Evaluating the Use of Affordable User Testing and Visualization Techniques in Level Design of a Hardcore 2D Platform Game

Arthur Silva Bastos, Emanuele Santos and George Allan Menezes Gomes
Universidade Federal do Ceará
Fortaleza, Brasil
arthursb@lia.ufc.br, emanuele@dc.ufc.br, george@virtual.ufc.br

Marcos Arruda Mourão
Centro Universitário 7 de Setembro
Fortaleza, Brasil
71882343@uni7.edu.br

***Abstract*—As the videogame industry evolves, with more diverse and demanding players, making games becomes an increasingly complex task. Modern developers apply Games User Research (GUR) methods to make informed game design decisions based on their target audience. Traditional methods include observation, interview, and questionnaires. In order to obtain detailed user or gameplay information, complementary methods might be required. We analyze the inclusion of two affordable complementary methods, namely webcam-based eye-tracking and telemetry, along with data visualization in a playtesting routine. By developing three versions of a hardcore 2D platform game that demands multitasking abilities using different GUR methods, we were able to find that the chosen complementary methods cover a significant amount of gameplay issues. The metrics and eye-tracking data visualization provided insights about multitasking and level design. Furthermore, we discuss the challenges of evaluating prototypes regarding a more enjoyable experience when frustration is a desirable gameplay element.**

***Keywords*-Level Design; Games User Research; Data Visualization.**

I. Introduction

In general, it is the game designer's duty to assure that a game successfully serves its purpose. However, relying solely on the designers' intuition may lead to unsatisfactory results. Developers employ Games User Research (GUR) methods to obtain meaningful information about players, which ultimately helps them in making informed decisions towards an optimal game experience [1]. The most adopted GUR methods are direct observation, interviews, and questionnaires [2]. These methods excel in answering "why" questions, but may not be reliable for the "what" and "how" questions, which often need quantifiable or measurable information. To collect more nuanced feedback, user testing routines often include complementary methods such as biometrics, eye-tracking, and telemetry (also called game metrics).

There has been an increasing interest in biometric methods since they provide information about emotional states [3]. These methods, however, often require expensive, intrusive equipment. Consequently, small studios, students, and independent developers might not take advantage of them [4]. Eye-tracking methods, however, bring the advantage of assessing players' attention [5] and nowadays can be executed with a simple webcam [6]–[8]. Development platforms, such as Unity [9], offer free data analytics solutions, which makes telemetry accessible as well. Game metrics alongside data visualization techniques can be used to better understand users' in-game behaviors and to balance game mechanics [10].

In this paper, we discuss the inclusion of webcam-based eye-tracking and telemetry, along with data visualization as affordable complementary methods in a development routine. To do this, we developed three different versions of a hardcore 2D platform game and analyzed feedback from user testing. We developed the first version (v1) without user feedback, modified the second version (v2) according to feedback from participant observation, interviews, and a questionnaire; and modified the third version (v3) based on the same traditional methods plus telemetry and eye-tracking. Versions v2 and v3 were then compared using translated edits of the Game Experience Questionnaire (GEQ) [11] and the Positive And Negative Affect Scale (PANAS) [12] to determine which one provided users with the best experience.

Our key contributions include:

- The development of a hardcore 2D platform game and the fundamental design choices that influenced its creation;
- The description of a methodology that uses different accessible GUR methods to improve the level design;
- The development of a novel visualization technique that aggregates in-game data with eye-tracking data;
- A discussion of using affordable user testing and visualization techniques in level design of a 2D platform game.

The remainder of the paper is organized as follows. We review related work in Section II. In Section III we describe the developed 2D platform game. We explain the methodology used to develop the modified versions in Section IV and present our results in Section V. In Section VI, we discuss our results and finally conclude in Section VII, where we

also outline directions for future work.

II. Related Works

Game development is a complex task. Comprehending what makes a good game and how different methodologies contribute to achieving it is therefore of key importance. Gameplay research about improving games tend to focus on games with a purpose (e.g., education) and how to assess processes such as learning [13]. Research on games for entertainment often translates into design recommendations. Examples of this include understanding what makes a game induce *affective* states [14], [15], engage players in a story [16], facilitate immersion [17], or better teach mechanics [18], [19].

Game development can be accessible for everyone using free tools and low-cost methods [20]. Moreover, visualization has also helped to better analyze and understand the data with techniques such as heatmaps [21], user flow [22], and clusterization [23]–[25].

Data analysis can be useful both for developers and players. In online multiplayer games, for example, developers can use data analysis to predict when players will leave the game or cancel their subscriptions [26]. It can also be used by competitive players to better understand the metagame[1] and plan strategies against common enemies [27].

In this sense, the most important related works to our research are the approaches by Mirza-Babaei and coauthors [28] and by Gomez-Maureira and coauthors [29]. In both works, there is the development of multiple versions of a 2D platform game, which are then compared to assess the benefits of different GUR techniques. In the first work, the authors show that biometrics complements the traditional methods well [28], whereas the other contradicts this by showing that psychophysiological data provides fewer insights than the combination of game metrics and direct observation in the specific case of 2D platform games [29].

Our work is similar to those of Mirza-Babaei and Gomez-Maureira, but we replace psychophysiology with eye-tracking, as the latter can be applied reliably using a webcam for this specific case. We want to assess whether telemetry and webcam-based eye-tracking complement traditional user testing methods in a 2D platform game. Telemetry is a viable low-cost solution to assess game balance. Eye-tracking, in the context of games, is usually applied to user interface design [21] and scenery analysis [30]. Including eye-tracking as a level design assessment tool is justified by the fact that attention itself plays a major role during the gameplay in the proposed game prototype described in Section III.

[1]Any information that transcends the game itself and might influence matches, such as knowing which characters are the most common in a fighting game, thus training to countermeasure these characters before competing in a tournament.

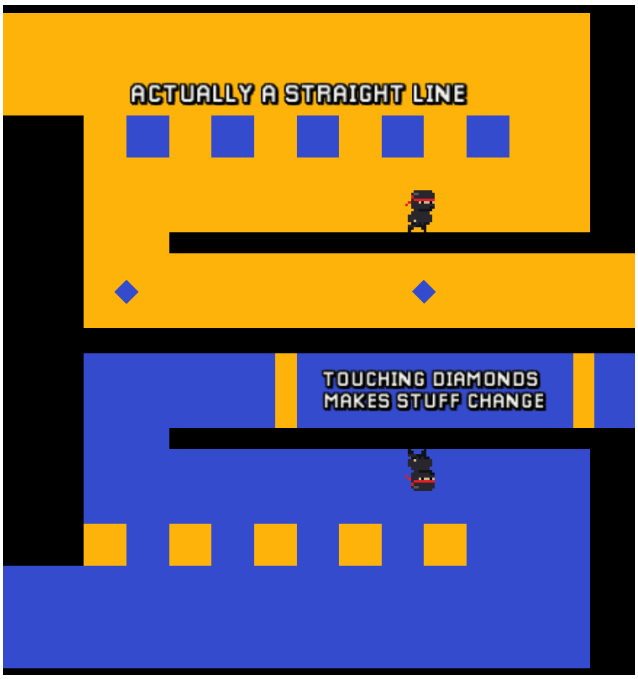

Figure 1. The main concept of *Downside Up* is that players must avoid obstacles and overcome challenges in two mirrored dimensions simultaneously. The game is available at https://arthursb.github.io/Downside-Up/.

III. Case Study - Downside Up

Before describing our research methodology, in this section, we briefly describe *Downside Up*, the game used in our research, and our specific design choices when developing it.

Downside Up is a hardcore 2D platform game in which players must guide an avatar through two mirrored screens. Objects may appear in only one of those screens, besides existing in both, so players must use their gaze to understand and overcome challenges. Fig. 1 shows a screenshot of the game. The game is composed of three levels, and each level consists of smaller challenges. There are 18 challenges altogether. We labeled each challenge with a number and a letter and sorted them alphabetically, so "3D", for example, refers to the fourth challenge in level 3. We expect players to finish the game in 40 minutes.

Regarding the game genre, data from 2017 show that action games are among the most popular genres in the United States [31], which is the largest gaming market in the world. Based on this, we chose the 2D platform genre, which is a subset of the action genre. Its simplicity and well-known mechanics (walk and jump) bring the advantages of reducing time spent by players learning basic mechanics, which in turn make the user testing process faster and more manageable.

Inspired by the games *Celeste* [32] and *I Wanna be The Guy* [33] we decided to follow the game design principle

of either having seemly unexpected traps that stimulate trial and error or having difficult challenges that require fast and precise reaction times. Games like these are informally referred as "platform hell" or "masocore". Similar to *Chronos Twins* [34] and *Binaries* [35], users coordinate two characters simultaneously, but face different obstacles with them (see Fig. 1). As an offline game, *Downside Up* did not provide means of interaction between players. As we thought this was a significant trait of fun, we simulated it by adding a narrator that makes constant comments taunting or provoking players, as well as providing hints (see the texts in the background of Fig. 1), similar to *Binaries*. Also, inspired by the game *Sometimes You Die* [36] we added a mechanic of not undoing players' actions when the avatar dies. We believe these mechanics subvert the usual 2D platform genre design expectations, thus making it enjoyable and an exciting opportunity for research.

By using SteamSpy [37], we found that players of *Celeste* and *Binaries* described these games with common tags such as "difficult", "puzzle-platformer", "minimalist", "funny" and "fast-paced". This not only suggested these games had a common audience but provided us design directions as well.

We followed a minimalist approach and developed the game with the least distracting visual elements possible to keep players focused on the challenges. Colors were kept simple, favoring high contrast. To avoid possible motion sickness issues, most challenges were designed in a way to avoid constant eye movement between the up and down portions of the game.

Our team, composed of a game designer and a programmer, developed the prototypes of *Downside Up* using Unity and a third-party customizable controller [38]. We developed three levels with increasing difficulty in three months. Art assets were either provided by the controller's examples or generated using Unity itself.

IV. Methodology

Our approach intends to evaluate to what extent GUR improves the level design of 2D games using *Downside Up* as a case study. We consider in the context of this research that overall level design quality is directly related to what players perceive as "fun", which is the positive emotional response to learning, puzzle-solving and overcoming challenges [39]. As a consequence, the most successful prototype is the one users report to have the most positive responses and the least negative ones. Other important aspects of a fun game, such as artistic properties, are left out the scope, therefore not analyzed despite the presence of game design decisions made towards them.

Digital game experience is often self-reported. It can be categorized in dimensions such as enjoyment, flow, immersion, suspense, competence, positive and negative affect, control, and social presence [40]. Fun is a complex subject strongly dependant on context and users' preferences, gaming background, and mood [41]. We do not rigorously evaluate aspects of the game experience that could be considered fun, such as flow, immersion, and attention. Our work assesses fun through user-reported emotions. To assess it, we applied two questionnaires: GEQ [11] (core module and post-game module) and PANAS [12]. We edited GEQ, so the core module kept only questions that assess positive or negative affect, and the post-game module kept only questions that assess positive or negative experience. Both questionnaires were translated to Portuguese. Our methodology is divided into three phases and is summarized in Fig. 2.

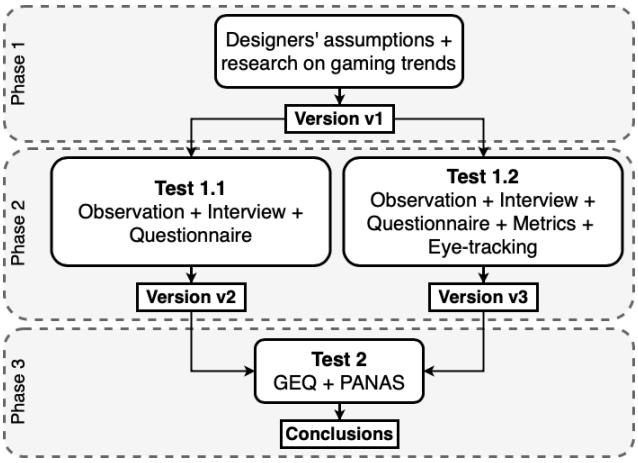

Figure 2. A diagram summarizing the development and research processes.

A. Phases

In Phase 1, there is the development of the first version of *Downside Up* (v1) based on assumptions about the game's target audience. In Section III, we described the game design choices and justifications.

Then, in Phase 2, there is the first user test with v1 (Test 1.1), in which we adopted direct observation, interview, and a custom challenge difficulty assessment questionnaire. Results from this test provide data for a series of level design changes, which are applied to v1 to generate a new version (v2). There is also another test of v1 adopting the same methods plus in-game metrics and webcam-based eye-tracking (Test 1.2). This other test generates a new report, which in turn provides design recommendations that, after applied to v1, will generate v3.

Finally, in Phase 3, players play either v2 or v3 and fill the GEQ and the PANAS questionnaires (Test 2). The results are then compared to determine which of these versions is the most fun to play.

B. User Testing Procedure

In Test 1.1, users are invited to participate and fill a consent form after agreeing to it. They play the game after

Figure 3. Five consecutive frames of a replay visualization. A custom avatar (circle) replaying an anonymous user test session. The colored background indicate to which portion of the screen the user was looking at in any given moment. Conversely, a dark background indicates that the player was not looking to that portion at the time. Observe that in this example (Challenge 2C) the colored portions alternate, indicating rapid eye movement. You can check video samples of this visualization at https://arthursb.github.io/Downside-Up/.

a simple explanation of the game's mechanics and controls. While they play, researchers may take notes about their behavior. After finishing the game, players participate in an informal interview to provide qualitative information. When the interview is over, participants fill the difficulty assessment questionnaire and receive a chocolate candy as compensation for participating in the test.

In Test 1.2, besides the activities described above, a Unity script tracks certain player's properties and saves them in a text file. Before playing each level, players are asked to calibrate WebGazer [6], the tool we used for eye-tracking.

Test 2 consists of the same procedure of Test 1.1 with three differences: there are no interviews, no observation notes, and participants fill the GEQ and PANAS instead of the custom difficulty assessment questionnaire.

C. Collected Data

We categorize the data we collected into four types which we explain below.

1) ***Observation and Interviews****:* observation notes describe players' behavior or in-game difficulties. Interviews' questions are informally structured and vary according to what was observed. They address specific problems, for example, why a specific user had difficulty in a specific challenge, or what was the cognitive process a user took to solve a specific problem.

2) ***Difficulty Assessment Questionnaire****:* as we mentioned before, we fragmented each level into smaller challenges, totaling 18 challenges. Participants respond in a 5 item Likert scale how difficult was each challenge, ranging from 1 ("too easy") to 5 ("too hard"). We compare the results from the Likert scales with the designers' expectations.

3) ***Telemetry****:* when required, the following metrics were tracked: avatar position and location of in-game deaths. These metrics help to understand if challenges are taking too long because: (1) they have usability issues; (2) they are hard to understand; (3) they are hard to see; (4) they require movements that are hard to perform.

4) ***Webcam-based eye-tracking****:* initial experiments showed that WebGazer's precision was low. For this project, however, it is only necessary to track the Y-axis on the computer's screen, thus making it more reliable. We modified the application[2] to log when users were looking to the upper or lower half after the calibration process is over.

[2]Available at: https://webgazer.cs.brown.edu/calibration.html

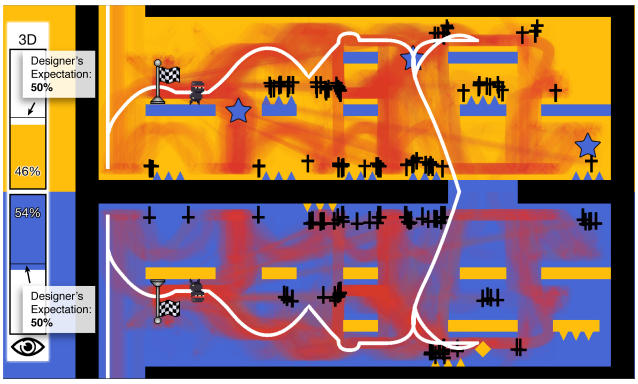

Figure 4. Visualization of challenge 3D. Crosses around trigger-controlled spikes suggest that these deaths happened because players did not see them. The white path represents the optimal solution. Red paths represent actual players' movements (154 paths in this segment).

D. Visualizations

Through the combination of the collected data, we produced two types of visualizations, one for individual users and another that aggregates the data of all participants.

By syncing timestamps, we can see replays of gameplay sessions and change colors of portions of the screen according to where players were looking at the time. This visualization, however, is only suitable for specific segments of gameplay. Some full replays might last hours, therefore requiring an unreasonably long time to watch. Fig. 3 shows this visualization.

To better understand overall behavior, we developed a visualization that aggregated for all users every variable relevant for this research.

For trajectory data, we adapted previous techniques [23], [42] and built a visualization shown in Fig. 4: it shows a white path that represents the ideal trajectory for each level along with red paths with a low alpha value, which are the actual paths performed by the participants (seven, in this case). Superposed red lines get progressively darker, meaning users repeated the same path frequently.

Black crosses represent players' deaths. A mark in the lower screen, for instance, represents a death that happened while the player was looking down. Using this, we could

check which sections of the game contained potential accidental deaths.

Finally, we determined the distribution of gaze per challenge by joining Unity's spatial data with eye-tracking data. We represented this with a bar chart in a sidebar. The two mirrored portions represent the percentage of time spent looking at that portion. The black solid line inside the bars indicates what designers expected.

V. Results

In this section, we present the results of the tests present in the three phases of the methodology.

A. Phase 2 - Development of Version v2 (Test 1.1)

In this subsection, we present the results from the first test performed in Phase 2, which involved observation, interviews, and a questionnaire. Participants were recruited through mailing lists, social media groups or in person. The tests were conducted in a room with only a desk, chairs, and the necessary components for the game and eye-tracking procedures. Each participant played the game in a Late 2014 Macbook Retina[3]. The tests took one month with 12 participants. The majority of them aged between 21 and 26 (8 players), had the habit of playing games for 8 or more hours a week (7 players) and preferred to play on the computer (9 players).

1) ***Observation notes and interviews****:* players seemed to take a little more time than expected to understand the basic mechanics of the game, only understanding it completely in level 2. They reported having spent more time looking at the upper half of the screen and having perceived the bottom one as "shadow". This behavior suggests that the game tutorial is either not challenging enough or not teaching basic mechanics properly.

We also identified a few usability issues: players felt the controls were unresponsive in certain sections, there were abrupt camera shifts, many spikes were positioned in a way such that it felt unfair to touch them and served no purpose besides adding frustration. One player had to ask for help to overcome some dexterity challenges. In challenge 2D, which consisted of a room filled of triggers and spikes, a new mechanic is introduced with no explanation, so players had to understand it with trial and error.

Players seemed to die the most in the last level. They had not explicitly reported it, but we observed that participants were, in general, annoyed with constant accidental deaths. Therefore, it is necessary to reevaluate the position and quantity of spikes.

Players praised the overall mechanics. The game design principle of tricking players into traps was well-received, as many participants played smiling, yelling at themselves or looking surprised after completing a challenge. They also reported having liked the subtle hints given by the narrator.

[3]Technical specifications: https://support.apple.com/kb/SP704

2) ***Questionnaire****:* responses for Test 1.1's difficulty assessment questionnaire are summarized in Fig. 5. For all the three levels, responses tended to match designer's expectations, with the exceptions of challenges 1A, 1B, 2A, 2C, 2D, 3B and 3F. Some of them had design issues that will be explained next.

We identified that challenges 1A and 1B, which were supposed to teach all jumping mechanics, didn't teach air jumping properly. This caused an increased perceived difficulty (*Easy* labeled more times than *Too Easy*).

Challenge 2C was labeled as *Too Hard* most likely due to an usability issue. To solve this challenge, players had to find a secret passage in the ceiling, that would be revealed when the avatar gets close to it. It was the first time that the camera could move vertically despite the game not presenting any indications of this.

Challenge 2D presented a new mechanic with no explanations or safe places for exploration. Players had to understand this mechanic through trial-and-error. Interestingly, the majority of them labeled the challenge *Just Right*.

We believe few players labeled 3B as *Too Easy* because it happens entirely in the reflected portion of the game.

The last noteworthy challenge is 3F, which requires careful planning and precise motor skills to maneuver through a narrow corridor filled with spikes. We theorize this dexterity challenge was perceived as not too hard because the more they repeated it, the more skillful they became.

B. Phase 2 - Development of Version v3 (Test 1.2)

Here we present the results from the second test performed in Phase 2, which involved observation, interviews, a questionnaire, in-game metrics, and eye-tracking. Tests happened under a similar environment to Test 1.1's (see subsection V-A) and took two weeks. We had 7 participants. The majority aged between 18 and 20 (3 players), had the habit of playing games for 2 to 8 hours a week (6 players) and preferred to play on the computer (3 players).

1) ***Observation notes and interviews****:* results were noticeably similar to those reported in Section V-A. A higher percentage of players had difficulties understanding the core mechanics. The same usability issues identified in Test 1.1 were identified in Test 1.2. Based on this, we expect that v3 will be easier than v2. No player mentioned the narrator or any other aspect of the game besides the level design itself, which was praised for creativity.

2) ***Questionnaire****:* results for this second test did not differ from the results of the first one and can also be seen in Fig. 5. We consider that both groups of players perceived difficulty similarly. Notably, no player's response matched designers' expectations for challenges 2B and 3B. Both of these challenges were expected to be the easiest in their respective levels.

3) ***Metrics and eye-tracking****:* replays of challenges (see subsection IV-D) displaying color flickering suggested that

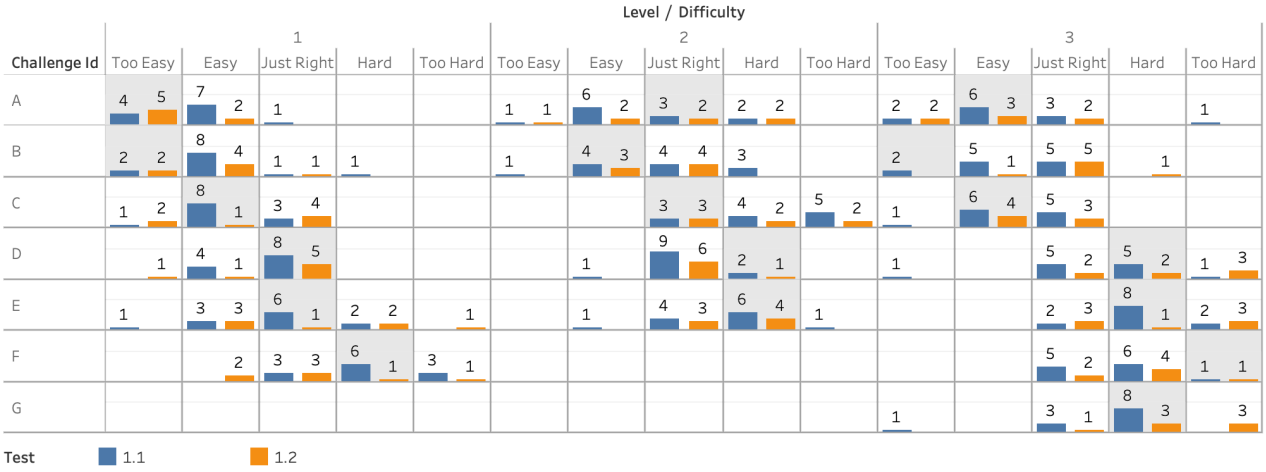

Figure 5. Responses from Tests 1.1 and 1.2 regarding difficulty of version v1. Cells highlighted in gray represent the designers' expectation for the difficulty of each challenge.

there was constant eye-movement, i.e., players were either confused about how to manage information in both screens or were exploring the whole scenario before trying to solve the puzzle. In some cases, such as challenge 1E, this is intentional, as this is the final challenge of the first level and was designed to check if players learned that it is necessary to understand both screens to progress. In other cases, such as challenge 2C, where there was a lesser need for attention to the lower portion, flickering was not expected, therefore reinforced the usability issues observed by the tester and reported by users in interviews. This flickering issue happened, for instance, when players tried to jump to the platform that led to a secret passage (see Fig. 3).

In challenge 3D (Fig. 4), players must explore the environment to find the correct trigger that unlocks a blue platform on the floor. Using the aggregated view, we found that many unwanted deaths occurred around trigger-controlled spikes. Moreover, we expected that both portions of the screen would share attention equally, but players spent a little more time looking to the lower half. This indicates that either the elements in the lower screen were more distracting than the upper screen or they were perceived as more dangerous than the ones in the upper screen. A solution to this was to simplify the challenge and remove spikes.

Interestingly, conclusions from data visualizations not only reinforced what was observed and discussed with players but provided more insight about the multitasking nature of the game as well. We could determine unfair spikes that might have caused unnecessary frustration, unwanted platforms, or texts in the background that caught attention and challenges that should have elements better mixed between the two screens. We found difficulties regarding maneuvering the avatar in the lower screen.

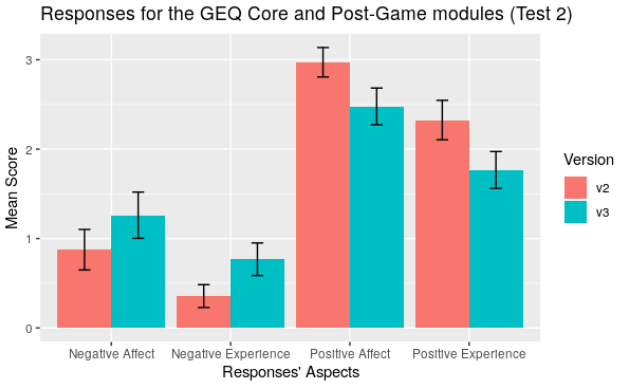

Figure 6. Responses for the GEQ Core and Post-Game modules (Test 2) with 95% confidence intervals.

C. *Phase 3 - Comparing prototypes (Test 2)*

In this subsection we present the results from the last test, performed in Phase 3, in which players fill the GEQ and the PANAS questionnaires after playing either v2 or v3 (randomly assigned). We recruited participants through the same means of Phase 2. This time, participants performed the tests in a laboratory equipped with Late 2015 21.5 inch iMac computers[4]. Tests consisted of 45 participants and took one week. The majority aged between 21 and 25 (20 players), had the habit of playing games for more than 8 hours a week (21 players) and preferred to play on the computer (34 players). Among the 45 players, 22 played v2, and the other 23 played v3.

The GEQ results, displayed in Fig. 6, point that players felt more positive affect and less negative affect playing v2

[4]Technical specifications: https://support.apple.com/kb/SP733

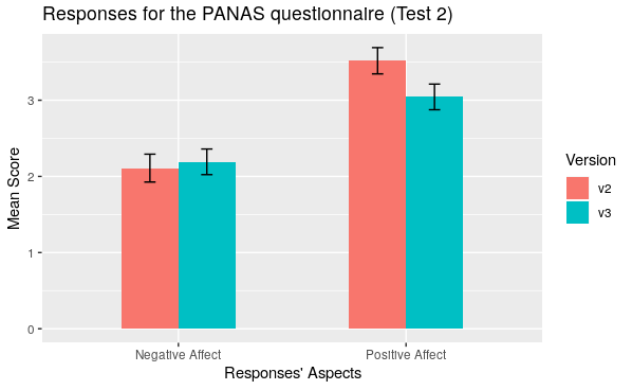

Figure 7. Responses for the PANAS questionnaire (Test 2) with 95% confidence intervals.

Table I
WILCOXON'S RANK SUM TEST RESULTS AND EFFECT SIZES FOR BOTH QUESTIONNAIRES' RESPONSES.

Questionnaire	Aspect	Wilcoxon's rank sum test result and effect size
GEQ	Positive Affect	$W = 7947, p < .001, r = -.23$
	Negative Affect	$W = 3337.5, p = 0.03, r = -.16$
	Positive Experience	$W = 11393, p < .001, r = -.22$
	Negative Experience	$W = 7084.5, p < .001, r = -.13$
PANAS	Positive Affect	$W = 30616, p < .001, r = -.19$
	Negative Affect	$W = 23905, p = 0.2825, r = -.05$

($M = 2.97$, 95% CI $= 2.81, 3.13$ and $M = 0.88$, 95% CI $= 0.65, 1.11$, respectively) than playing v3 ($M = 2.48$, 95% CI $= 2.27, 2.69$ and $M = 1.26$, 95% CI $= 1.0, 1.52$, respectively). Furthermore, their experiences playing v2 were both more positive (v2: $M = 2.33$, 95% CI $= 2.11, 2.55$ and v3: $M = 1.77$, 95% CI $= 1.56, 1.98$) and less negative (v2: $M = 0.36$, 95% CI $= 0.30, 0.42$ and v3: $M = 0.77$, 95% CI $= 0.59, 0.95$).

The PANAS results (Fig. 7) show that players felt more positive emotions in v2 ($M = 3.52$, 95% CI $= 3.35, 3.69$) than in v3 ($M = 3.04$, 95% CI $= 2.87, 3.21$). They also felt slightly less negative emotions in v2 ($M = 2.11$, 95% CI $= 1.93, 2.29$) than in v3 ($M = 2.19$, 95% CI $= 2.02, 2.36$).

To check whether these results are significant, we performed Wilcoxon's rank-sum test [43], a non-parametric test for independent samples, after using the Shapiro-Wilk test to verify that our data were not normally distributed. The results of the Wilcoxon's test are displayed in Table I and we can observe that v2 did not differ significantly from v3 only in the Negative Affect in the PANAS responses because its p-value ($p = 0.2825$) was above .05. Based on these results, v2 would be considered the version players enjoyed the most and so the most successful version. We discuss many aspects of these results in the following section.

VI. DISCUSSION

In this section, we will discuss many aspects of our study to explain the results described in Section V. We divided our discussion into three different topics: *Downside Up*'s design, Games User Research, and limitations.

A. *Downside Up's design*

Downside Up, like the so-called "masocore" games, purposely taunts players and intersperses tough challenges with easy, tutorial-like challenges. Although the game provides no elements of randomness or chance, traps were positioned in ways that felt unfair. It is an admittedly controversial game design principle as it forces players through trial and error behavior in some challenges even with checkpoints and infinite lives, thus making it hard to evaluate whether a challenge is adequate.

The "masocore" games are a subset of platform games, which in turn are a subset of action games. Finding participants who enjoy this genre and fit the designated public within our time constraints was therefore troublesome. Considering all participants, the majority preferred to play RPG or MOBA (Multiplayer Online Battle Arena) games, which deviates from our intended audience.

Furthermore, players consistently demonstrated "love it or hate it" reactions to the game. Although players were angry, during all tests, many participants informally complimented it, expressing that it instigates a sense of mastery when playing. Since the narrator not only taunts but also congratulates, players found overcoming difficult challenges rewarding. Similarly, many players reported the game to be excessively hard, and the difficulty alone would justify them abandoning the game if they were playing casually. Also, though we had no data to confirm this in Phase 3, we believe, based on observation notes and interviews during Phase 2, that "good" frustration happens when the game taunts players or when players feel it is their fault they could not overcome a challenge, while "bad" frustration happens when players blame the game for not being able to progress. Examples of this include: spikes positioned in places where unfair deaths occur, sections where it is hard to maneuver the avatar, platforms that require jumps hard to perform, sections where the camera hinders puzzle-solving and sections that require quick reflexes and memory about elements on both sides of the screen at the same time.

One disadvantage of introducing intentional frustration challenges is that some of them were made to be unusually hard in their context. This potentially reduced the effectiveness of GUR methodologies. Observation and data visualization played a major role in this case, as the designer could observe if players considered intentionally hard sections

difficult or if there were glitches or issues such as camera transitions.

Another fact we observed in *Downside Up* is related to dealing with two screens simultaneously. We found that players tend to focus on the upper screen, which behaves like traditional platformers. The lower screen mirrored the upper screen horizontally. Interesting enough, in challenge 3F, that consisted of an S-shaped corridor filled with spikes, the first two horizontal portions were the same, but all spikes could be seen in only one of the two sides of the screen. Players reported the second portion, which happened entirely in the lower half of the screen, was considerably harder. We believe that the extra cognitive load caused by the lower screen might affect players' performance.

B. Games User Research

The combination of metrics and eye-tracking not only confirmed data from participant observation and interviews but also provided more insightful means of understanding level design and multitasking in *Downside Up*. As a drawback, they did not provide any data about what players felt or how was their cognitive process clearing challenges. We believe we could solve this issue by adding space for comments in the questionnaire. We concluded that, for this experiment, the combination of metrics, eye-tracking, and the questionnaire could cover all important aspects of the game.

Phase 2 consisted of testing v1 and finding areas of improvement. As already discovered in Mirza-Babaei's experiment [28], we also found that the inclusion or modification of challenges should not have gone untested before Phase 3. Even backed up by data, the game design process is still subjective. We theorize that a modification in challenge 3F made the final section undesirably frustrating, which led to lower scores for v3 in the GEQ and the PANAS. Fig. 8 shows these modifications.

Although we defined an audience that is composed of players experienced with platform games, most of our participants were not. Inexperienced players, however, provided better feedback regarding introductory challenges and helped to highlight usability issues. We found that feedback from expert users was not as rich for tutorials as they quickly learned mechanics. Their feedback provided better means to tune more complicated challenges towards the end of each level.

We found that visualizing data from different sources facilitated comprehension of how our participants played the game in a simpler, less time-consuming manner than observing and annotating every participant's behavior. It also provided data that could not be easily grasped by other GUR methods, such as the identification of accidental deaths in-game.

Our results point that, for this specific game, using WebGazer as a level design tool led us to valuable data about attention. As already mentioned, WebGazer's inaccuracies might render it unfeasible for certain genres of games. The tool worked for *Downside Up* mainly because we could split the screen into two areas of interest. We theorize that this solution might be useful as well for games with few or no camera changes and a screen space that can be safely discretized in up to four portions.

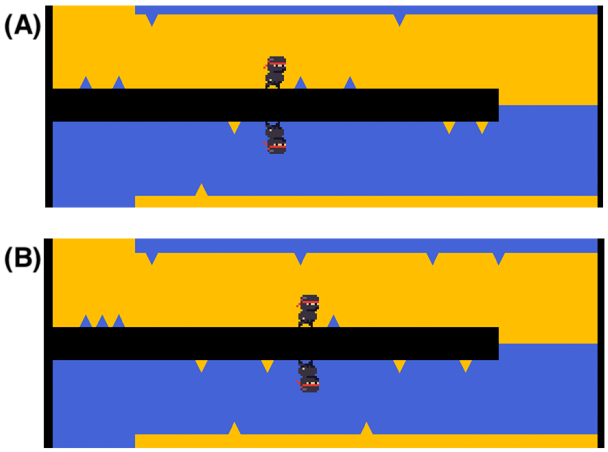

Figure 8. (A) A portion of challenge 3F in version v1, consisting of an S-shaped maze with spikes, is present unmodified in v2. (B) The challenge was modified in v3 with the addition of more spikes on both screens. This untested modification might have led to low scores for v3 by players in Phase 3 as it made the challenge considerably harder.

Our questionnaires of choice, GEQ and PANAS, treat negative affects as undesirable, which may not be in line with the game design. We believe that the addition of a subjective field in these questionnaires so that participants have the chance to explain what the game made them feel would improve upon identification of unwanted frustration.

The final results suggest that observation, interviews, and a questionnaire alone may improve the level design. Even though v2 scored higher than v3, we still advocate the inclusion of in-game metrics and eye-tracking to evaluate attention when design features require it. From the developers perspective, spatial data mixed with eye-tracking data helped to identify usability issues without depending on the observer's or the user's memories.

C. Limitations

This project builds on the premise of affordable and accessible methods, thus excluded biometrics and dedicated eye-trackers. For some developers, however, these limitations are not a problem. Moreover, eye-trackers such as Tobii's[5] may be affordable in some countries.

[5]Technical specifications: https://gaming.tobii.com/product/tobii-eye-tracker-4c/

Players might have felt pressured to finish the game with the knowledge that they were participating in a test. Despite some participants asking for a commercial version of the game, we can not know for sure how many of them would finish the game instead of abandoning it if they were playing in a casual environment.

Many games do not value the principles of "masocore" platformers such as high difficulty, precision, and potentially misleading elements. We can not claim that our finds apply to 2D platform games, or other genres, in general.

Another limitation of our study is that in Phase 2, we enhanced our prototype primarily based on difficulty. We then compared the improved versions using the criteria of fun. We assumed these two criteria were related, but our results indicate they might not be, as we considered version v3 easier than v2 and yet v2 scored higher in Phase 3.

VII. Conclusions and Future work

We evaluated the use of affordable methods in level design of a 2D platform game. Moreover, we found questions intrinsic to the hardcore platformer genre, such as the difficulty of evaluating fun through positive and negative affects. We discussed how frustration could be a desirable element, and because of that, evaluating it using questionnaires such as the GEQ and PANAS might be a challenge. We developed a novel data visualization that combines spatial data with eye-tracking data specific to the game studied in this project.

We expect that our findings might help developers of hardcore 2D games, as well as provide information about the importance of data visualization for game development.

There are many possible directions for future work. First, we intend to expand this research and further investigate multitasking in games through data visualization. Second, investigate how fun and difficulty in 2D hardcore games are related so we can improve our methodology. Third, we noticed further in development that frustration is not always a bad indicator, especially for hardcore platformers. Future research could improve upon it by adding biometric methods, as these might better separate "good" and "bad" frustration. If the budget is still a concern, there are existing affordable webcam-based methods that could be tested [44].

Another possibility for future research is to compare the webcam based solutions with dedicated eye-trackers to determine in which situations each of them is the most appropriate.

Finally, we chose a game that satisfies a particular subset of players. Choosing a simpler platform game or a puzzle game, for instance, could provide other information about using affordable methods.

VIII. Acknowledgment

We would like to thank FUNCAP for funding the author Arthur Silva Bastos through a scholarship.

References

[1] P. Mirza-Babaei, V. Zammitto, J. Niesenhaus, M. Sangin, and L. Nacke, *Games user research: practice, methods, and applications*. ACM, 2013.

[2] M. Rajanen and J. Tapani, "A survey of game usability practices in north american game companies," in *Designing Digitalization (ISD2018 Proceedings)*, 2018.

[3] A. Drachen, P. Mirza-Babaei, and L. E. Nacke, *Games user research*. Oxford University Press, 2018.

[4] P. Mirza-Babaei, N. Moosajee, and B. Drenikow, "Playtesting for indie studios," in *Proceedings of the 20th International Academic Mindtrek Conference*. ACM, 2016, pp. 366–374.

[5] M. S. El-Nasr and S. Yan, "Visual attention in 3d video games," in *Proceedings of the 2006 ACM SIGCHI international conference on Advances in computer entertainment technology*. ACM, 2006, p. 22.

[6] A. Papoutsaki, P. Sangkloy, J. Laskey, N. Daskalova, J. Huang, and J. Hays, "Webgazer: Scalable webcam eye tracking using user interactions," in *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*. AAAI, 2016, pp. 3839–3845.

[7] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba, "Eye tracking for everyone," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2176–2184.

[8] R. Valenti, J. Staiano, N. Sebe, and T. Gevers, "Webcam-based visual gaze estimation," in *International Conference on Image Analysis and Processing*. Springer, 2009, pp. 662–671.

[9] U. Technologies. (2019) Products - unity. Accessed: 2019-09-30. [Online]. Available: https://unity3d.com/unity

[10] M. S. El-Nasr, A. Drachen, and A. Canossa, *Game analytics*. Springer, 2016.

[11] W. IJsselsteijn, Y. De Kort, and K. Poels, "The game experience questionnaire," *Eindhoven: Technische Universiteit Eindhoven*, 2013.

[12] D. Watson, L. A. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect: the panas scales." *Journal of personality and social psychology*, vol. 54, no. 6, p. 1063, 1988.

[13] S. Çiftci, "Trends of serious games research from 2007 to 2017: A bibliometric analysis," *Journal of Education and Training Studies*, vol. 6, no. 2, pp. 18–27, 2018.

[14] L. E. Nacke and C. A. Lindley, "Affective ludology, flow and immersion in a first-person shooter: Measurement of player experience," *arXiv preprint arXiv:1004.0248*, 2010.

[15] A. Canossa, A. Drachen, and J. R. M. Sørensen, "Arrrgghh!!!: blending quantitative and qualitative methods to detect player frustration," in *Proceedings of the 6th international conference on foundations of digital games*. ACM, 2011, pp. 61–68.

[16] J. T. Murray, R. Robinson, M. Mateas, and N. Wardrip-Fruin, "Comparing player responses to choice-based interactive narratives using facial expression analysis," in *International Conference on Interactive Digital Storytelling*. Springer, 2018, pp. 79–92.

[17] A. S. Bastos, R. F. Gomes, C. C. dos Santos, and J. G. R. Maia, "Assessing the experience of immersion in electronic games," in *2017 19th Symposium on Virtual and Augmented Reality (SVR)*. IEEE, 2017, pp. 146–154.

[18] I. Iacovides, J. Aczel, E. Scanlon, and W. Woods, "Making sense of game-play: How can we examine learning and involvement?" *Transactions of the Digital Games Research Association*, vol. 1, no. 1, 2013.

[19] E. Andersen, E. O'Rourke, Y.-E. Liu, R. Snider, J. Lowdermilk, D. Truong, S. Cooper, and Z. Popovic, "The impact of tutorials on games of varying complexity," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2012, pp. 59–68.

[20] M. Bosnjak and T. Orehovacki, "Measuring quality of an indie game developed using unity framework," in *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE, 2018, pp. 1574–1579.

[21] A. E. İlhan, "Eye-tracking to enhance usability: A race game," in *Proceedings of SAI Intelligent Systems Conference*. Springer, 2018, pp. 201–214.

[22] A. R. Gagné, M. Seif El-Nasr, and C. D. Shaw, "Analysis of telemetry data from a real-time strategy game: A case study," *Computers in Entertainment (CIE)*, vol. 10, no. 1, p. 2, 2012.

[23] G. Wallner, N. Halabi, and P. Mirza-Babaei, "Aggregated visualization of playtesting data," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 2019, p. 363.

[24] M. Mozgovoy, "Analyzing user behavior data in a mobile tennis game," in *2018 IEEE Games, Entertainment, Media Conference (GEM)*. IEEE, 2018, pp. 1–9.

[25] L. A. L. Rodrigues and J. D. Brancher, "Improving players' profiles clustering from game data through feature extraction," in *2018 17th Brazilian Symposium on Computer Games and Digital Entertainment (SBGames)*, Oct 2018, pp. 177–186.

[26] E. S. Siqueira, C. D. Castanho, G. N. Rodrigues, and R. P. Jacobi, "A data analysis of player in world of warcraft using game data mining," in *2017 16th Brazilian Symposium on Computer Games and Digital Entertainment (SBGames)*. IEEE, 2017, pp. 1–9.

[27] V. R. Feitosa, J. G. Maia, L. O. Moreira, and G. A. Gomes, "Gamevis: Game data visualization for the web," in *2015 14th Brazilian Symposium on Computer Games and Digital Entertainment (SBGames)*. IEEE, 2015, pp. 70–79.

[28] P. Mirza-Babaei, L. E. Nacke, J. Gregory, N. Collins, and G. Fitzpatrick, "How does it play better?: exploring user testing and biometric storyboards in games user research," in *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 2013, pp. 1499–1508.

[29] M. A. Gómez-Maureira, M. Westerlaken, D. P. Janssen, S. Gualeni, and L. Calvi, "Improving level design through game user research: A comparison of methodologies," *Entertainment Computing*, vol. 5, no. 4, pp. 463–473, 2014.

[30] S. Almeida, Ó. Mealha, and A. Veloso, "Video game scenery analysis with eye tracking," *Entertainment Computing*, vol. 14, pp. 1–13, 2016.

[31] "Genre breakdown of video game sales in the united states in 2017," https://www.statista.com/statistics/189592/breakdown-of-us-video-game-sales-2009-by-genre/, accessed: 2019-05-08.

[32] M. M. Games, "Celeste," 2018. [Online]. Available: http://www.celestegame.com/

[33] M. O'Reilly, "I wanna be the guy," 2007. [Online]. Available: http://kayin.moe/iwbtg/index.php

[34] E. Games, "Chronos twins," 2007. [Online]. Available: http://www.enjoyup.com/

[35] A. W. Ltd, "Binaries," 2016. [Online]. Available: http://binaries.ant-workshop.com/

[36] Philipp Stollenmayer, "Sometimes you die," 2014. [Online]. Available: http://www.kamibox.de/sometimesyoudie

[37] S. Galyonkin. (2019) Steamspy - all the data and stats about steam games. Accessed: 2019-09-30. [Online]. Available: https://steamspy.com/

[38] C. J. Kimberlin. (2019) Unity 2d platformer controller. Accessed: 2019-09-30. [Online]. Available: https://github.com/cjddmut/Unity-2D-Platformer-Controller

[39] R. Koster, *Theory of fun for game design*. " O'Reilly Media, Inc.", 2013.

[40] K. Poels, Y. De Kort, and W. Ijsselsteijn, "It is always a lot of fun!: exploring dimensions of digital game experience using focus group methodology," in *Proceedings of the 2007 conference on Future Play*. ACM, 2007, pp. 83–89.

[41] L. C. Vieira and F. S. C. da Silva, "Assessment of fun in interactive systems: A survey," *Cognitive Systems Research*, vol. 41, pp. 130–143, 2017.

[42] G. A. Gomes, E. Santos, C. A. Vidal, T. L. C. da Silva, and J. A. F. Macedo, "Real-time discovery of hot routes on trajectory data streams using interactive visualization based on gpu," *Computers & Graphics*, vol. 76, pp. 129–141, 2018.

[43] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[44] A. Dingli and A. Giordimaina, "Webcam-based detection of emotional states," *The Visual Computer*, vol. 33, no. 4, pp. 459–469, 2017.