How psychophysiology can aid the design process of casual games: A tale of stress, facial muscles, and paper beasts

Stefano Gualeni, Dirk Janssen, Licia Calvi
Academy for Digital Entertainment
NHTV Breda University of Applied Sciences
P.O. Box 3917
4800 DX Breda
The Netherlands
{gualeni.s; janssen.d; calvi.l}@nhtv.nl

ABSTRACT
Psychophysiological measurements have so far been used to express player experience quantitatively in game genres such as shooter games and race games. However, these methods have not yet been applied to casual video games. From a development point of view, games developed in the casual sector of the games industry are characterized by very short production cycles which make them ill-suited for complex and lengthy psychophysiological testing regimes.

This paper discusses some methodological innovations that lead to the application of psychophysiological measurements to enhance the design of a commercially released casual game for the Apple iPad, called ‘Gua-Le-Ni’; or, The Horrendous Parade’. The game was tested in different stages of its development to dry-run a cycle of design improvements derived from psychophysiological data. The tests looked at the correlation between stress levels and the contraction of facial muscles with in-game performance in order to establish whether ‘Gua-Le-Ni’ offered the cognitive challenge, the learning curve, and the enjoyment the designers had in mind for this product. In this paper, we discuss the changes that were made to the game and the data-analysis that led to these changes.

Categories and Subject Descriptors
H.5.3 [Information Systems]: Group and Organization Interfaces – evaluation/methodology

General Terms

Keywords
Casual games, Psychophysiological measurements, Game evaluation, User experience evaluation.

1. INTRODUCTION
There is a growing interest in applying psychophysiological (biometric) methods to the design of interfaces and digital experiences in general, so that their design can be adjusted to better align with the user experience of the targeted audience. Biometric technology has already successfully been applied for different media purposes, such as the tuning of video games and the testing of efficacy in advertisement or strategies like pervasive marketing.

Biometric Design for Casual Games (BD4CG) is a research project (supported by the industry and the Dutch government) that aims to develop a biometric-based methodology to specifically evaluate the experience of playing casual computer games. The results of this project should benefit casual game designers by providing them with an additional tool to detect and assess the implications of design choices.

Casual games can fit any canonical game genre and are typically characterized by a short learning curve and simple game mechanics with simple interrelations. Other defining traits of a casual game are the broadly accessible premises and narration and the fact that they do not require long-term commitment, specific skills, or training. For a number of reasons, they are normally low production and distribution when compared to the hardcore games in the same price-range [1].

The number of casual games has expanded and is still expanding at a very fast rate: they have upturned the games market in the last few years. Up to quite recently, the games market used to be dominated by large studios making
AAA games with multi-million dollar budgets. Presently, casual games are sold in larger numbers than AAA titles and they proved to be as profitable or even more profitable than large corporate productions. In the last ten years, an increasingly larger group of people who had never played computer games before have been attracted to various types of casual games, and a certain portion of gamers that were traditionally recognized as pertaining to the hard-core category have migrated to more casual territory [8]. As already mentioned, the revenues followed suit, even though casual games have a per-unit profit which is only a fraction of newly released AAA titles. The sheer size of their user base has compensated for this small per-unit profit.

Biometric observations already proved to be valuable tools capable of providing media content designers and analysts with different indices of the users’ physiological state which can be associated with psychological states (enjoyment, immersion, dislike). By correlating changes in the casual video game players’ physiological variables to their cognitive and emotional states, we intend to answer two fundamental questions:

1. How can accurate physiological and psychological measurements give us a deeper and more objective insight into the player experience?

2. How can we convert psychophysiological measures into actionable results in the game-design process?

In our research, we use four very popular biometric measures: skin conductivity (EDA), facial muscular activity (EMG), heart rate, and blood pressure (BVP). In the experiments we present in this paper, participants played development builds of an action-puzzle video game for the Apple iPad in two different stages of its development, starting from a very early alpha-build. To foreshadow the results of this study: we were able to make informed design decisions based on the results of these biometric experiments. We will further elaborate on the design cycle that was deployed, its benefits and shortcomings, and we will emphasize the changes we made that were specific to casual games development as a result of the analysis of biometric responses of players in the intended target audience for the game (as indicated by the developers).

The paper is structured as follows. In the next section, we will provide a brief review of the research so far on the evaluation of casual games using psychophysiological measurements. Then, we will describe the game that has been evaluated, the experimental protocol, and the results that were collected. In the concluding section, suggestions for further research on psychophysiological measurements in casual games and possible design decisions stemming from research will be provided.

1.1 Psychophysiological Measurements in Casual Games

Any review of recent publications on biometrics and computer games will reveal a marked over-representation of first-person shooter games, racing games, and sport simulation games [5,11,15,16], as well as a paucity or absence of reports on game genres which are typical of the casual sector of the games industry (puzzles, platformers, matching games, hidden object games, etc.). This research tendency is motivated by funding reasons: large video game-development studios that expressed interest in and funded psychophysiological research tend to produce action games in the shooter, race, and sport-simulation genres. This is also due to practical reasons: action games give rise to easily detectable physiological patterns, such as raised heart beat and decreased skin conductivity, because of the increased levels of arousal they aim to generate [5].

However, the findings and methods from this field do not carry over to casual games in a straightforward manner because these games employ different game-mechanics, have different objectives, and evoke different psychological and physiological reactions. Likewise, methods that are viable for large studios developing action games do not apply so well to the small and medium-sized studios that author casual games: casual video game development studios tend to have very short development cycles of approximately two weeks, which do not fit easily with a normal experimental cycle (data harvesting, data analysis and production of design deliverables) that can easily last twice as long.

The absence of studies so far on measuring psychophysiological indices while playing casual video games may be an indicator of the fact that casual games do not provoke as much arousal or excitement as the “high-intensity games” do [5]. Also, the emphasis in casual video games lies not on arousal and action but rather on subtle time pressure, decision-making processes, and incremental discoveries and rewards. These dimensions are all harder to measure with psychophysiological equipment than the mapping from heart beat to arousal in action video games.
1.2 Measurements
Facial EMG (electromyography) is a sensitive and precise method to measure changes in the activation of facial muscles, which are very good indicators of emotional state [6]. It uses small electrodes glued to the skin surface to record the electrical activity of the underlying muscle tissue. Facial EMG can be used to measure both positive and negative emotions [9,15]: positive emotions are often gauged by the activity of the Zygomaticus Major (ZM, involved in laughing) and negative emotions are often gauged by the activity of the Corrugator Supercilii (CM, involved in frowning).

Another dimension can be added by also measuring the skin response, which is widely considered more accurate for arousal [3,18]. This was complemented by heart rate indices, also considered a useful measure for the evaluation of arousal [12,13,17]. Therefore, we added a measure of heart rate by means of a reflective blood volume pressure sensor (BVP). Prior research has been carried out on the effects of game difficulty on heart rate [16,19,20,21], and this has shown that a more difficult game is very likely to cause increased heart rate. In the experiments reported in this paper, we have used heart rate to measure the relative stress level provoked by the casual game. Stress can also be 'good' stress, and constructive stress and people's heart rate can rise for many different reasons. However, we have referred to studies by Cacioppo, by Drachen and by Mandryk [2,5,11] to sustain our claim that an increase in players’ heart rate was causing a higher stress experience while playing the game. In our experiments, such an increase was also correlated to skin conductivity and self-reported stress.

2. EXPERIMENTAL DESIGN
2.1 Participants
We recruited a convenience sample of 24 participants, mostly from the NHTV University of Breda. We balanced the number of men and women and the amount of player experience to be in accordance with the target audience of the game as envisaged by the game developers, which according to their briefing needed to focus mostly on females between 25 and 45 years of age without specific video game experience (although some basic literacy of gaming, for example on social networks or mobile devices, was considered to be a desirable trait).
During the first test, participants were asked to play three games for 12 minutes each. During the second test, our test-subjects were busy with two games, which they needed to play for 20 minutes each. Besides the different builds of ‘Guà-Le-Ni’, the other two games that were tested were a freeware action-puzzle casual video game involving a kite and a simple race game developed by video game development students at NHTV University of Breda. The second test only focused on ‘Guà-Le-Ni’ and the kite game.

The data reported here, however, only focus on the two versions of ‘Guà-Le-Ni’ (i.e., an earlier and the later, improved version) of the iPad-based casual action-puzzle game. This was a comparative study in which three variables were analysed as a possible cause of stress:
1. the initial speed of the in-game beasts,
2. the magnitude of the beasts’ progressive acceleration, and
3. the players’ ease in associating head parts or body parts represented on the cubes with the ones that formed the paper beasts walking across the screen.

During gameplay, participants’ EMG, heart rate and skin response were measured using a Procomp Infinity device coupled to software that was developed in-house. We also recorded the participant's eye movements, but we will not report on those results here. Finally, we used short and long questionnaires (both during and after the game) to collect the participants’ reflection on the game.
The long questionnaire was the Game Experience Questionnaire (GEQ [10,14]), which was developed as a way for players to self-report their playing experience. It is a Likert-scale-based questionnaire analyzing seven features that have been identified as relevant to describe the gameplay. Such features are: 1) Immersion (sensory and imaginative), 2) Flow, 3) Competence, 4) Tension, 5) Challenge, 6) Negative affect and 7) Positive affect. Although the GEQ is not formally overall accepted as the way to self-report somebody’s game experience, it has, however, become a benchmark mainly due to its validation against psychophysiological measurements [13,14].

Two video cameras and high quality audio recording were used to record each gaming session. One camera was used to capture the screen; the second one recorded the participant's facial expressions. At the beginning of each session, participants were briefly interviewed in order to identify possible medical conditions that could interfere with the test, were fully informed of the procedures, and were asked to sign a consent form. After the test, the participants were debriefed and asked to fill out the GEQ.

2.2 The Game
‘Guà-Le-Ni’ [4,7] is a casual action-puzzle videogame released for the iPad in Autumn 2011 (Fig. 1). It requires players to match mock paper beasts that cross the screen by moving and manipulating the ‘taxonomic cubes’, i.e., the modular and three dimensional interface for the game which displays the different modules that can compose the fantastic beasts that walk through the game-space. In ‘Guà-Le-Ni’ a successful match is a sequence of cube faces that correspond in type and order to the parts of the beast currently crossing the screen. If a successful match is achieved before the beast disappears from the screen, points are scored. For the first test, there was no game-over condition. For the second test, the game was over if no
match was made before the beast had crossed the screen (presenting the game-over condition that was planned for the commercial release of the game).

As part of its biometrics-guided tuning and design iterations, the ‘Gua-Le-Ni’ videogame underwent two consecutive sessions of psychophysiological measurements. From a game design perspective, the two separated sets of experiments were aimed at tackling particular aspects of the game which had a specific hierarchy and relevance within the time line of its design iterations. Consequently, in order to be effectively applicable to the design process, dedicated tests needed to be run in specific phases of the design process.

![Figure 1. A screenshot of ‘Gua-Le-Ni; or, The Horrendous Parade’](image)

We wanted to test game difficulty and we did so by altering the speed at which the beasts were crossing the screen and the difficulty in recognizing them through a match with the taxonomic cubes. Much like in [11], we built several game experiences (two in our case: an easy and a more difficult game condition) that we tested on different players. Our goal was to identify the game difficulty level (in terms of speed and ease of beast recognition) that was “just right” [11] for our players. Our assumption was that if the players were not in their “just-right” condition, they would feel overly stressed. Such a relative stress increase was recorded by means of a set of psychophysiological measurements (i.e., heart rate and skin conductivity). The use of GEQ and of post-gameplay interviews served to confirm the biometric results collected.

We carried out two series of tests. The initial tests were aimed at answering the question of the optimal initial speed of the basic matching mechanics. The second tests, building on the results of the first ones, were not only useful as a validation of the first tests but also served to evaluate the relationship between relative stress levels and in-game performance in the progression of the game in terms of speed increase (see point 2 among the identified causes of stress for this specific game).

3. RESULTS
As mentioned earlier, design decisions were based on the biometric data, the in-game and post-game questionnaires using GEQ [10], and the informal interviews with the participants after the test.

3.1 First Test
The question of initial game speed was asked with psychophysiological tools to subjects within the expected target audience of the game: players with no experience with ‘Gua-Le-Ni’ and without prior exposure to tablet devices (these tests were run in Fall 2010). As we wanted to tune this aspect of the game early on, an alpha-version of the game was used.

The focus was on the first few minutes of gameplay as a crucial factor in casual games for their accessibility (in terms of the game being maximally inviting and non-threatening). In this sense, the results of the first test became a cornerstone for all the subsequent tuning decisions. Two parallel tests were run on two slightly different versions of the game: the first had the beast cross the screen at a speed which was decided on by the design team in the prototyping phase (hard version: 24s crossing time). The second version was made significantly slower, and thus easier, in order to assess the difference in the reaction of test subjects (medium version: 30s crossing time). The second version registered less signs of stress on the participants compared to the faster version when analyzing their stress pattern and correlating them with the in-game questionnaires answered by the test subjects during the game.

Even in the slower version of the game, the recorded stress levels were much higher at certain points than at others; that is they indicated that “just-right” condition for the players to simply enjoy the game. We could infer this outcome by combining the biometric data with the self-reported data. Having received and discussed the results, the initial speed of the beta game was set to 34 seconds crossing time by the design team. This was further refined after the second set of tests to the value of 36 seconds, with which the game was released [7].

We analysed the data from two standard skin conductivity sensors, the inter-beat interval derived from a finger photoplethysmographic sensor, and the log-RMS of the ZM (Zygomaticus M.) and CS (Corrugator S.) muscles (measured from electrode pairs over the muscle belly and near the end point) over a 49 sample moving window, using data from five representative participants that showed no artifacts. Our dataset comprised of 140 game periods (delineated by game restarts, speed increases, and participants). We computed average psychophysic signals
and slopes by linear regression on the hundreds of measurement points in that period. The slopes speak to the increase or decrease in psychophysiological signals caused by single events and will be discussed below. The participant averages were computed by removing values above 4 standard deviations and then rescaled to the participant's grand average (ipsatized).

For the interbeat interval (ibi), a regression model was built that included the lagged ibi and the speed (recoded to slow and fast) by beast number (counted since last game start) interaction and main effects. All four effects and the intercept were highly significant (model $R^2$ 0.38, with near-linear residuals in a QQ plot), indicating a robust effect of speed on the ibi (an effect of -40.7, sd 8.7, $p<.001$ for fast over slow speeds), even after taking order effects (lagged ibi) and build-up of stress over time (beast number) into account. The effect of speed levels reduces as the game progresses, as seen from the beast by speed interaction (effect of 9.4, sd 2.8, $p<.002$).

Similar results were obtained for skin conductivity (SC), although the data of one participant had to be removed due to a ceiling effect. The model showed great fit ($R^2$ 0.44), with significant effects of lagged SC, beast number, speed, and beast x speed. Crucially, speed decreased SC (-0.136, $p<.01$) but the effect attenuates over time (speed by beast number 0.05128, $p<.01$). For ZM, lag was removed and all participants were included, but the model needed a participant factor to balance out subject differences. A well-fitting model was obtained (model $R^2$ 0.40) with significant effects including a decrease in ZM activity with speed (effect -0.50, sd 0.17, $p<.01$) and the familiar attenuation of the effect of speed over beast numbers (effect 0.15, $p<.01$). For CS, neither the general model nor the subject-specific model resulted in significant effects other than those for subjects.

The first test also functioned as a check on the comparability of the game elements: different body parts (head, body, tail) of different animals are used in the game and it was important to experimentally equate their difficulty level. For example, the head of the salmon appeared particularly harder to recognize and proved to be a constant cause for spikes in cognitive stress in our test subjects.

The informal interviews held at the end of the game had indeed highlighted that players had found some animal parts (for example, the salmon head) hard to recognize, so that it was difficult for them to reconstruct the puzzle correctly as to match it with the actual crossing beast. But it was only through the combination of biometrics and in-game performance data that we could pinpoint the problematic ones. In this way, we could precisely spot that when certain sides of the cubes had to be used, stress levels were going up more than average. This was indicated by our biometric sensors, specifically by the heart rate which increased when

four specific body parts were appearing and were, therefore, troublesome (like the previously mentioned salmon head).

### 3.2 Second Test

At this stage, we were still interested in perfecting the tuning of the initial speed of the game, but we also wanted to understand how players grew into the game. Specifically, the speed of the beasts will increase stepwise after a certain number of beasts. We wanted to ensure that players developed their skills while the speed stayed constant (decreasing stress levels) and were freshly challenged when the speed went up (increasing stress levels, increased frequency of gameovers).

Clearly, these tests needed to be done on a gameplay-complete version of the game. We chose an early beta-version of the game as a test, so that the experimental results could still be factored into the final development stages.

We found the desired pattern of slowly decreasing stress levels, interspersed by spikes in stress when the speed of the beasts was increased. As shown in Figure 2, this pattern was observed in most players, satisfying the design team in their choices.

![Figure 2: Plot of skin conductivity (SC) for subject 38, showing lower conductivity for slow game speed (blue circles) than fast game speed (red circles). Black dots show SC during pauses; solid vertical lines show game restarts.](image-url)
Additionally, a design goal was to let novice players reach a gameplay duration of 3 to 4 minutes within a reasonable amount of tries and without having to invest too much time. The results of this second batch of tests consistently showed that the development team’s goal was already met (most users reached 3 minutes playtime slightly earlier than expected). In response to this, the development team modified the in-game parameter ‘delta’, which is the multiplication factor applied to the crossing time when it is time to speed up the game. The value of ‘delta’ was originally planned as a constant of 0.7, meaning that at speed-up event (called ‘hunger’ within the game) the crossing time was reduced by 70%. After the biometric tests and interviews, the ‘delta’ was given an initial value of 0.74, with a second parameter ‘metadelta’ which determined a change in delta over time. We arrived at a metadelta of +0.015 and at each speed-up event (hunger), the value of metadelta was added to delta to make the game accelerate slower in general and to also soften the acceleration curve at higher game-speeds. Ultimately, this made the game more manageable both for entry-level players and for the experienced players of the in-house testing team.

4. DISCUSSION
We believe that the pilot set of experiments described in this study led to objective benefits for the developers. Such benefits are traceable in concrete changes made to ‘Gua-Le-Ni’ as a released product (adjustments to the basic design structure, optimization of the tuning of basic game variables, and a higher degree of clarity for the graphical interface). In general, the advantages granted by psychophysiology to the casual game sector can be identified with those outlined by academic applications of biometric observations to action video games and their modes [12]. Among such advantages are the obtaining of large quantities of objective data with a high sample frequency and the fact that such data can be easily compared among test subjects and associated with specific game events [11,12]. In the particular case of the application of biometrics to casual games, specific benefits can be identified in the possibility to explore and experiment with dimensions which are normally not foundational or apparent in action games but can be interpreted as combination of stress and concentration signals. Such dimensions include:

- The initial cognitive weight of the logical task(s) at hand
- The rate of increase of the cognitive weight of the logical task(s) at hand
- The initial pace of the game in terms of speed
- The acceleration of the game pace.

In a general sense, downsides of the application of medical technology to game development purposes are identified as the following:

- The invasive qualities of some of the sensors used, which often limit the player movement or ask the players to control the sharpness and surge of the movements of their hands and heads.
- The extra programming-time required from the developers, as the methodology we propose needs a game-logging system with qualities which are more comprehensive and sophisticated than the ones used for traditional game-metric purposes.
- The time and cost involved in the analysis both in terms of equipment and specialists involved (setting up the framework and the environment for the experiments, harvesting and collating data, the production of design-oriented deliverables for the industry partners/clients, etc.).

However, when applied to casual games, specific downsides also arise. Among them:

- The short iterations that characterize the production cycles of the casual sector of the game industry put a lot of pressure on the researchers to produce results in equally quick time-spans, which runs the risk of making the rigorous application of biometric methodologies impossible (for instance in the case of the sample size of test subjects).
- The current scarcity of academic references and studies involving the application of psychophysiological observations to game genres which are common in the casual sector of our industry, like puzzle video games, action puzzle video games, point-and-click adventure video games, hidden object games, time management video games, etc.

Against this background, the continuation of our research project will specifically focus on overcoming the economical, the academic, and the technical-logistical difficulties that are inherent to pioneering the application of biometrics to casual video games.

5. CONCLUSION
The case of ‘Gua-Le-Ni’ shows the value of using biometric (psychophysiological) testing in addition to more traditional quality-assurance methods, including playtesting, interviews, player metrics, and questionnaires. It was the combination of methods that allowed the design team to come up with the optimal value for the initial speed of the game and a player-friendly pattern of changes in speed over time. Likewise, the identity of those animal parts that were hard to match was easily established in the biometric stress measurements. Common design goals, such as playability within a certain number of tries or amount of time, could also have been measured using in-game metrics only. However, the addition of biometric data meant that the designers were alerted to the fact that the
participants were reaching this goal, but that it was pleasurably challenging for them.

A number of hurdles had to be overcome, and not all challenges of applying biometric methods to casual games have been solved. Yet, biometric hardware is becoming rapidly more affordable and with the measurement and interpretative frameworks that are being developed by our BD4CG project, we hope to make biometric testing available to any casual video game developers as an additional tool to increase and validate the quality and effectiveness of their game-design choices.

6. ACKNOWLEDGMENTS

The authors wish to thank Maartje Polman, Niki Bosman, Hans Bouwknegt, Marcello Gómez Maureira and all the testers who took part in our experiments; the Raak International Funding body and NHTV for their support in this research.

7. REFERENCES