Entertainment Computing 5 (2014) 463–473

Contents lists available at ScienceDirect

Entertainment Computing

journal homepage: ees.elsevier.com/entcom

Improving level design through game user research: A comparison of methodologies ☆

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

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

ARTICLE INFO

Article history:
Received 11 April 2014
Revised 10 August 2014
Accepted 28 August 2014
Available online 16 September 2014

Keywords:
Games user research
Quality assurance
Level design
Game industry
Combined methodologies
Physiological measures

ABSTRACT

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

1. Introduction

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

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

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

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

In this article we look at three methodologies, using three different sources of information, and compare which combinations are most productive in terms of the quality of the changes and the user evaluation of these changes. Through this comparison

☆ This paper has been recommended for acceptance by Haruhiro Katayose.

* Corresponding author.

E-mail addresses: ma.gomezmaureira@gmail.com (M.A. Gómez-Maureira), michellewesterlaken@gmail.com (M. Westerlaken), janssen.d@nhtv.nl (D.P. Janssen), gualeni.s@nhtv.nl (S. Gualeni), calvi.l@nhtv.nl (L. Calvi).

http://dx.doi.org/10.1016/j.entcom.2014.08.008
1875-9521/© 2014 Elsevier B.V. All rights reserved.

we want to illustrate how designers can gather and use GUR data to make informed decisions in their games. To simplify matters, we focus on 2-D level design: This is modular, fast and relatively easy to produce and iterate, and provides a clear basis for comparison among level-sets.

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

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

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

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

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

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

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

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

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

2. Related work

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

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

A recent study looked at the combination of biometrics with a think-aloud protocol [16]. These authors used four types of biometric data (GSR, heart rate, and activity of the facial muscles responsible for smiling and frowning). They concluded that think-aloud protocols and biometric data provided different and mostly independent sources of information. Like us, they found that there were various practical hurdles in combining data from a such a large number of sources with different timing characteristics.

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

[1] Super Mario is a registered trademark of Nintendo.