceblock’ enemy. ‘Flat’ indicates that the enemy can be kicked by the player, which causes the enemy to slide away. In this form the enemy acts as a projectile that can harm other enemies or the player on contact. If ‘Mr. Iceblock’ has not fallen off the play screen while sliding, it returns to a ‘normal’ state after some time.

E. Participant Information Sheet



Raak Project: Biometric Design for Casual Games
Experiments starting April 2012

Participant Information Sheet

Thank you for taking part in our research. In this experiment, we will ask you to play a simple computer game. You will play the game for 10 minutes in order to learn how to play, and another 15 minutes as part of a test research session.

The goal of this research is to learn more about what creators of games can do to improve the quality of game levels. We will therefore ask you some questions, both during the game and after the game. Please answer these questions as honestly as you can.

Note that it is the game that is being tested, not your performance as the player of the game.

Next to the questions, we will also take some simple biometric measurements. With simple electrodes, we will measure basic information like your heart rate, your skin conductivity, and the activity of two major muscles in your face. This procedure is very simple and completely harmless. The procedures have routinely been used in research for over 50 years.

Just to be sure, we will not continue with this research on you if any of the following apply:

- you have (had) epilepsy or any other brain-related problem
- you are claustrophobic
- you are otherwise not at ease with the procedure.

Remember that you can withdraw from the experiment at any given time, without giving a reason.

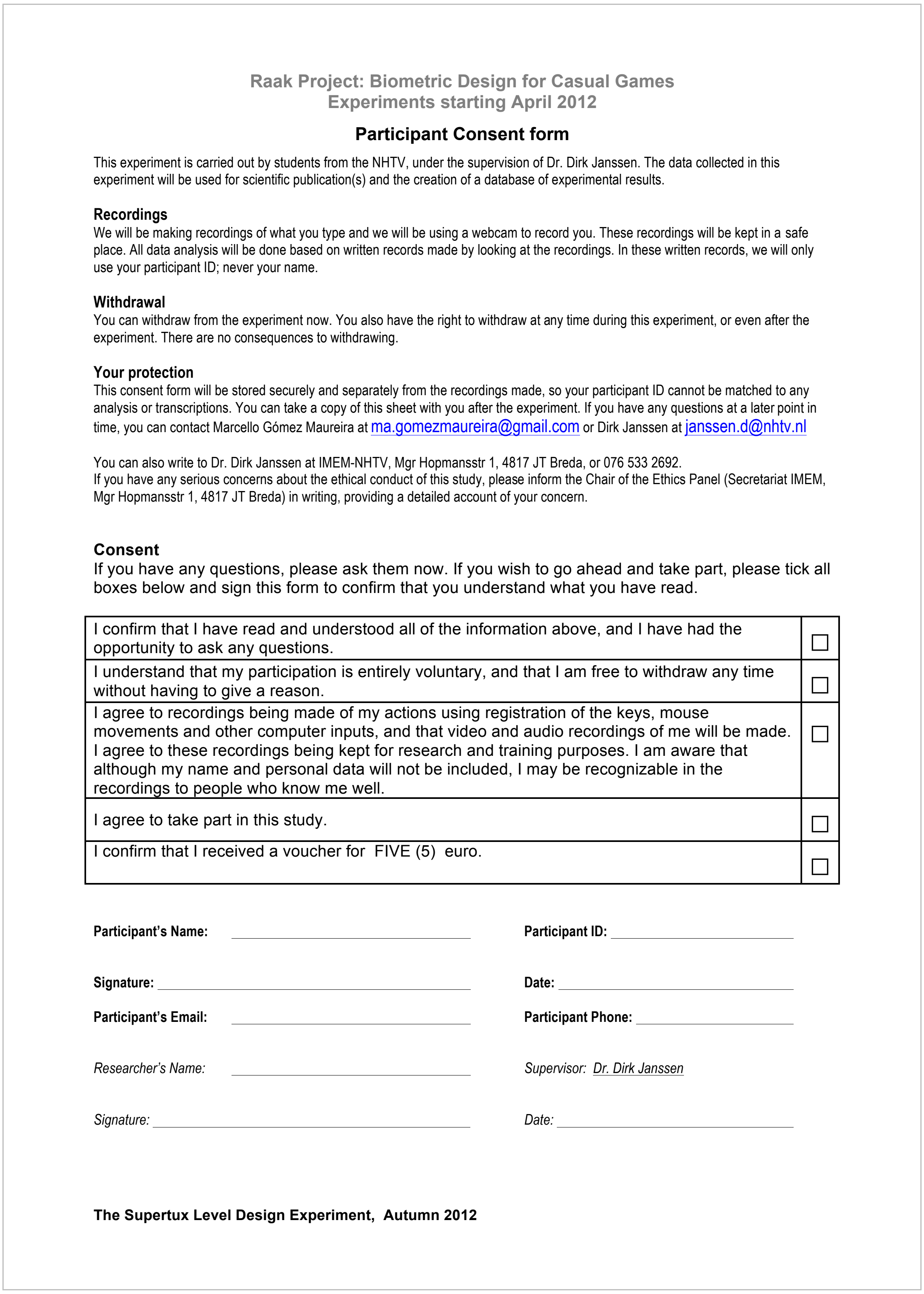
We will use your results for research into games and to improve the games that we investigate. We will never use your name, publish your results, or tell anybody that you participated in this research. After the data collection is over, you will be known only by your participant number and never by your name. We do make video and audio recordings to double check our results. We may show clips of those to fellow researchers or use them in publications (although we rarely do). Somebody who knows you may recognize you and deduce that you participated in this research, but we will never link your picture or voice to what you did in this test.

If you have any questions, please be so kind as to ask them now.

Thanks,
Marcello Gómez Maureira, student researcher &
Dirk P. Janssen, lead researcher

The Supertux Level Design Experiment, Autumn 2012

F. PARTICIPANT CONSENT FORM



Raak Project: Biometric Design for Casual Games
Experiments starting April 2012

Participant Consent form

This experiment is carried out by students from the NHTV, under the supervision of Dr. Dirk Janssen. The data collected in this experiment will be used for scientific publication(s) and the creation of a database of experimental results.

Recordings
We will be making recordings of what you type and we will be using a webcam to record you. These recordings will be kept in a safe place. All data analysis will be done based on written records made by looking at the recordings. In these written records, we will only use your participant ID; never your name.

Withdrawal
You can withdraw from the experiment now. You also have the right to withdraw at any time during this experiment, or even after the experiment. There are no consequences to withdrawing.

Your protection
This consent form will be stored securely and separately from the recordings made, so your participant ID cannot be matched to any analysis or transcriptions. You can take a copy of this sheet with you after the experiment. If you have any questions at a later point in time, you can contact Marcello Gómez Maureira at ma.gomezmaureira@gmail.com or Dirk Janssen at janssen.d@nhtv.nl

You can also write to Dr. Dirk Janssen at IMEM-NHTV, Mgr Hopmansstr 1, 4817 JT Breda, or 076 533 2692.
If you have any serious concerns about the ethical conduct of this study, please inform the Chair of the Ethics Panel (Secretariat IMEM, Mgr Hopmansstr 1, 4817 JT Breda) in writing, providing a detailed account of your concern.

Consent
If you have any questions, please ask them now. If you wish to go ahead and take part, please tick all boxes below and sign this form to confirm that you understand what you have read.

I confirm that I have read and understood all of the information above, and I have had the opportunity to ask any questions.	□
I understand that my participation is entirely voluntary, and that I am free to withdraw any time without having to give a reason.	□
I agree to recordings being made of my actions using registration of the keys, mouse movements and other computer inputs, and that video and audio recordings of me will be made. I agree to these recordings being kept for research and training purposes. I am aware that although my name and personal data will not be included, I may be recognizable in the recordings to people who know me well.	□
I agree to take part in this study.	□
I confirm that I received a voucher for FIVE (5) euro.	□

Participant's Name: ______________________ **Participant ID:** ______________________

Signature: ______________________ **Date:** ______________________

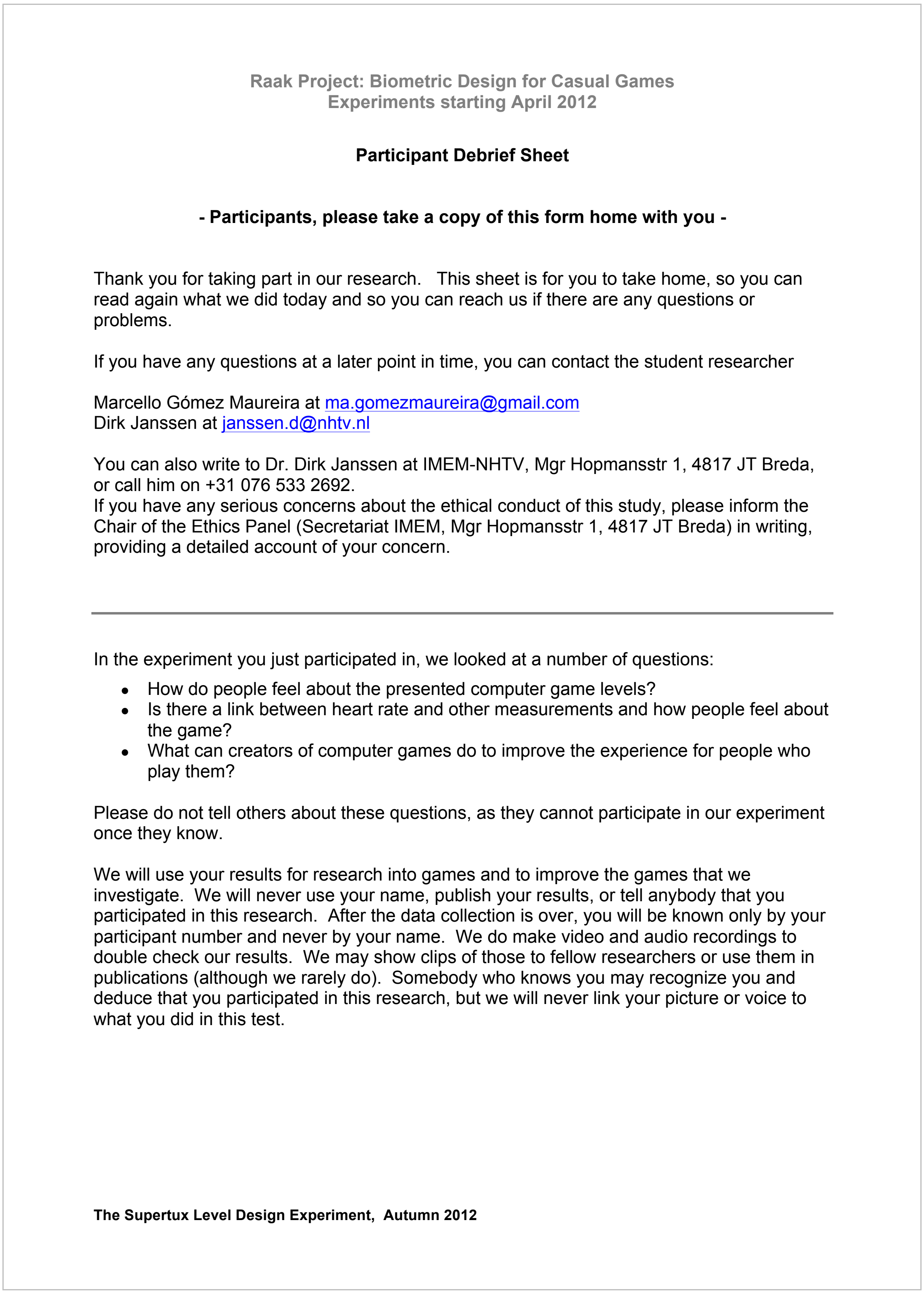
Participant's Email: ______________________ **Participant Phone:** ______________________

Researcher's Name: ______________________ *Supervisor: Dr. Dirk Janssen*

Signature: ______________________ *Date:* ______________________

The Supertux Level Design Experiment, Autumn 2012

G. Participant Debrief Sheet



Raak Project: Biometric Design for Casual Games
Experiments starting April 2012

Participant Debrief Sheet

- Participants, please take a copy of this form home with you -

Thank you for taking part in our research. This sheet is for you to take home, so you can read again what we did today and so you can reach us if there are any questions or problems.

If you have any questions at a later point in time, you can contact the student researcher

Marcello Gómez Maureira at ma.gomezmaureira@gmail.com
Dirk Janssen at janssen.d@nhtv.nl

You can also write to Dr. Dirk Janssen at IMEM-NHTV, Mgr Hopmansstr 1, 4817 JT Breda, or call him on +31 076 533 2692.
If you have any serious concerns about the ethical conduct of this study, please inform the Chair of the Ethics Panel (Secretariat IMEM, Mgr Hopmansstr 1, 4817 JT Breda) in writing, providing a detailed account of your concern.

In the experiment you just participated in, we looked at a number of questions:

- How do people feel about the presented computer game levels?
- Is there a link between heart rate and other measurements and how people feel about the game?
- What can creators of computer games do to improve the experience for people who play them?

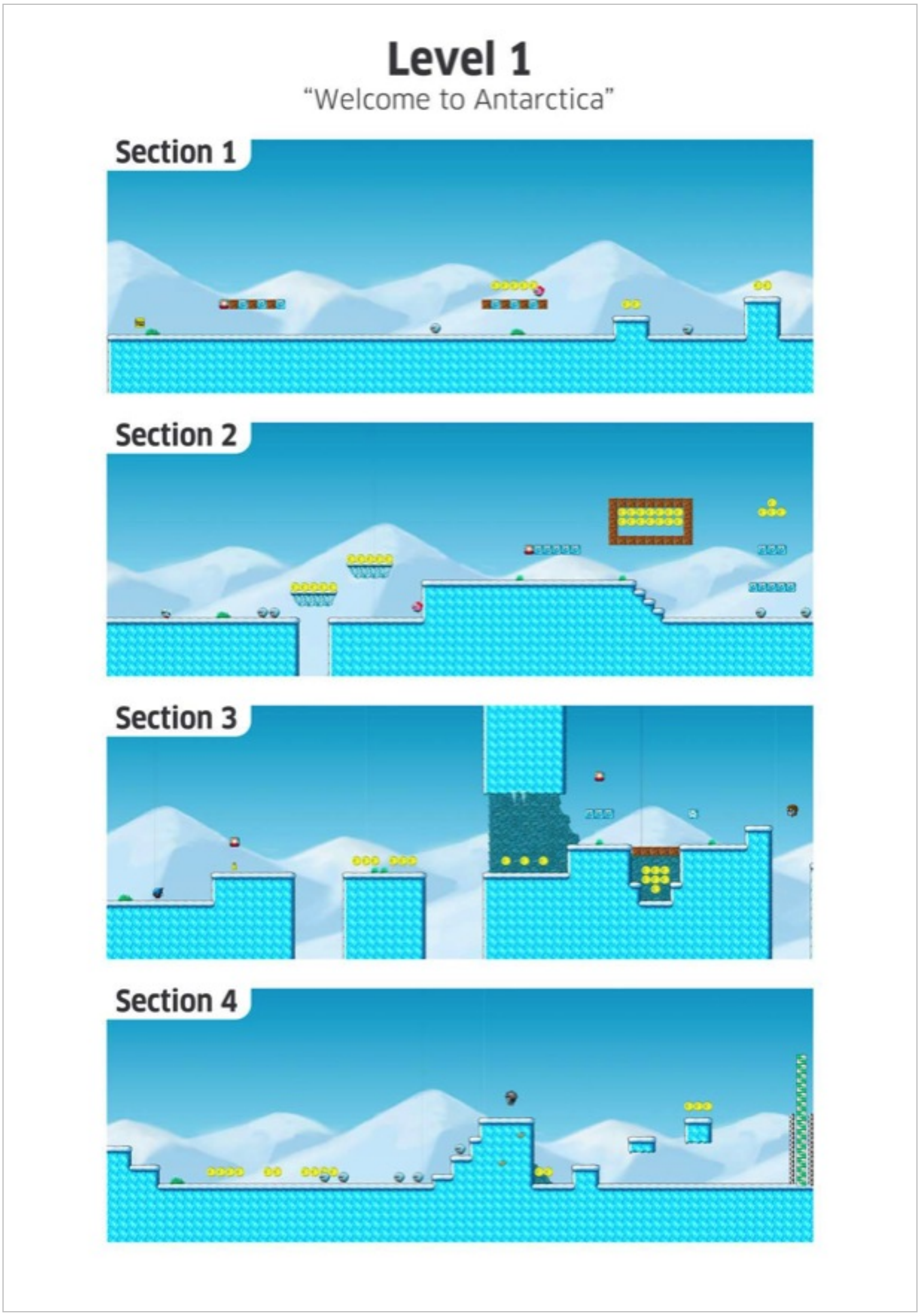
Please do not tell others about these questions, as they cannot participate in our experiment once they know.

We will use your results for research into games and to improve the games that we investigate. We will never use your name, publish your results, or tell anybody that you participated in this research. After the data collection is over, you will be known only by your participant number and never by your name. We do make video and audio recordings to double check our results. We may show clips of those to fellow researchers or use them in publications (although we rarely do). Somebody who knows you may recognize you and deduce that you participated in this research, but we will never link your picture or voice to what you did in this test.

The Supertux Level Design Experiment, Autumn 2012

H. Level Sheets for Player Interviews in Testing Round 1

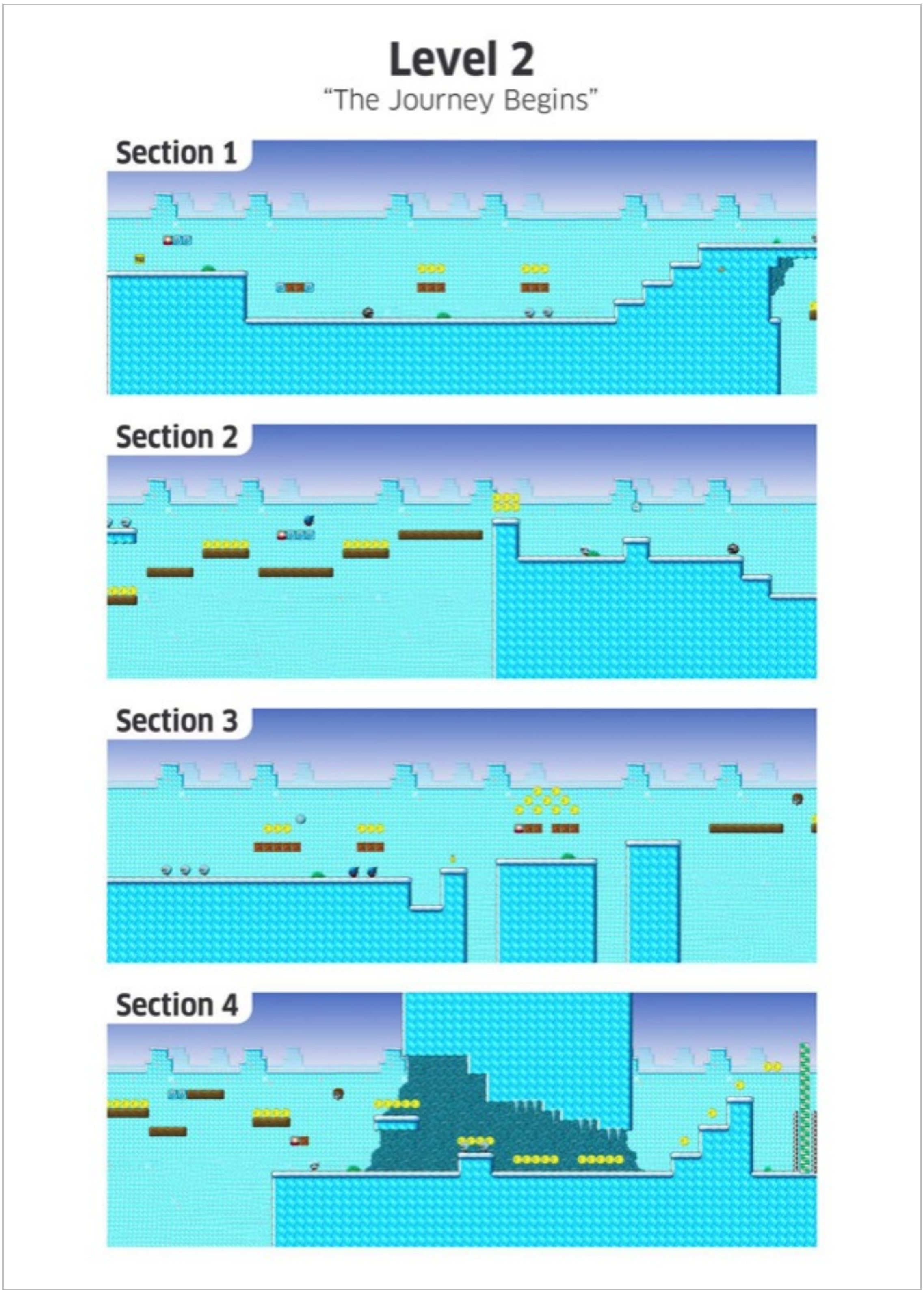
H.1 Level 1



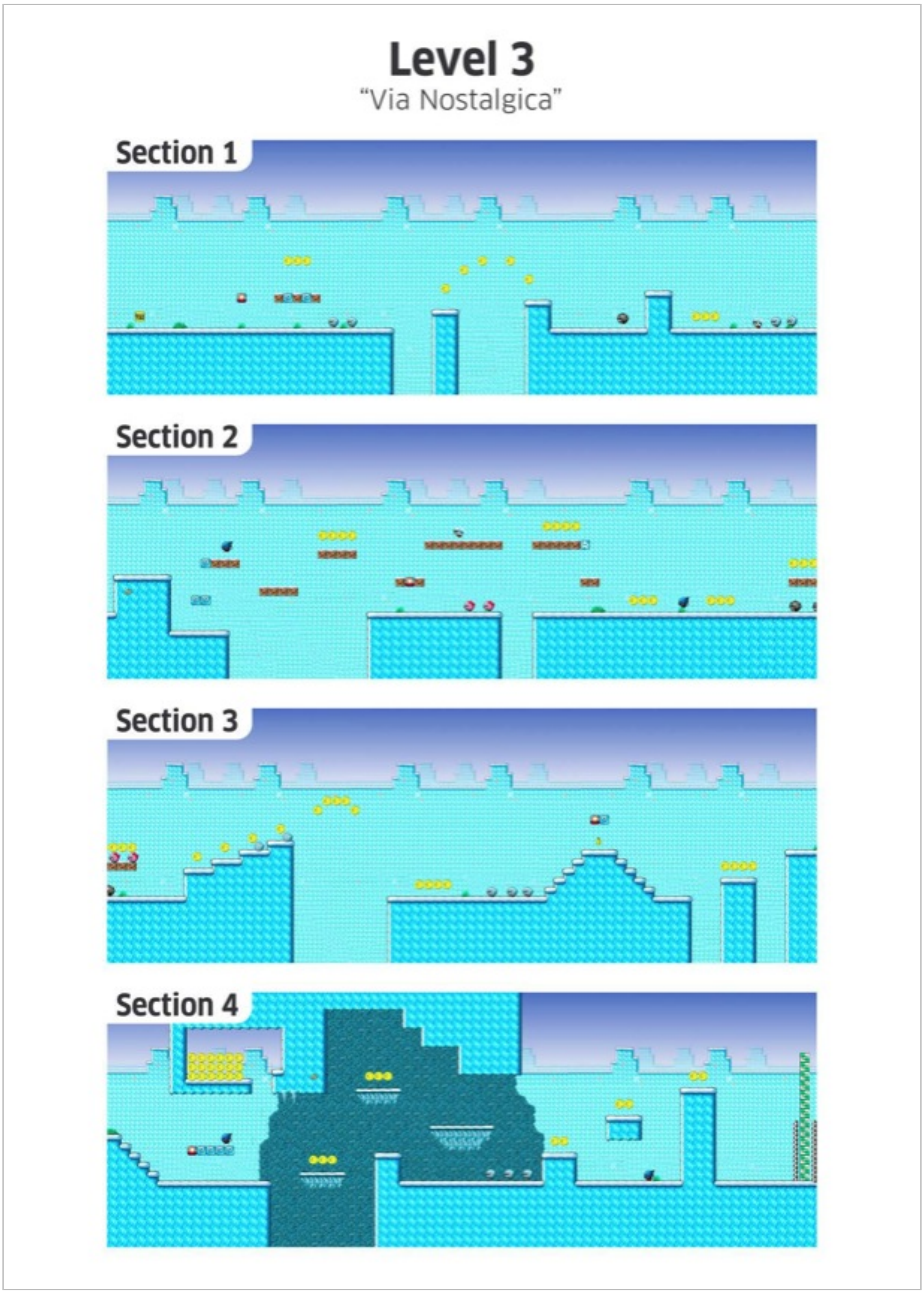
Level 1

“Welcome to Antarctica”

H.2 Level 2

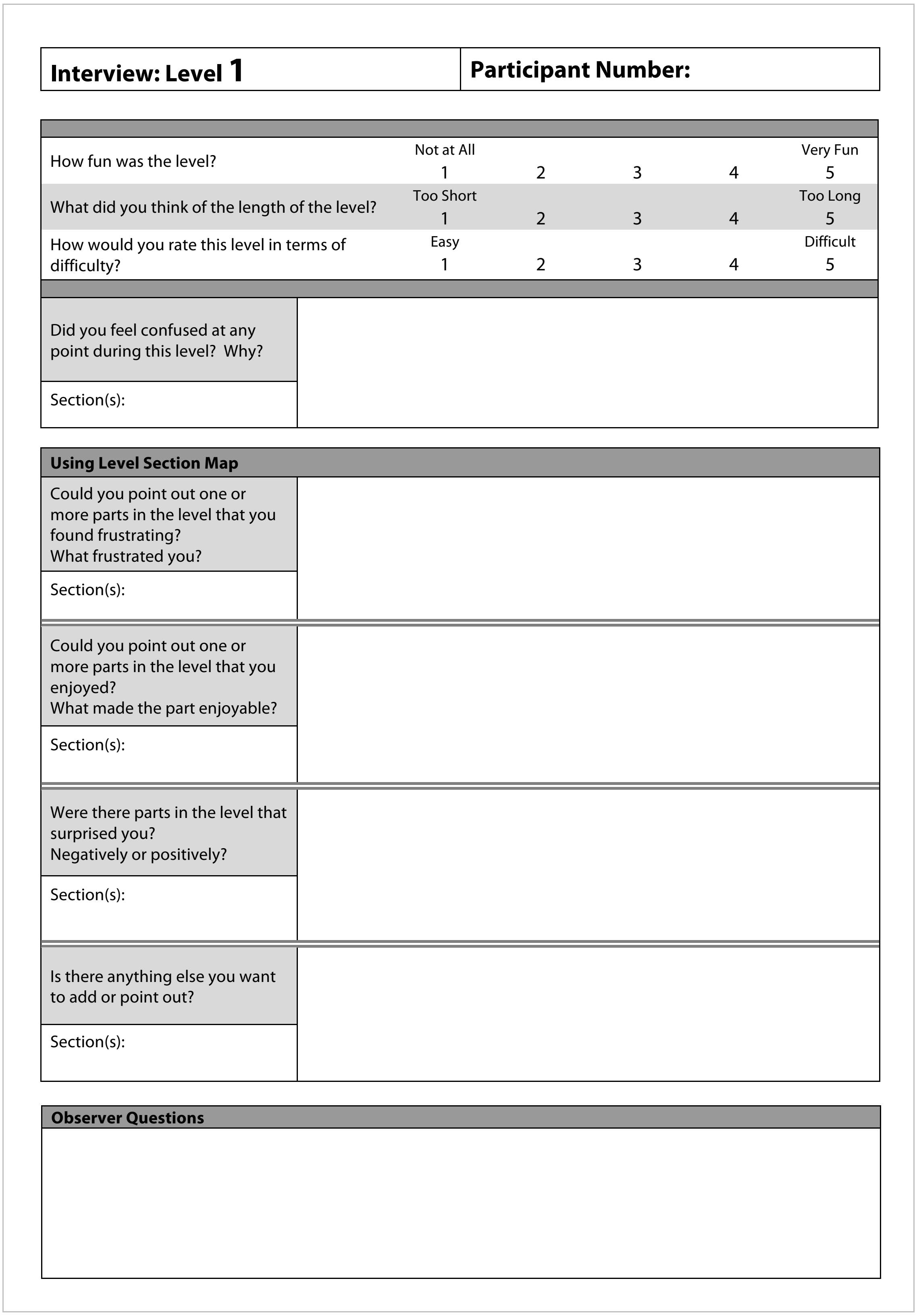

Level 2
"The Journey Begins"
Section 1
Section 2
Section 3
Section 4

H.3 Level 3


Level 3
"Via Nostalgica"
Section 1
Section 2
Section 3
Section 4

I. Player Interview Form Used in Testing Round 1

<u>Note:</u> The same form was used for three individual levels of a level-set.



Interview: Level 1	**Participant Number:**

How fun was the level?	Not at All 1	2	3	4	Very Fun 5
What did you think of the length of the level?	Too Short 1	2	3	4	Too Long 5
How would you rate this level in terms of difficulty?	Easy 1	2	3	4	Difficult 5

Did you feel confused at any point during this level? Why?	
Section(s):	

Using Level Section Map	
Could you point out one or more parts in the level that you found frustrating? What frustrated you?	
Section(s):	
Could you point out one or more parts in the level that you enjoyed? What made the part enjoyable?	
Section(s):	
Were there parts in the level that surprised you? Negatively or positively?	
Section(s):	
Is there anything else you want to add or point out?	
Section(s):	

Observer Questions

J. List of Derived Measures from Metric GUR Data

Note: Unless stated otherwise, all measures can be tracked per level and per level section.

J.1 Total Counts

- **Enemies Killed**
  A count of all enemies that were killed, regardless of how they were disposed of.
- **Modeshifts**
  'Modeshifts' are events in the game that cause the player avatar 'Tux' to change from one player state to another. States that are considered for modeshifts are 'small', 'grow', 'fire' and 'death'. For the amount of total modeshifts we considered it irrelevant whether the change in state was positive for the player (also referred to as 'modeshift up') or negative (also referred to as 'modeshift down')

J.2 Timespan between measured instances in seconds

Note: By looking at the timespan between two instances as opposed to just counts, we can decrease the influence of the amount of time that players spent in a level. Otherwise, we would have to ensure that players take about the same amount of time to finish.

- **Direction Change Instances**
  Measures the mean time between all instances of changes in direction, referring to the alignment of the player avatar, which faces the direction of the last pressed directional button.
- **Continuous Movement Instances**
  A measure of the mean time between instances that are defined as continuous movements. This measurement is based on another measure that counts all instances in the level at which the player avatar has remained in movement for longer than 3 continuous seconds.
- **Jump Instances**
  This measure returns the mean timespan between individual jump instances.
- **Ducking Instances**
  Returns the mean timespan between individual instances of ducking with the player avatar.
- **Modeshift Instances**
  A measure similar to the total count of modeshifts mentioned above, expressed in the mean amount of time that passes between individual modeshift instances.
- **Tux Deaths**
  Measures the mean timespan between instances in which the player avatar is defeated.
- **Enemy Kills**
  A measure for the mean time between instances in which the player takes out an enemy character.

J.3 PERCENTAGE OF TIME IN REGARDS TO THE TIME THAT WAS TAKEN TO COMPLETE THE LEVEL

- **Time Until Checkpoint** (tracked for whole levels only)
  This measure indicates how much percent of the players' time in the level was spent before reaching the level checkpoint.
- **Time Standing Still**
  A measurement for the percentage of time in the level that was spent standing still with the player avatar.
- **Time Spent Walking / Running / Jumping**
  Expresses the percentages of time in the level that were spent in the individual movement modes.
- **Time in 'Small / Big / Fire / Star' Tux Mode**
  Measures for the percentage of time in the level that was spent in the individual player states.
- **Time Spent Moving Right / Left**
  These measures return the percentage of time that was spent moving into either the right or left direction. The movement modes 'walking', 'running' and 'jumping' are all taken into consideration for this measure.

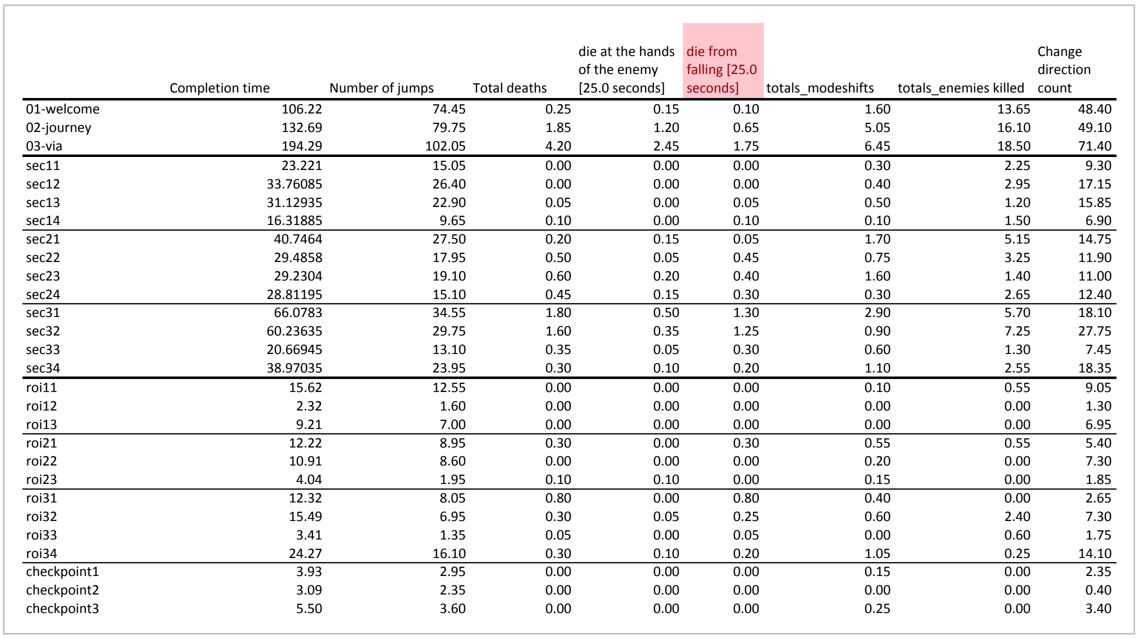
J.4 PERCENTAGE OF INSTANCES IN REGARDS TO THE TOTAL INSTANCES OF A MEASURE IN A LEVEL

- **Tux Deaths Caused by Enemies**
  This measurement returns the percentage of Tux deaths that were caused by enemy contact given all Tux deaths in a level.
- **Tux Deaths Caused by Falls**
  Similar to the previous measurement, however with a focus on Tux deaths that are caused by falling into pits instead of deaths caused by enemy contact.
- **Tux Deaths Before Checkpoint** (tracked for whole levels only)
  A measure that expresses in a percentage how many Tux deaths in a level occurred before a player reached the level checkpoint.
- **Frustration Tux Deaths**
  Returns a percentage for how many Tux deaths in a level were caused by 'frustration'. This measure was added to see if we would get interesting results. For the evaluation we decided to look for repeated deaths within a short amount of time. Such instances of repeated deaths were then logged as 'frustration' deaths, as observation of participants showed that frustrated players tended to play more risky and as a result had more Tux death instances as a result.
- **Enemies Killed** (tracked for whole levels only)
  This measurement returns a percentage for how many of all existing enemies in a level were killed by the player.

- **Enemies Killed by ‘Squashing / Fire / Star / Iceblock’**
  These measurements express the percentages of the individual methods of disposing enemies in relation to all enemies killed in a level.
- **Coins Found** (tracked for whole levels only)
  Shows the percentage of coins that were collected given the full amount of coins that could be collected. In the end we were not able to use this measure since individual coins reappear if the player avatar dies. This is true even for already collected coins. It is therefore possible to collect more than the total amount of coins in a level and consequently percentages exceeded 100% in some cases.
- **Secrets Found** (tracked for whole levels only)
  Returns the percentage of secrets found of all secret areas in a level. It should be noted that only the third level featured two, and therefore more than one secret area.
- **‘Question-Mark Blocks’ Hit**
  This measurement shows a percentage for how many of the total amount of ‘question-mark blocks’ were hit by the player.
- **‘Crate Blocks’ Hit**
  A measure that shows the percentage of ‘crate blocks’ that were hit out of the total amount of ‘crate blocks’ in a level or level section

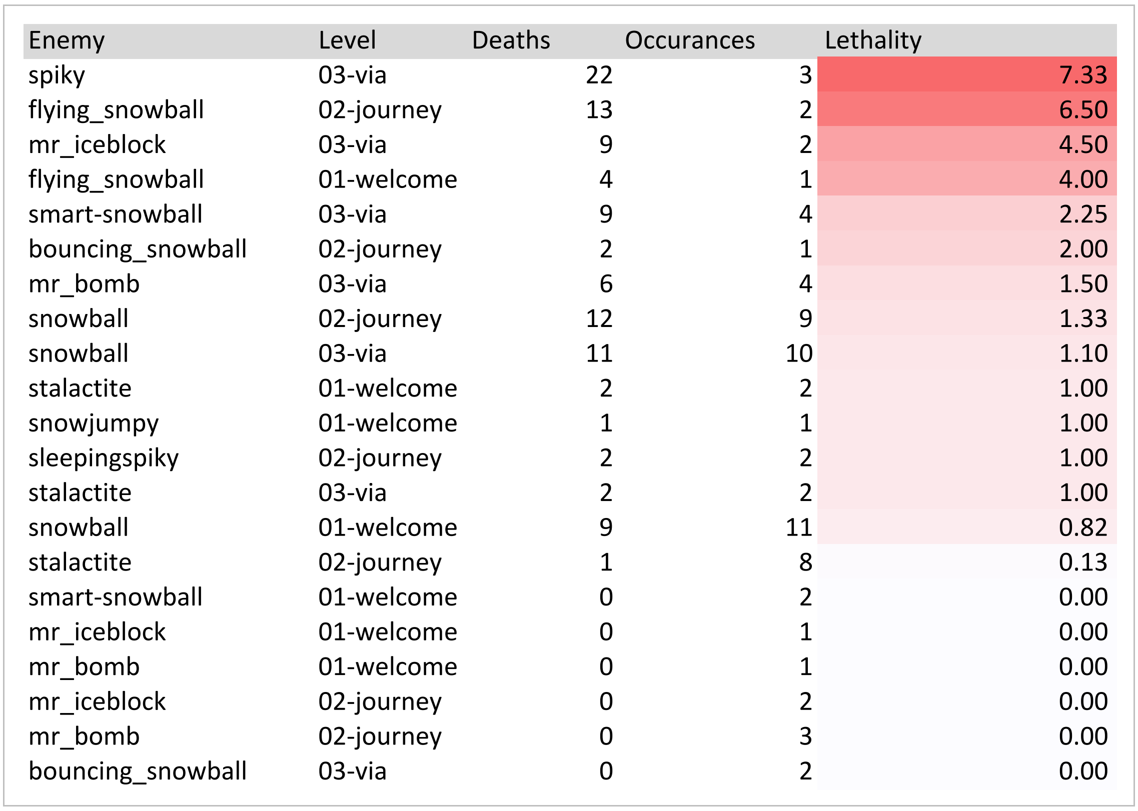
K. Metric Data Results

K.1 Mean Values



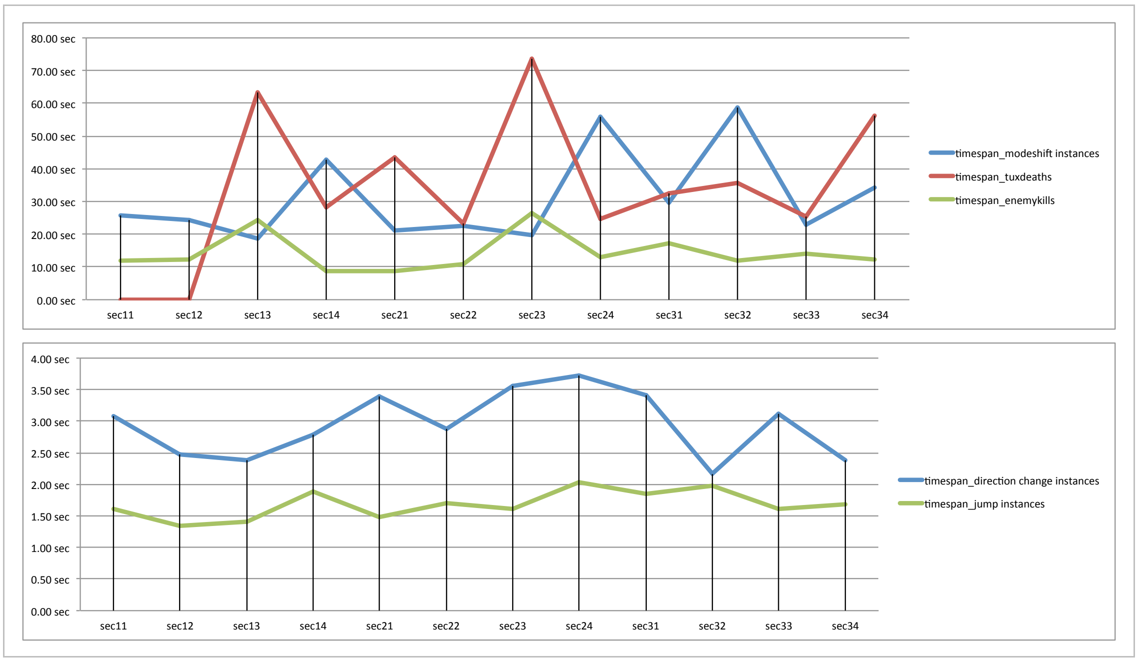
	Completion time	Number of jumps	Total deaths	die at the hands of the enemy [25.0 seconds]	die from falling [25.0 seconds]	totals_modeshifts	totals_enemies killed	Change direction count
01-welcome	106.22	74.45	0.25	0.15	0.10	1.60	13.65	48.40
02-journey	132.69	79.75	1.85	1.20	0.65	5.05	16.10	49.10
03-via	194.29	102.05	4.20	2.45	1.75	6.45	18.50	71.40
sec11	23.221	15.05	0.00	0.00	0.00	0.30	2.25	9.30
sec12	33.76085	26.40	0.00	0.00	0.00	0.40	2.95	17.15
sec13	31.12935	22.90	0.05	0.00	0.05	0.50	1.20	15.85
sec14	16.31885	9.65	0.10	0.00	0.10	0.10	1.50	6.90
sec21	40.7464	27.50	0.20	0.15	0.05	1.70	5.15	14.75
sec22	29.4858	17.95	0.50	0.05	0.45	0.75	3.25	11.90
sec23	29.2304	19.10	0.60	0.20	0.40	1.60	1.40	11.00
sec24	28.81195	15.10	0.45	0.15	0.30	0.30	2.65	12.40
sec31	66.0783	34.55	1.80	0.50	1.30	2.90	5.70	18.10
sec32	60.23635	29.75	1.60	0.35	1.25	0.90	7.25	27.75
sec33	20.66945	13.10	0.35	0.05	0.30	0.60	1.30	7.45
sec34	38.97035	23.95	0.30	0.10	0.20	1.10	2.55	18.35
roi11	15.62	12.55	0.00	0.00	0.00	0.10	0.55	9.05
roi12	2.32	1.60	0.00	0.00	0.00	0.00	0.00	1.30
roi13	9.21	7.00	0.00	0.00	0.00	0.00	0.00	6.95
roi21	12.22	8.95	0.30	0.00	0.30	0.55	0.55	5.40
roi22	10.91	8.60	0.00	0.00	0.00	0.20	0.00	7.30
roi23	4.04	1.95	0.10	0.10	0.00	0.15	0.00	1.85
roi31	12.32	8.05	0.80	0.00	0.80	0.40	0.00	2.65
roi32	15.49	6.95	0.30	0.05	0.25	0.60	2.40	7.30
roi33	3.41	1.35	0.05	0.00	0.05	0.00	0.60	1.75
roi34	24.27	16.10	0.30	0.10	0.20	1.05	0.25	14.10
checkpoint1	3.93	2.95	0.00	0.00	0.00	0.15	0.00	2.35
checkpoint2	3.09	2.35	0.00	0.00	0.00	0.00	0.00	0.40
checkpoint3	5.50	3.60	0.00	0.00	0.00	0.25	0.00	3.40

K.2 Tux Deaths by Enemies

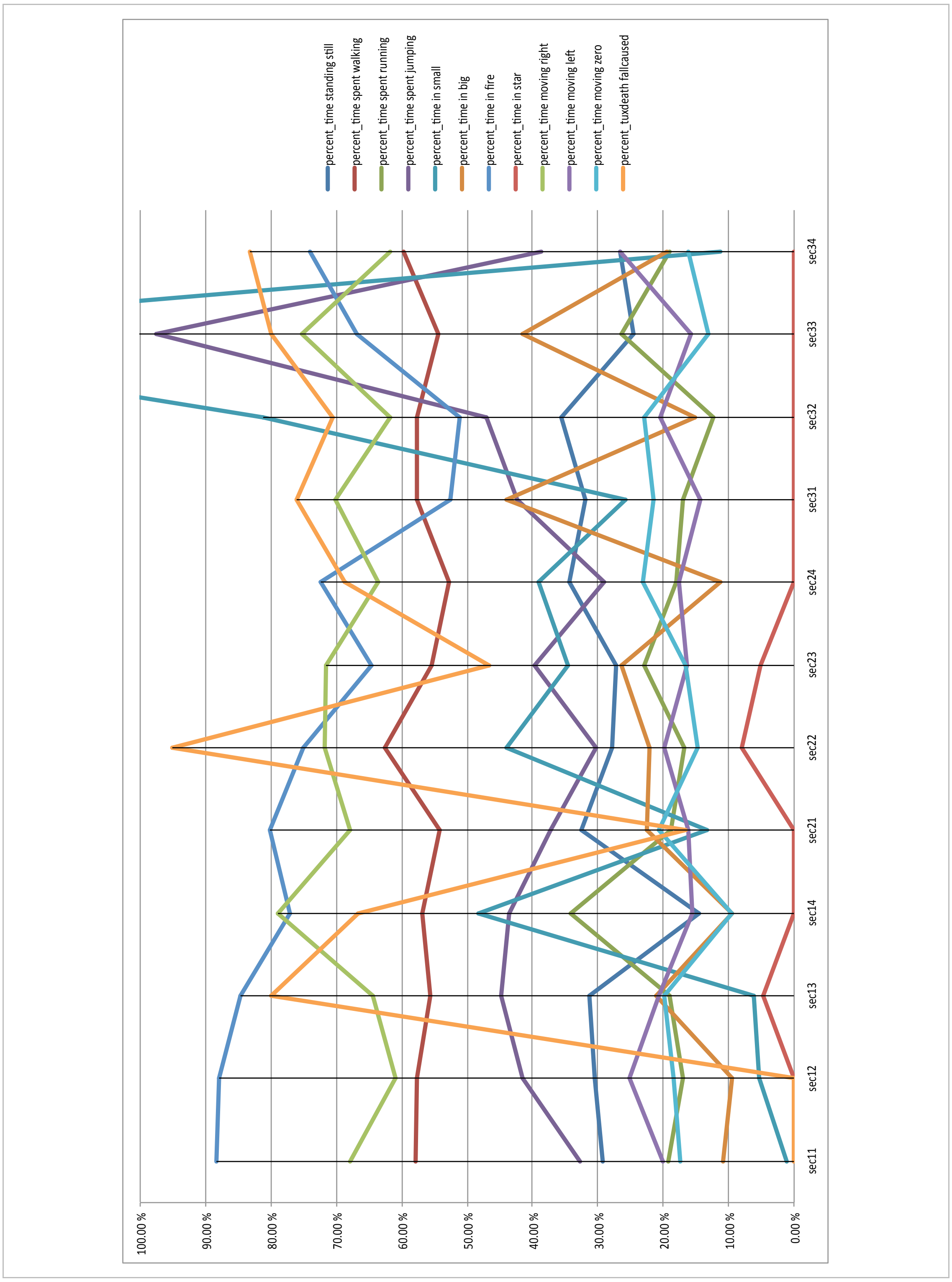


Enemy	Level	Deaths	Occurances	Lethality
spiky	03-via	22	3	7.33
flying_snowball	02-journey	13	2	6.50
mr_iceblock	03-via	9	2	4.50
flying_snowball	01-welcome	4	1	4.00
smart-snowball	03-via	9	4	2.25
bouncing_snowball	02-journey	2	1	2.00
mr_bomb	03-via	6	4	1.50
snowball	02-journey	12	9	1.33
snowball	03-via	11	10	1.10
stalactite	01-welcome	2	2	1.00
snowjumpy	01-welcome	1	1	1.00
sleepingspiky	02-journey	2	2	1.00
stalactite	03-via	2	2	1.00
snowball	01-welcome	9	11	0.82
stalactite	02-journey	1	8	0.13
smart-snowball	01-welcome	0	2	0.00
mr_iceblock	01-welcome	0	1	0.00
mr_bomb	01-welcome	0	1	0.00
mr_iceblock	02-journey	0	2	0.00
mr_bomb	02-journey	0	3	0.00
bouncing_snowball	03-via	0	2	0.00

K.3 Graphs for Timespans

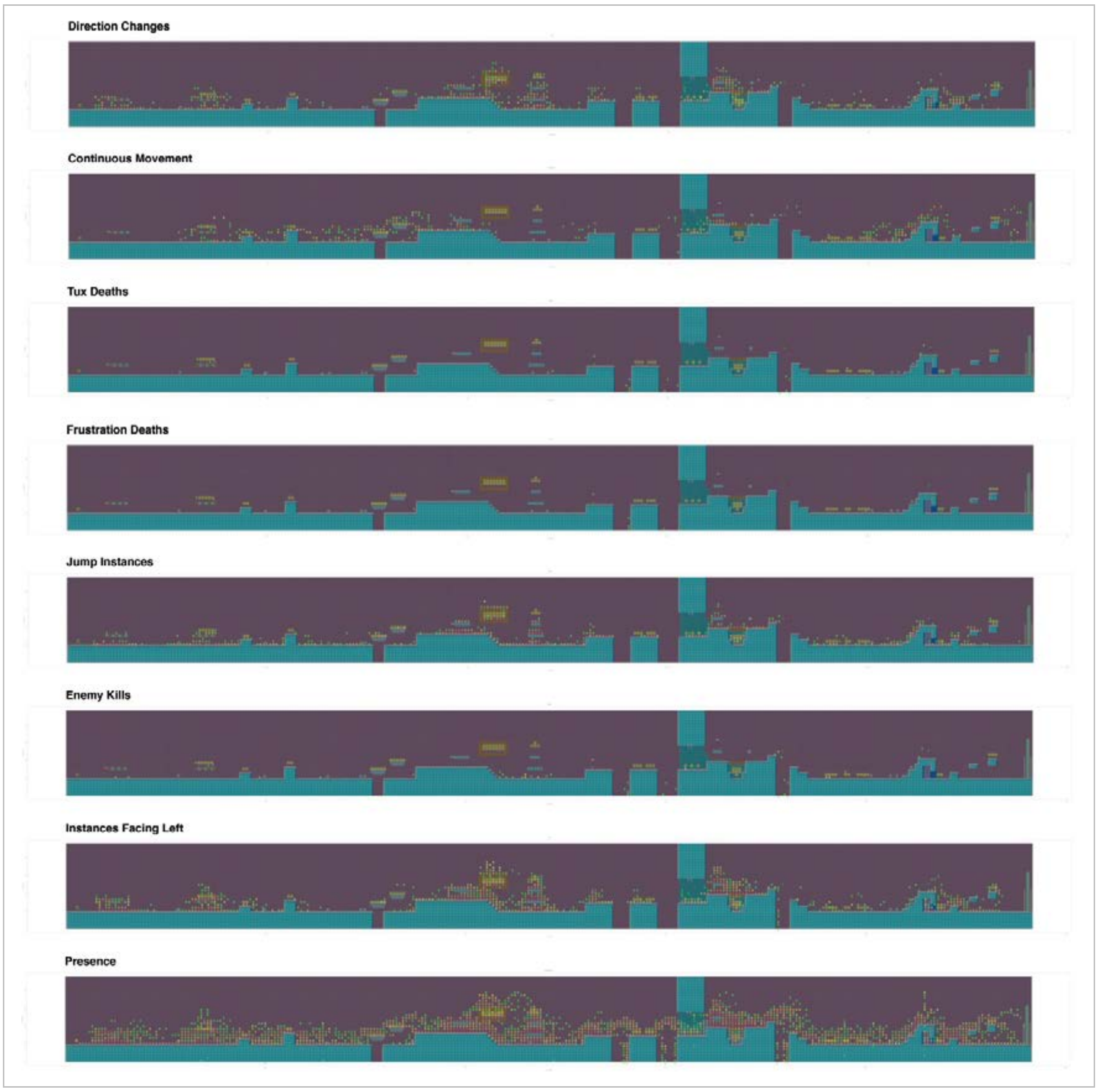


K.4 Graphs for Percentages

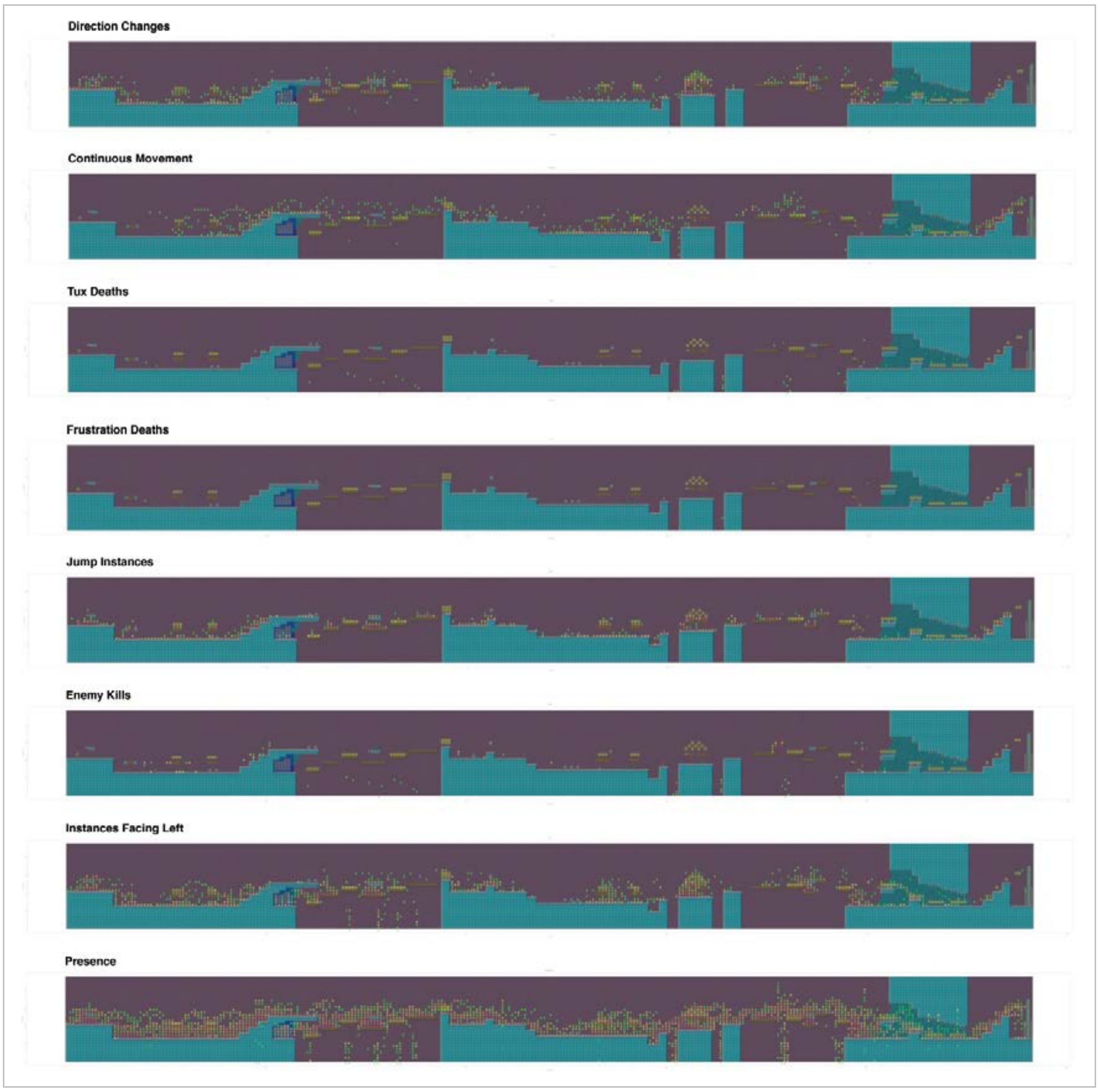


L. HEATMAPS

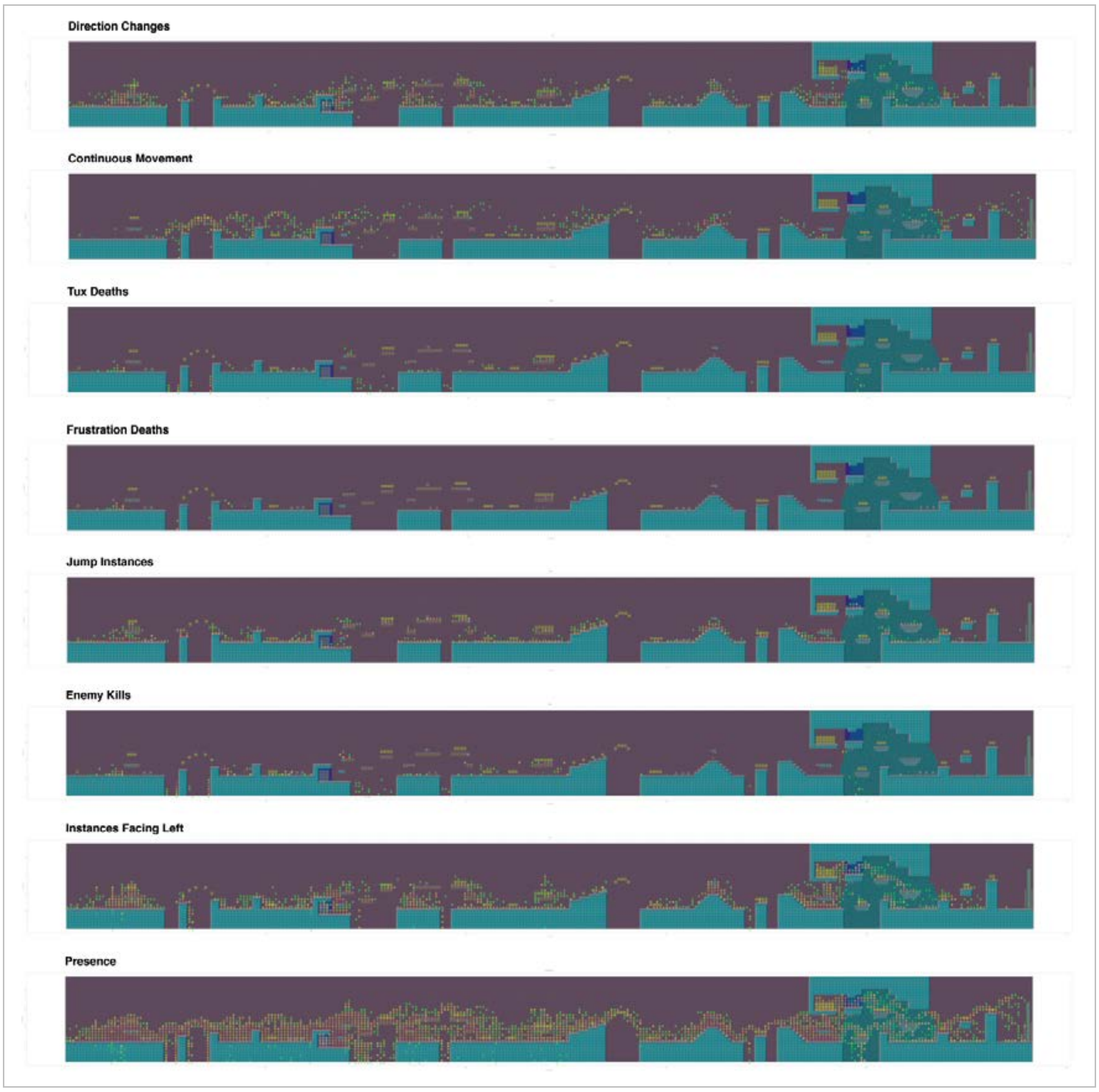
L.1 LEVEL 1



L.2 Level 2


Direction Changes
Continuous Movement
Tux Deaths
Frustration Deaths
Jump Instances
Enemy Kills
Instances Facing Left
Presence

L.3 Level 3



Direction Changes

Continuous Movement

Tux Deaths

Frustration Deaths

Jump Instances

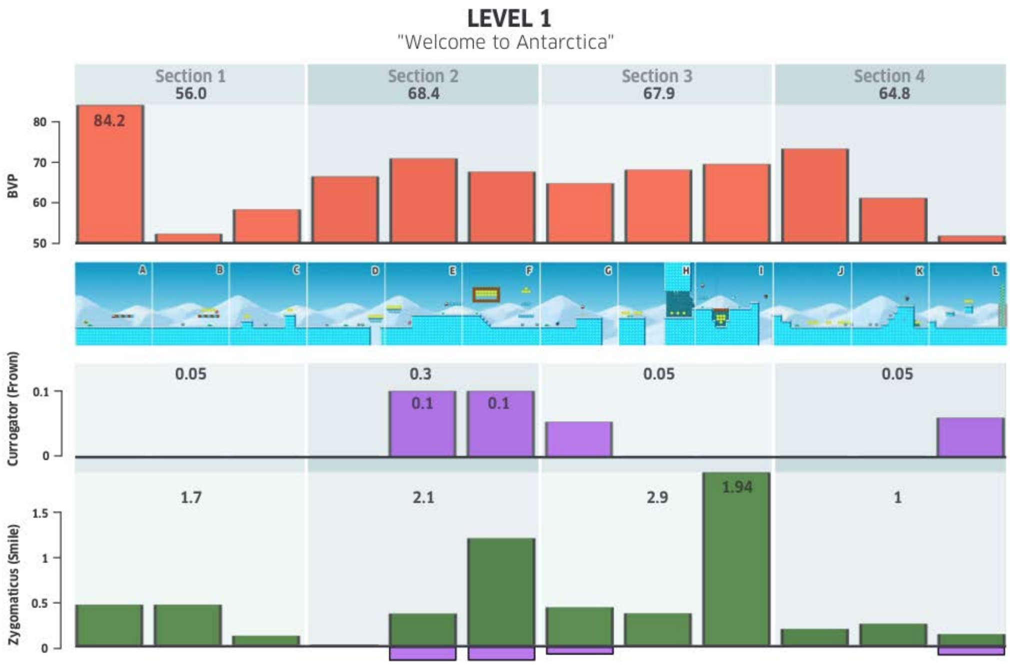
Enemy Kills

Instances Facing Left

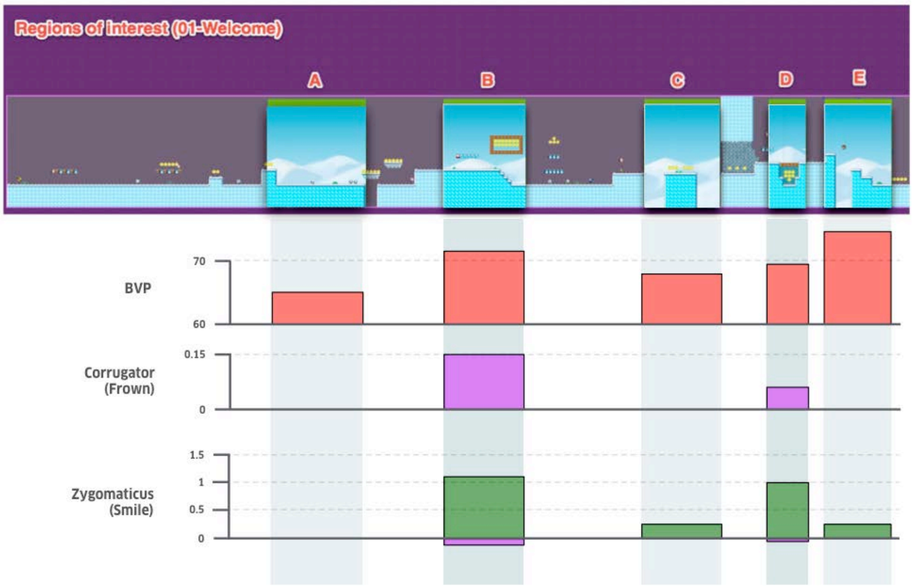
Presence

M. Biometric Result Graphs

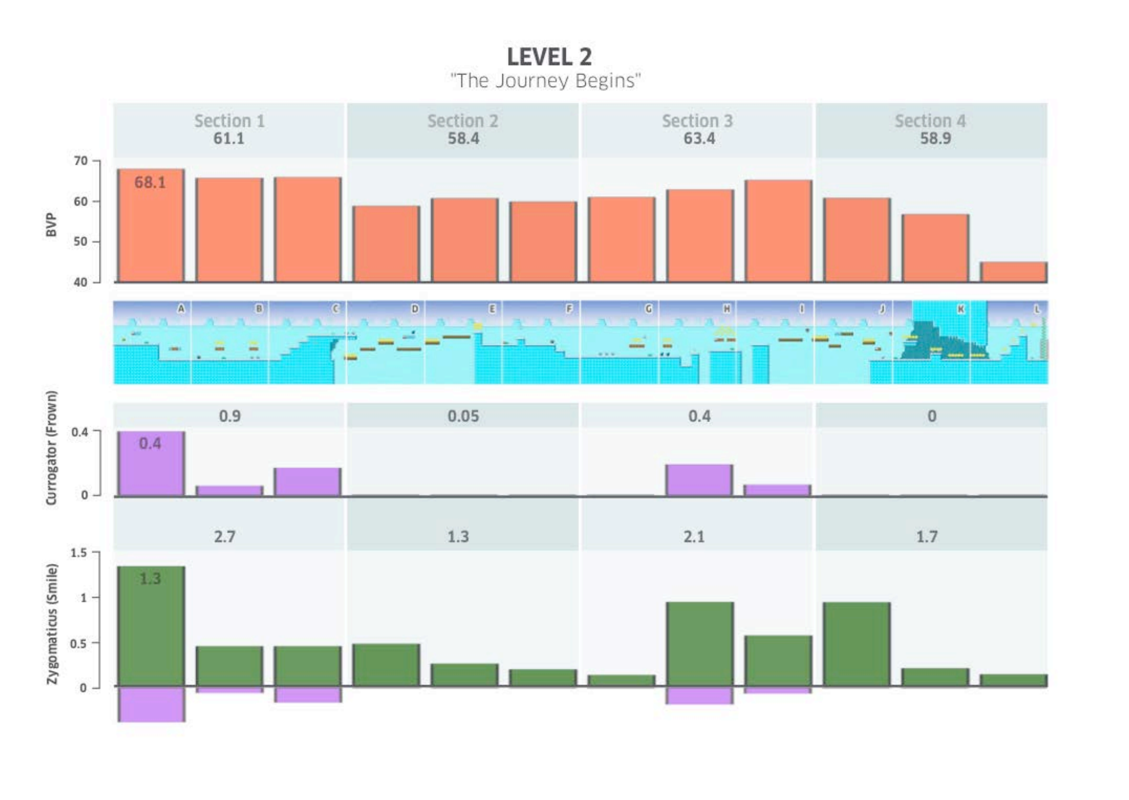
M.1 Level 1 Sections



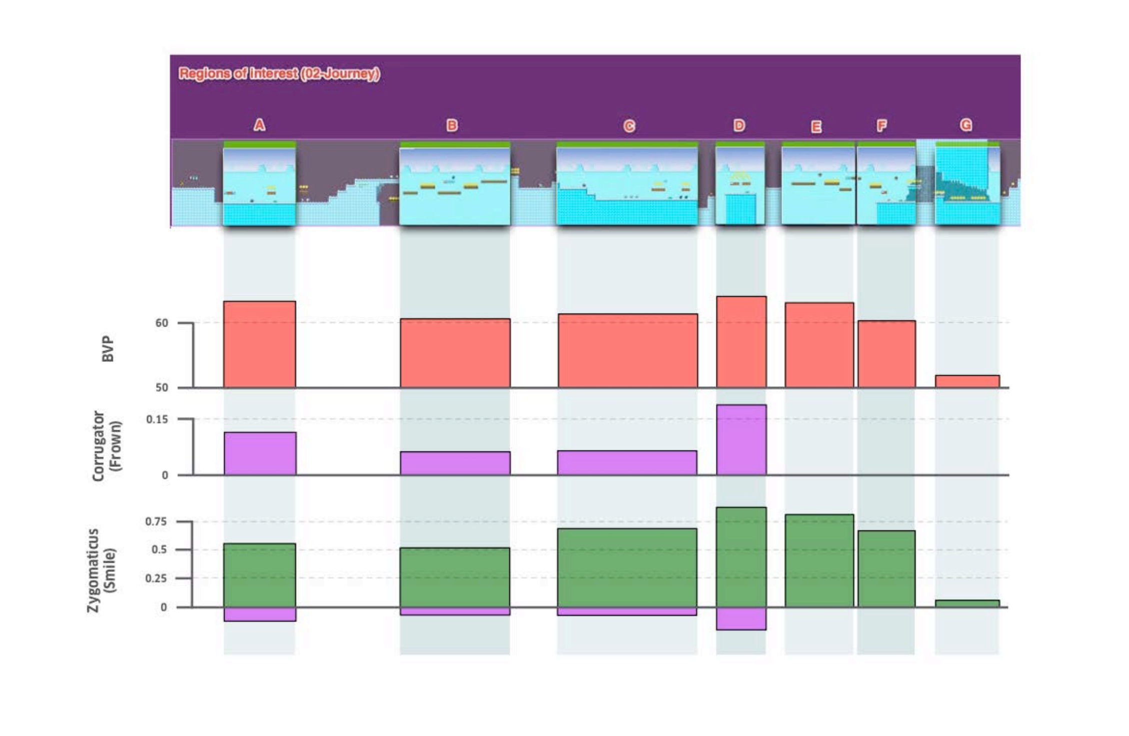
M.2 Level 1 Regions of Interest



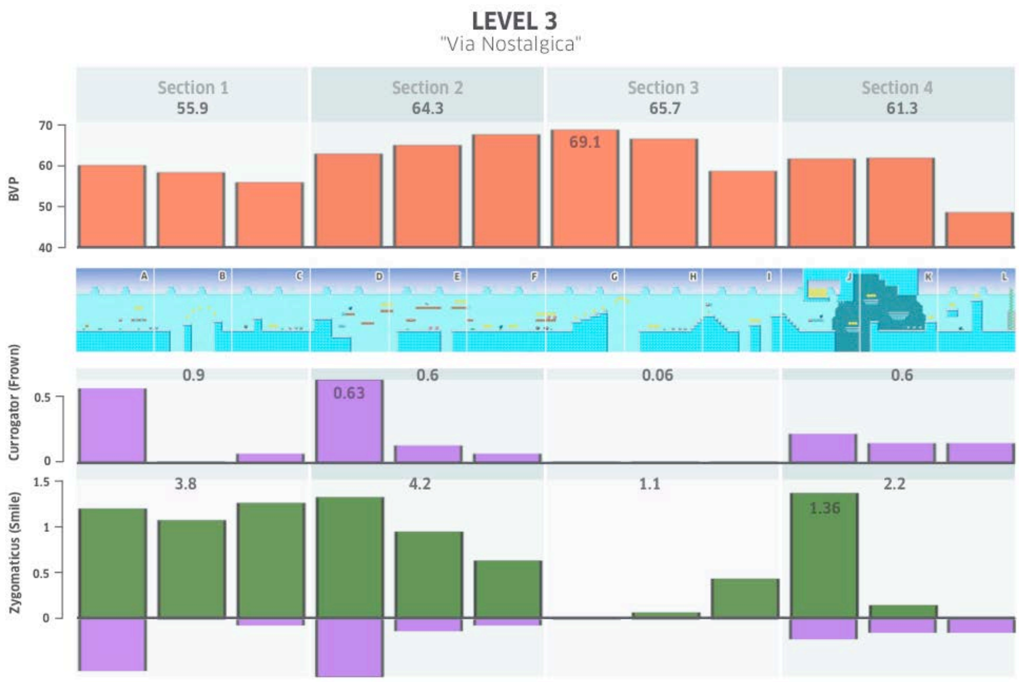
M.3 Level 2 Sections



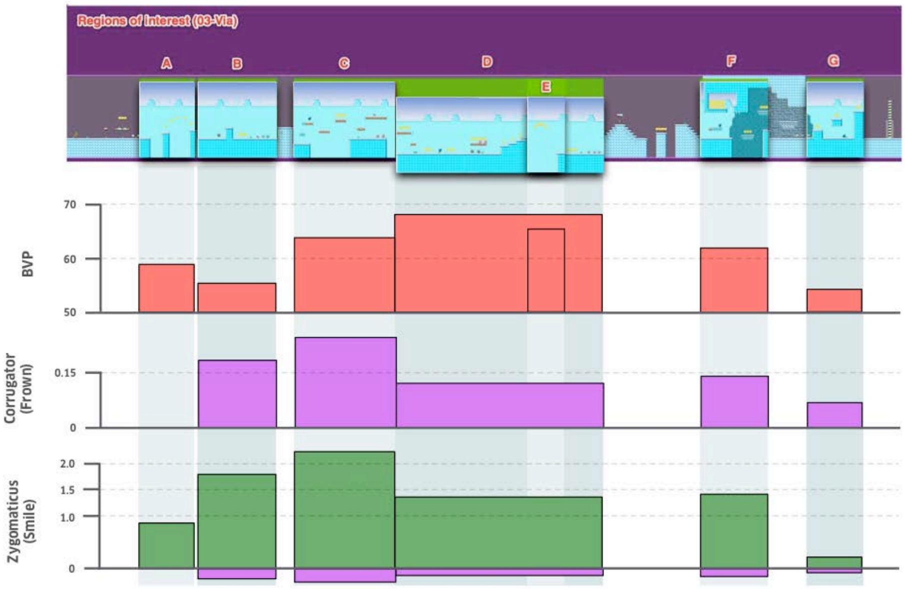
M.4 Level 2 Regions of Interest



M.5 Level 3 Sections



M.6 Level 3 Regions of Interest


Regions of Interest (03-Via)
A
B
C
D
E
F
G
BVP
Corrugator (Frown)
Zygomaticus (Smile)