<http://www.diva-portal.org>

Postprint

This is the accepted version of a paper published in *Entertainment Computing*. This paper has been peer-reviewed but does not include the final publisher proof-corrections or journal pagination.

Citation for the original published paper (version of record):

Bevilacqua, F., Engström, H., Backlund, P. (2018)

Changes in heart rate and facial actions during a gaming session with provoked boredom and stress.

Entertainment Computing, 24: 10-20

<https://doi.org/10.1016/j.entcom.2017.10.004>

Access to the published version may require subscription.

N.B. When citing this work, cite the original published paper.

Permanent link to this version:

<http://urn.kb.se/resolve?urn=urn:nbn:se:his:diva-14267>

Changes in heart rate and facial actions during a gaming session with provoked boredom and stress[☆]

Fernando Bevilacqua^{a,b,*}, Henrik Engström^a, Per Backlund^a

^a*University of Skövde, Skövde, Sweden*

^b*Federal University of Fronteira Sul, Chapecó, Brazil*

Abstract

This paper presents an experiment aimed at exploring the relation between facial actions (FA), heart rate (HR) and emotional states, particularly stress and boredom, during the interaction with games. Subjects played three custom-made games with a linear and constant progression from a boring to a stressful state, without pre-defined levels, modes or stopping conditions. Such configuration gives our experiment a novel approach for the exploration of FA and HR regarding their connection to emotional states, since we can categorize information according to the induced (and theoretically known) emotional states on a user level. The HR data was divided into segments, whose HR mean was calculated and compared in periods (boring/stressful part of the games). Additionally the 6 hours of recordings were manually analyzed and FA were annotated and categorized in the same periods. Findings show that variations of HR and FA on a group and on an individual level are different when comparing boring and stressful parts of the gaming sessions. This paper contributes information regarding variations of HR and FA in the context of games, which can potentially be used as input candidates to create user-tailored models for emotion detection with game-based emotion elicitation sources.

Keywords: Games, Boredom, Stress, Facial Expression, Multifactorial, Heart rate

[☆]This article is an extension of previous work [1]. ©2017. This manuscript version is made available under the CC-BY-NC-ND 4.0 license <http://creativecommons.org/licenses/by-nc-nd/4.0/>. DOI: <https://doi.org/10.1016/j.entcom.2017.10.004>

*Corresponding author. Address: Fernando Bevilacqua, University of Skövde, P.O Box 408, SE-541 28, Skövde, Sweden.

Email addresses: fernando.bevilacqua@his.se (Fernando Bevilacqua), henrik.engstrom@his.se

1. Introduction

A general definition of emotions is that they are biologically based action dispositions that have an important role in the determination of behavior [2]. Emotions are multimodal responses to a situation causing changes in expressive behavior [3], e.g. facial activity, and physiological activity [4], e.g. heart rate (HR). Methods for processing such changes to recognize user emotions have been proposed in the domain of human-computer interaction [5], affective computing [6] and game research [7], often involving the mapping of signals into emotional states using machine learning models [8].

In the context of games, a variety of physiological signals have been used to automatically assess different emotional states [9, 10, 11]. Significant differences in HR, for instance, are reported at stressful periods of gameplay [12, 13, 14], including different levels of game difficulty in *Tetris* [15] and fast/slow conditions in an adapted version of *Pacman* [16]. Additionally the analysis of facial behavior shows that increased activity of the zygomatic [16] and the corrugator [17] muscles, associated with smiling and frowning respectively, are more frequent during particularly emotional game events. The variations of those psychophysiological signals have the potential to be used as sources for emotion detection.

However, while previous work explored the use of games as elicitation sources for recognizing user emotions, relying on the emotional states a person can experience [18] and which physiological signals are better predictors of such states [19], they lack a more user-tailored approach for studying the variations of signals. Emotional states such as stress and boredom are often inducted by administering a game with the same particular setup, e.g. high/low difficulty, to all subjects. People respond differently to media according to their personality [20], and they differ in social, learning and play styles [21]. A game session labeled as stressful, for instance, assumes that all subjects have the same expectations and behave similarly, which dilutes the individuality of each person as some might experience the interaction as not being stressful as intended. Additionally the analysis usually involves the interaction of subjects with some game levels (from the same game) featuring a constant difficulty scale,

(Henrik Engström), `per.backlund@his.se` (Per Backlund)

which does not contemplate the variations of signals in a context where the game difficulty is constantly increasing in the same game level/session.

In this paper, we propose a different approach to explore the relation between facial activity, HR and emotional states, particularly stress and boredom, during the interaction with games. We designed and carried out an experiment involving games as emotion elicitation sources, which were deliberately designed to cause the aforementioned emotional states in a novel configuration. Our approach consists of recording participants while they play three different games that were carefully designed and developed to have a difficulty level that constantly and linearly progresses over time without a pre-defined stopping point. At the beginning the games are highly predictive, without novelties, changes or surprises and with emphasis on the passage of time during a wait, which leads to an emotional state of boredom [22, 23, 24]. The game difficulty is then periodically increased until the subject is not able to cope with the challenges at hand, which happens at different times for different subjects. The ever-growing game difficulty leads to an emotional state of stress towards the end of the interaction.

The purpose of our experiment is to investigate how responses related to physiological activity, i.e. HR, and facial actions (FA), defined by us as being any facial movement different from a neutral face, e.g. lips contraction, relate to emotional states in a game context featuring constant changes in difficulty. As a result, we present an analysis regarding the changes in the HR mean and annotated FA that happened during the phases of the games that were perceived as being boring and stressful. Our main contribution is twofold: firstly we introduce a different structure for emotion elicitation in our games, which account for personal differences among subjects when inducing an emotional state of stress. Secondly we present information, on group and individual level, about the variations of HR and naked-eye recognizable FA that happened during the interactions with such different game structure, especially under situations that were designed to provoke boredom and stress. The aim of this paper is to provide information regarding the variations of HR and FA when compared in boring and stressful moments of a game, which can potentially be used as input candidates to create user-tailored models for emotion detection when games are

used as emotion elicitation sources. Additionally we highlight the heterogeneous nature of our group of subjects that have significantly different ages and gaming experiences, which produces a diverse sample for the investigation of changes in HR, FA and emotional states.

2. Related Work

2.1. Games and emotional states

A game is defined as a system in which players engage in an artificial conflict, defined by rules, that results in a goal [25]. The difficulty level of the challenge affects the emotional state of players, e.g. moments of boredom or anxiety/stress [24]. A challenge beyond the player's skill to address and overcome it causes anxiety, while the opposite results in disinterest, leading to boredom [26]. An ideal challenge/skill balance produces an optimal experience and concentration state called flow [27], which is vastly connected to engagement/immersion [28] and sense of presence [29]. Game design may also be described as the effort to induce states of flow and presence [30, 31].

2.2. Heart rate, stress and frustration

Physiological signals are considered reliable sources of information since they are hard to fake (because of their link to the autonomic nervous system), differently from facial expressions, for instance [32]. The use of physiological signals, such as HR, have been demonstrated as player input for games [33], indication of perceived interest and confusion in mobile applications [34], triangulation of psychophysiological emotional reactions to digital media stimuli [35], and as measurement of frustration in a game [13]. Those approaches rely on physiological arousal [36] and its connection to emotion regulation [37, 38]. Vandeput et al. [39] and Garde et al. [40] demonstrate that higher HR and differences in heart rate variability (HRV) are present in mentally demanding tasks when compared to a rest period. Similarly Bousefsaf et al. [10] show that the stress state curve tends and HR to decrease during the rest period and increase during stress sessions, in accordance with previously mentioned works. The standard deviation of HRV has also been reported as significant in the monitoring of arousal, however low and high frequencies of HRV present less weight in

the monitoring compared to other signals, e.g. skin conductance [41]. Finally McDuff et al. [42, 43] also use HR and its variants in order to measure cognitive stress during computer tasks. According to the authors, the average HR and breathing rate are not significantly different in any case, which differs from the findings of previously mentioned work.

2.3. Facial features and emotional state

The main manifestation of anxiety in the human face involves activities related to eyes (e.g. blinking), mouth (e.g. lip deformation), cheeks, and head movements [44]. The Facial Action Coding System (FACS) [45, 46] is a widely used taxonomy for characterizing facial activity, however facial analysis based on Euclidean distances of points has also been reported as successful [47]. Emotion detection based on extracted facial features are often performed by machine learning models [48, 49] that achieve significantly different results, which highlights the complexity of correlating facial features and emotions. Heylen et al. [50] report the occurrence of a variety of expressions in a pilot experiment, but most of the time subjects remain with a neutral facial expression. Grafsgaard et al. [51] presents brow lowering as a positive predictor of student frustration in the context of a tutoring system, but the same facial activity has been correlated with confusion by previous work. Bailenson et al. [52] show that a machine learning model for emotion detection built from a combination of facial and physiological information is more efficient than a model built with either one alone. Additionally a person-specific model outperforms a model trained with data from all subjects, suggesting that a user-tailored model might be more effective in identifying features (even the more subtle ones) than a general-purpose model.

2.4. Summary

Some of the previously mentioned works rely on subjects performing tasks on a computer, e.g. Vandeput et al. [39], Garde et al. [40], or interacting with gamified cognitive tests, e.g. McDuff et al. [43], Bousefsaf et al. [10], Grundlehner et al. [41], to study the relation between signals and emotional states. Those are game-like emotion elicitation sources that are less likely to evoke the same emotions a subject has when interacting with a game,

which has decisions and consequences that might produce a deeper emotional involvement. When games (as defined in Section 2.1) are used, they are often commercial off-the-shelf (COTS) games whose difficulty level has been designed to serve and entertain a broad audience, not to induce boredom or stress. Sharma et al. [12], for instance, used three unchanged COTS games in an assessment of computer games as a psychological stressor, which relied on subjects gaming skills to perceive a game as stressful. Similarly Ravaja et al. [14] used an unchanged COTS game to measure phasic emotional responses. Chanel et al. [15] used an unchanged version of Tetris whose level of difficulty for each subject was selected based on repeated interactions with the game followed by analysis of self-reported levels of engagement. It familiarizes subjects with the game and prepare them for the chosen difficulty level. When COTS games are adapted to the needs of researchers, e.g. increased speed in *Pacman* to cause stress [16], subjects interact with pre-defined game modes (e.g. slow, fast and normal mode), which assumes that all players behave similarly and have the same expectations. Finally Rodriguez et al. [13] used a custom-made game designed to induce frustration in the subjects, however the experiment aim was to evaluate emotional regulation strategies in adolescents, not variations of psychophysiological signals in a context involving games as emotion elicitation sources. As opposed to previously mentioned works, our approach consists of using a induced boring to stressful mechanics in games to produce variations in the emotional state of participants. Our experiment uses custom-made games with a linear and constant progression from a boring to a stressful state, without pre-defined levels, game modes or stopping conditions. We believe such configuration gives our experiment a novel approach for the exploration of facial actions and HR regarding their connection to emotional states, since we can categorize information according to the induced (and theoretically known) emotional states on a user level. To the best of our knowledge, this is the first experiment where games with linear boring-to-stressful progression are used to deliberately induce emotional reactions.

3. Experiment Setup

Twenty adult participants of both genders (10 female) with different ages (22 to 59, mean 35.4, SD 10.79) and different gaming experience gave their informed and written consent to participate in the experiment. The study population consisted of staff members and students of the University of Skövde, as well as citizens of the community/city. When asked how skilled subjects believe they are at playing video games, 1 subject (5%) reported no skill, 10 (50%) reported not very skilled, 7 (35%) reported moderately skilled and 2 (10%) reported very skilled. When asked the number of hours per week they had played any type of video game over the last year, 2 subjects (10%) reported more than 10, 6 (30%) reported 5 to 10, 2 (10%) reported 3 to 4, 2 (10%) reported 1 to 3, 4 (20%) reported 0 to 1, and 4 (20%) reported no activity. Those numbers indicate that our population has a diversity of gaming experience and playing frequency, which provides us with information that is less skewed towards specific profile of players, e.g. hardcore players. Subjects were seated in front a computer, alone in the room, while being recorded by a camera and measured by a heart rate sensor. The camera was attached to a tripod placed in front of the subjects at approximately 0.6m of distance; the camera was slightly tilted up. A spotlight, tilted 45° up, placed at a distance of 1.6m from the subject and 45cm higher than the camera level, was used for illumination; no other light source was active during the experiment. The experiment setup is based on similar emotion assessment experiments conducted by previous work [53, 43, 15, 10]. Figure 1 illustrates the setup.

The participants were each recorded for about 25 minutes, during which they played three games (detailed in section 3.1). Each game was followed by a computerized questionnaire related to the game and to stress/boredom. The first two games were followed by a 140 seconds rest period, where the subjects listened to calm classic music. The last game was followed by an additional questionnaire about age and gaming experience/profile. The order in which the games were played was randomized among subjects. Before the start, participants received instructions from a researcher that they should play three games, answer a questionnaire after each game and rest. They were told that their gaming performance

Figure 1: Experiment setup. Subjects were recorded while alone in the room during the gaming session

was not being analyzed, that they should not give up in the middle of the games and that they should remain seated during the whole process. The transition among games, resting periods and questionnaires was completely automated by software.

3.1. Games and stimuli elicitation

The three games¹ used in the experiment were 2D and casual-themed, played with mouse or keyboard in a web browser. The games were carefully designed to provoke boredom at the beginning and stress at the end, with a linear progression between the two states (adjustments of such progression are performed every 1 minute). The game mechanics were chosen based on the capacity to fulfill such linear progression, along with the quality of not allowing the player to instantly kill the main character (by mistake or not), e.g. by falling

¹Source code available at: <https://github.com/Dovyski/face-tracking-games>

into a hole. The mechanics were also designed/selected to ensure that all subjects would have the same game pace, e.g. a player must not be able to deliberately control the game speed based on his/her will or skill level. We used well-established genres for the games, but tailored them to meet the previously mentioned requirements, e.g. controllable pace, linear challenge progressions from extremely easy to extremely hard. We assumed subjects would be more familiarized with such genres. The game Tetris was included under the assumption that it was very well known by subjects. We did not rely on COTS games because their difficulty level is usually designed to serve and entertain a broad audience, not to induce boredom or stress. Additionally they are less likely to present mechanics that meet our requirements, such as a controllable pace, prevent players from instantly killing the main character, and a linear challenge progression from a boring to a stressful state, which are the foundation of our games as emotion elicitation.

The **Mushroom** game, illustrated in Figure 2 (left), is a puzzle where the player must feed a character by dragging and dropping mushrooms in rounds. In a given round, M mushrooms are displayed in a grid and the player has K seconds (a decreasing time bar at the top informs the remaining time) to collect good and discard bad (poisonous) mushrooms. At the upper-right corner of the screen, a sign informs the player about the bad/poisonous mushroom of the round. The player must drag and drop all good mushrooms (the ones different from the poisonous indication) into the character, while dragging and dropping the bad ones into the trash can. Mushrooms are differentiated by the colors of their features (circles). The player is rewarded with score points, a health bar increase (HB_I) and a pleasant sound when a right move is performed. In case of mistake a health bar decrease (HB_D) and an annoying/aggressive alarm sound is applied. If the time K is over and the player has not finished moving all the mushroom of the round, each remaining mushroom in the grid is counted as a mistake. If the grid is clean and there is still time available, the player must wait until the time is over. The values of M , K , HB_I and HB_D are used to induce boredom/stress. At the beginning, M is low (starts with 2) and K is high (starts with 45 seconds), so the player spends a significant amount of time waiting for the game to continue; every 1 minute the value of M is increased and K is decreased. The changes

continue to happen until the player is unable to deal with the amount of mushrooms within the available time. This leads to mistakes that will eventually decrease the health bar to zero, terminating the game. After the mark of 6 minutes, the game becomes virtually impossible to beat.

The **Platformer**, illustrated in Figure 2 (center), is a side-scrolling, endless runner game where the player must control the main character while collecting hearts and avoiding obstacles (skulls with spikes). The character can jump (by pressing the up arrow key in the keyboard) or slash (S key), however the player is not able to move the main character left or right, it remains in the same position on the screen (towards the left side of the screen). The character moves on top of platforms, which are always perfectly connected, so there are no gaps (holes) among them; the height of the platform can vary, however, so there might be a slope up/down connecting two platforms, for instance. If the character hits an obstacle, the health bar is decreased (HB_D) and a sound effect related to pain is played. If any heart is collected, the health bar increases (HB_I) and a pleasant sound effect is played. The position where the hearts appear on each platform is adjustable (defined by HH), so they can appear close to the platform (no action is required to collect the heart) or a bit higher from the ground (jump action is required to collect the heart). The speed of the character (S , which is the velocity at which elements are moving on the screen), the height variation of each new platform that appears on the screen (HV), the amount of hearts (G) and obstacles (E) per platform are all controlled by the game and used to adjust boredom/stress. At the beginning, boredom is induced by keeping all previously mentioned parameters with low values, which means the game is slow, the character moves from platform to platform at the same height and almost no hearts or obstacles appear on the screen. The few hearts that are available are placed close to the ground to destimulate jumping actions. As time progresses, the values of S , E , HV , HB_D and HH increase, while G and HB_I decrease to induce the player to a stressfull state. At the mark of 5 minutes, for instance, the game is significantly fast, with several obstacles on the screen and almost no hearts to collect; the damage caused to the character when hit by an obstacle is also higher than the beginning of the game. The linear increase in difficulty will eventually result in consecutive hits (mistakes), which will

Figure 2: Mushroom (left), Platformer (center) and Tetris (right). In Mushroom, player has to drag and drop the correct mushrooms into the character, discarding the wrong ones into the trash. In Platformer, the player has to jump over or slide under obstacles while collecting hearts. In our version of Tetris, there are no hints about the next piece to be added to the screen

decrease the health points until zero, when the game ends.

Finally the game **Tetris**, shown in Figure 2 (right), is a modification of the original Tetris game. In our version of the game, the next block to be added to the screen is not displayed, so the player is unable to predict future moves. Additionally, the down key, usually used to speed up the descendant trajectory of the current piece, is disabled. The keyboard controls are the arrow keys to move the piece left/right and the R key to rotate the piece. The game is also modified to ensure that all subjects received the same sequence of pieces (we use the same seed for the generation of random numbers). The speed that the pieces fall (S) is used to control boredom and stress; at the beginning of the game, boredom is induced by using a low value for S , which makes the game slow since the pieces are falling slowly and the player is unable to speed them up. As time progresses, S increases linearly making the game faster and harder to play, which should induce stress. At the mark of 5 minutes, for instance, a single piece takes almost 1 second to traverse the whole screen.

3.2. Data collection

During the whole experiment, subjects were recorded using a Canon Legria HF R606 video camera. At the same time, their HR was measured by a TomTom Runner Cardio watch (TomTom International BV, Amsterdam, Netherlands), which was placed on the left

arm, approximately 7cm away from the wrist, like a regular wrist watch. The usage of the watch was unobtrusive, so it did not affect the movements of the subjects, who could still use both hands to play the games. The watch recorded the HR at 1 Hz.

3.3. Questionnaires

After each game, subjects answered a questionnaire in order to provide self-reported stress and boredom measurements. The questionnaire had six questions: the first four were a 5-point Likert scale related to how the player felt related to stress/boredom at the beginning/end of each game (1: not stressed/bored at all, 5: extremely stressed/bored); a question to identify the part of the game that best describes the moment the subject enjoyed the most (very beginning, after beginning and before middle, middle, after middle and before end, very end); finally a question asking if the subject understood the game. Before the end of the experiment, subjects answered a final questionnaire with nine questions, which were related to: age; gender; number of hours per week spent with games over the last year (question from the video game experience questionnaire [54]); how proficient or skilled the subject believe themselves to be at playing video games (question from the Survey of Spatial Representation and Activities - SSRA [55]); familiarity with puzzle, platform and Tetris games; current state of mind compared to other days (e.g. normal, unusually stressed, etc.); and gaming profile (like, dislike challenging games).

4. Analysis

The following subsections present the methodology we used to analyse the data regarding HR, FA and questionnaire answers. Based on previous work regarding variations of HR and FA in the context of emotions, we expect that both will vary and be different when comparing the beginning (boring part) and the end (stressful part) of the games. Such difference, if confirmed, shows the potential of HR and FA as input candidates for emotion detection models founded on game-based emotion elicitation. Subject 9 had problems to play the Platformer game, so data from this subject, in this particular game, was not used in any analysis nor results.

4.1. Heart rate

Firstly we removed from the set of all HR readings obtained during the experiment values that were equal to zero assuming they were miss-readings. After we calculated the baseline HR value for each subject (B_s) as:

$$B_s = \frac{1}{2}(\overline{HR}_{r1,s} + \overline{HR}_{r2,s}) \quad (1)$$

where s indicates the subject and $\overline{HR}_{r1,s}$, $\overline{HR}_{r2,s}$ are the mean HR during the first and the second resting period (for subject s), respectively. B_s is assumed to be the “expected” HR of a given subject while resting. The average difference between $\overline{HR}_{r1,s}$ and $\overline{HR}_{r2,s}$ for each subject was 2.34 bpm.

We then calculated the HR mean coefficient $C_s^{g,t}$, which is the HR mean of a subject while playing a game during a given period of 60 seconds:

$$C_s^{g,t} = \frac{1}{60} \sum_{n=1}^{60} HR_{s,g}(t \cdot 60 + n) \quad (2)$$

where s is the subject, g is the game being played (M for Mushroom, P for Platformer or T for Tetris), t is the period and $HR_{s,g}(k)$ is the HR measured from subject s , in game g at the mark of k seconds. Since each subject played each game for more than 60 seconds, there is more than one period for each subject for a given game. The t component of $C_s^{g,t}$ specifies which of such periods the HR mean refers to. For instance, $t = 0$ comprehends the period from time 0:00 until time 1:00 of a given game, $t = 1$ is the period from time 1:01 until 2:00, and so on. As an example, the HR mean coefficient $C_2^{P,1}$ is the HR mean of subject 2 while playing the Platformer game from time 1:01 to 2:00.

HR values are specific to each individual, so we calculated the relativized HR mean coefficient, $V_s^{g,t}$, by subtracting $C_s^{g,t}$ from B_s as:

$$V_s^{g,t} = C_s^{g,t} - B_s \quad (3)$$

$V_s^{g,t}$ accounts for values that are related to changes instead of absolute HR measurements, which are significantly more suitable for comparison among different subjects, or within the

same subject.

Based on previous work regarding variations of HR and emotions, we expect differences in the HR mean between the last and the second minutes of gameplay. The reason why we choose $t = 1$ (second minute of gameplay) instead of $t = 0$ (first minute of gameplay) for our analysis is because we believe the first minute of the game might not be ideal for a fair comparison. Firstly during the first minute of gameplay, subjects are less likely to be in their usual neutral emotional state. They are more likely to be stimulated by the excitement of the initial contact with a game soon to be played, which interferes with any feelings of boredom. Secondly subjects need a basic understanding and experimentation with the game in order to judge if it is boring or not. As per our understanding, such conjecture is less likely to be fulfilled during the first minute of gameplay than it is to be during the second minute of gameplay.

4.2. Facial actions

The recordings of all subjects were analyzed by the main author who took notes of any FA that were different from a neutral (resting) face, e.g. lips contraction, brow movement, etc. Annotations were not performed periodically, e.g. every 5 seconds, instead they were made only when the subject's face changed from its neutral/resting state; as a consequence, if the subject remained with a neutral face for a long period of time, no annotations were made during that period.

We decided to use an empirical and non-standard approach for facial annotation as we are not interested in facial expressions *per se*. We want to explore any facial action patterns (standardized or not) that might be used as inputs for emotion detection models to infer boredom/stressful states. This approach is not without its limitations, however it provides a reasonable empirical perception of facial activity that is different from a neutral face, which is satisfactory for our investigation. We believe FA are subtle and not necessarily part of a complete facial expression, e.g. surprise face, so they might be better identified in a context where annotations are made only when facial changes happen, as opposed to a frame-by-frame analysis/annotation of a video, for instance.

Figure 3: Annotated facial actions (FA). (a) Smile not showing teeth; (b) Smile showing teeth; (c) Lip pucker; (d) Lip stretch; (e) Lip suck; (f) Lip press; (g) Lips parted; (h) Tongue touching lips; (i) Mouth movement right; (j) Mouth movement left; (k) Lower lip bite; (l) Frown; (m) Brow raiser; (n) Lid tightener; (o) Brow lowering

According to our design of the games, subjects are supposed to perceive the experience during the beginning of the games as being more boring than the one at the end, while the experience at the end should be perceived as more stressing than the one at the beginning of the games. As a result, if we divide in half the game sessions of each subject, in theory one of the two resulting parts is more likely to be perceived as more boring by the subjects,

while the other is more likely to be perceived as more stressful. Using that assumption, FA annotations were divided in two groups, the ones made during the period that corresponds to the first half (H_0) of the games and the ones made in the second half (H_1). Our division of the annotation aimed to identify any pattern regarding FA happening during periods theoretically perceived as boring or stressful. After all annotations were made, we identified which ones were unique and, based on that information, we counted the repetitions of such unique actions across the games for all subjects. As a result, we obtained the frequency that each FA appeared during all game sessions, as well as when they happened (in H_0 or H_1). We excluded from the list any FA that appeared just a single time during the whole 6 hours recording, assuming that such action was noise or probably part of another action. As a result we identified 17 unique FA that appeared in the recordings at least twice. Excluding the talking and laughing FA, Figure 3 illustrates all our annotated FA. Finally after all annotations were counted and categorized according to the period in the game, we conducted a per-subject evaluation regarding the frequency of FA. For each subject, we inspected which FA appeared in H_1 of all three games with a higher frequency than in H_0 , and vice versa (appeared in H_0 of all three games with a higher frequency than in H_1).

4.3. Self-reported questionnaire answers

Regarding the self-reported levels of stress and boredom provided by subjects after each game, the answers were given according to the participant’s own interpretation of such levels. As a consequence we are not able to treat the answers as a uniform scale with a defined degree of difference among values. Because of that we decided to use a Wilcoxon Signed-rank test to statistically check if the reported boredom levels at the end are significantly different from the ones at the beginning of the games, as well as if the reported stress levels at the end are different from the ones at the beginning of the games.

5. Experimental Results

5.1. Self-reported questionnaire

We analyzed the difference between the reported boredom and stress levels at the beginning and at the end of each game, according to the procedures described in Section 4. For the Mushroom game, the reported boredom levels at the beginning (median 3.5) were significantly higher than the reported levels at the end (median 1) of the game, $Z = -2.69$, $p < 0.01$. Regarding the reported stress levels, values at the beginning (median 1) were significantly lower than the reported levels at the end (median 3), $Z = 3.63$, $p < 0.01$. For the Platformer game, boredom levels at the beginning (median 3) were higher than the ones at the end (median 1), $Z = -2.47$, $p < 0.05$. Regarding stress levels, values at the beginning (median 1) were lower than the ones at the end (median 4), $Z = 3.79$, $p < 0.01$. Finally for the Tetris game, boredom levels at the beginning (median 4) were higher than the ones at the end (median 2), $Z = -2.97$, $p < 0.01$. Regarding stress levels, values at the beginning (median 1) were lower than the ones at the end (median 4), $Z = 3.95$, $p < 0.01$. The self-reported answers support the idea that subjects perceived the three games as being boring at the begin and stressful at the end, which was the intended result of our design process.

5.2. Heart Rate

Table 1 presents the values of V , the relativized HR mean coefficient, for all subjects in all games, grouped by intervals of 1 minute, calculated according to the description in Section 4. Column g is the game being played, t is the period in the game and s is the subject. Since all games were constantly changing in difficulty and subjects have different gaming skills, there are subjects with no data entry for some t intervals, which means he/she was defeated by the game earlier than other subjects were. Subject 9 had problems playing the Platformer game, so data for that subject in that game was not used in the calculations.

A positive value in Table 1 represents a V (HR mean) that is above the subject's baseline B_s (mean HR while resting) for an specific period t . A negative value indicates that V in

Table 1: Values of $V_s^{g,t}$, the relativized HR mean coefficient, for all subjects (s) in a given game g (M is for Mushroom, P for Platformer and T for Tetris), grouped by intervals (t) of 1 minute

g	t	s																			
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
M	0	-3.8	2.2	-2.5	-3.1	-3.4	-2.5	0.4	3.5	-4.9	-2.8	-3.4	-0.2	-1.8	5.9	-4.9	6.5	-0.4	-3.3	4.4	2.1
	1	-9.1	-1.4	0.3	-2.6	0.2	-3.7	-1.5	2.7	-4.9	-4.1	-10.3	0.0	-3.1	-3.1	0.2	5.6	2.5	-0.8	0.5	3.2
	2	-4.8	-1.3	-0.1	-0.6	7.0	0.9	2.0	4.5	-0.8	-3.0	-9.2	4.1	0.1	-0.5	-0.1	5.4	2.8	2.4	2.6	4.8
	3	-4.9	-0.7	-2.8	-1.5	1.5	0.3	2.4	5.1	-2.4	1.7	-4.6	2.4	0.4	-0.2	0.7	4.5	3.4	3.8	2.4	3.5
	4	-3.9	-1.1	0.9	1.5	5.3	0.8	4.5	6.3	2.0	1.3	-3.6	3.3	1.6	6.9	0.0	4.5	2.9	3.4	9.9	9.1
	5	0.3	2.4	1.4	1.7	11.9	-1.2	-	-	4.4	10.2	1.6	6.2	-	-	3.2	-	5.9	3.2	17.9	5.8
	6	-	-	-1.2	-0.1	-	-	-	-	-	-	-	-	-	-	-	-	3.3	-	-	6.7
P	0	-1.7	1.3	0.4	-0.2	0.1	9.9	2.4	-1.7	- ^a	-1.9	-2.7	0.8	5.7	19.7	6.0	0.2	5.9	-1.2	4.2	5.5
	1	-1.6	-0.4	3.4	-0.9	-0.3	2.2	2.4	5.5	- ^a	2.4	-4.4	0.7	-1.5	4.1	4.9	-1.0	0.8	-0.2	3.7	7.7
	2	1.9	9.7	0.8	-0.6	3.0	-1.1	1.3	3.9	- ^a	15.1	0.2	3.5	3.9	3.8	11.7	-0.7	2.8	0.4	4.0	10.6
	3	3.0	9.3	2.5	-2.6	2.8	10.3	4.9	5.2	- ^a	21.6	3.2	5.4	9.2	4.6	9.9	-	2.1	2.6	7.9	10.4
	4	5.9	6.8	-	8.0	5.3	-	-	-	- ^a	-	-	-	4.9	-	13.5	-	-	-	-	-
T	0	2.1	6.5	-2.1	-1.3	-4.0	5.7	3.4	4.2	8.3	3.2	2.1	-0.1	3.5	3.4	4.4	-1.2	7.8	-3.9	5.8	4.7
	1	-2.7	0.0	-3.3	-1.2	-4.9	-0.1	4.3	4.2	2.7	2.9	0.0	2.6	2.2	-2.5	5.9	-1.3	4.2	-0.4	5.7	0.0
	2	-1.7	2.6	2.4	-0.1	-2.3	4.3	3.5	-0.4	2.7	5.1	2.6	5.9	1.1	-1.1	5.3	-1.8	7.4	0.1	8.1	4.3
	3	-1.9	-0.2	0.3	-2.2	0.8	5.4	2.1	3.8	-	5.2	2.2	5.4	-0.5	-2.5	4.7	-1.2	10.6	1.5	3.8	2.3
	4	-0.8	3.0	-	0.9	-	-	-	7.8	-	9.2	0.2	6.6	-	3.4	5.6	-1.2	-	2.0	6.8	4.3
	5	1.5	7.4	-	-0.4	-	-	-	-	-	12.9	7.4	4.5	-	3.5	6.7	-	-	-	6.9	-

^a Subject 9 had problems playing the Platformer game, so data from this subject during this game was excluded.

that period is below the subject's baseline B_s . Assuming n as the last minute of gameplay of a given subject in a game, by comparing the values at $t = 0$ (first minute of the gameplay, perceived as boring) and $t = n$ (last minute of gameplay, perceived as stressful) in the Mushroom game, 19 subjects (95%) presented $V^{M,n}$ greater than $V^{M,0}$. The same comparison regarding the Platformer game indicates that 16 subjects (84.2%) had higher $V^{P,n}$ than $V^{P,0}$. In the Tetris game 13 subjects (65%) presented higher $V^{T,n}$ than $V^{T,0}$.

As previously mentioned, our expectation is that the true difference in means between V at the last minute of gameplay ($t = n$) and at the second minute of gameplay ($t = 1$) is less than or equal to zero. Table 2 shows the mean of the differences of a one-tailed paired t-test on the values of $V^{g,n}$, i.e. last minute of gameplay for a given game g , and $V^{g,1}$, i.e. second minute of gameplay for a given game g , for all games and subjects. Results

Table 2: Mean of the differences of $V^{g,t}$ at the periods $t = 1$ (second minute of gameplay) and $t = n$ (last minute of gameplay), for all subjects in each game (g). Values in bpm (beats per minute). Significance was tested with a one-tailed paired t-test

Game (g)	Mean of the differences between $V^{g,n}$ and $V^{g,1}$
Mushroom (M)	6.11 ***
Platformer (P)	5.10 ***
Tetris (T)	3.33 ***

*** $p < 0.001$

Table 3: Mean of the differences of $V^{g,t}$ at key periods, for all subjects in a given game g (M is for Mushroom, P for Platformer and T for Tetris). Values in bpm (beats per minute)

	Pairs			
g	$V^{g,1},$ $V^{g,0}$	$V^{g,n},$ $V^{g,n-1}$	$V^{g,n},$ $V^{g,0}$	$V^{g,n-1},$ $V^{g,1}$
M	-0.87	2.39	5.23	3.71
P	-1.31	2.57	3.78	2.52
T	-1.71	1.22	1.62	2.10

indicate the difference is greater than zero with statistical significance for all games. For the Mushroom game, the mean of the differences between the last ($V^{M,n}$) and the second ($V^{M,1}$) minutes of gameplay is 6.11 bpm ($p < 0.001$). For the Platformer game, the mean of the differences of $V^{P,n}$ and $V^{P,1}$ is 5.1 bpm ($p < 0.001$). Finally, for the Tetris game, the mean of the differences of $V^{T,n}$ and $V^{T,1}$ is 3.33 bpm ($p < 0.001$). Those numbers support our experimental expectations that the HR mean during the last minute of gameplay is greater than the HR mean during the second minute of gameplay, for all games.

In order to further explore the mean variation of HR at key periods other than the

previously mentioned ones, we calculated the mean of the differences involving $V^{g,0}$, $V^{g,1}$, $V^{g,n}$ and $V^{g,n-1}$, for all games and subjects. Results are shown in Table 3. $V^{g,0}$ and $V^{g,1}$ are the values of V for a given game g during the first and the second minute of gameplay, respectively. $V^{g,n}$ and $V^{g,n-1}$ represent the values of V for a given game g during the last and the immediately before the last minute of gameplay, respectively. As previously mentioned, the value of n , the last minute of gameplay, is different for each subject since subjects might have been defeated by the game at different moments due to personal skill levels.

In the first two minutes of gameplay ($t = 0$ and $t = 1$), the mean of the differences between $V^{g,1}$ and $V^{g,0}$ is negative for all games. The mean of the differences is -0.87 bpm for the Mushroom game ($V^{M,1}$ and $V^{M,0}$), -1.31 bpm for the Platformer game ($V^{P,1}$ and $V^{P,0}$) and -1.71 bpm for the Tetris game ($V^{T,1}$ and $V^{T,0}$). Those numbers suggest a higher HR mean during the first minute of the games ($t = 0$) than during the second minute ($t = 1$). At the last two minutes of gameplay ($t = n$ and $t = n - 1$), the mean of the differences between $V^{g,n}$ and $V^{g,n-1}$ is positive for all games. The mean of the differences is 2.39 bpm for the Mushroom game ($V^{M,n}$ and $V^{M,n-1}$), 2.57 bpm for the Platformer game ($V^{P,n}$ and $V^{P,n-1}$) and 1.22 bpm for the Tetris game ($V^{T,n}$ and $V^{T,n-1}$). Those numbers suggest a higher HR mean during the last minute of the game ($t = n$) compared to the penultimate minute ($t = n - 1$).

Regarding the last ($t = n$) and the first ($t = 0$) minutes of gameplay, the mean of the differences between $V^{g,n}$ and $V^{g,0}$ is 5.23 bpm for the Mushroom game ($V^{M,n}$ and $V^{M,0}$), 3.78 bpm for the Platformer game ($V^{P,n}$ and $V^{P,0}$) and 1.62 bpm for the Tetris game ($V^{T,n}$ and $V^{T,0}$). Regarding the penultimate ($t = n - 1$) and the second ($t = 1$) minutes of gameplay, results show that the mean of the differences between $V^{g,n-1}$ and $V^{g,1}$ is 3.71 bpm for the Mushroom game ($V^{M,n-1}$ and $V^{M,1}$), 2.52 bpm for the Platformer game ($V^{P,n-1}$ and $V^{P,1}$) and 2.1 bpm for the Tetris game ($V^{T,n-1}$ and $V^{T,1}$). Both sets of numbers suggest a higher HR mean during the last minute of the game ($t = n$) compared to the first minute ($t = 0$), as well as a higher HR mean during the penultimate minute of gameplay ($t = n - 1$) compared to the second minute ($t = 1$).

Table 4: The amount of FA annotations made for all subjects during the games

Game	Period	
	H_0	H_1
Mushroom	90	98
Platformer	88	181
Tetris	110	159

5.3. Facial actions

We analyzed the number of subjects that featured a particular FA, alongside with the number of repetitions of such FA, for all three games. The analysis is also divided according to the period of the game. Only FA featured by two or more subjects were considered, since it produces an analysis that is connected to more frequent FA among the whole group of subjects instead of the peculiarities of a single person. Table 4 shows the amount of FA annotations made for all subjects during the games. According to the results, the amount of FA annotations made during H_1 (second half) of all three games was greater than the amount of annotations made during H_0 (first half). The increase in annotations during H_1 compared to H_0 was 8.8%, 105.6% and 44.5% higher for the Mushroom, Platformer and the Tetris game, respectively.

Regarding the FA annotated during each game, for the Mushroom game, the three most frequent FA in H_0 were frown (repeated 16 times among 5 subjects), talking (12 times, 3 subjects) and tongue touching lips (9 times, 3 subjects). The three most frequent FA in H_1 were frown (repeated 16 times among 3 subjects), talking (13 times, 5 subjects) and lips parted (13 times, 5 subjects). By comparing most frequent FA in the two periods, both frown and talking are present, however they were not featured by a significant number of participants. In fact no more than 5 subjects (25% of the participants) featured one of those FA. It suggests that individuals present distinct facial behaviors that are not easily generalizable, even in the same context. Curiously, two particular FA presented a significant

change in the amount of repetitions and subjects between the two periods: lip pressors (from 7 to 11 repetitions, 2 to 4 subjects) and lips parted (from 5 to 13 repetitions, 2 to 5 subjects). When compared to the whole group of participants, such increase is not significant (again they represent less than 25% of the participants), but it might be the indication of a pattern for two or three subjects. As suggested by previous work, the combination of such particular changes with another physiological signal, e.g. HR, might produce an acceptable detector for boredom/stress emotional state.

For the Platformer game, the three most frequent FA for H_0 were frown (19 repetitions among 3 subjects), tongue touching lips (12 repetitions, 3 subjects) and smile not showing teeth (11 repetitions, 3 subjects). For H_1 , the FA were frown (49 repetitions, 5 subjects), smile not showing teeth (21 repetitions, 7 subjects) and lips parted (17 repetitions, 5 subjects). By comparing the FA in both periods, frown was featured by more subjects (5, representing 25%) during the stressful part of the game, however more participants (7, representing 35%) also featured smiles not showing teeth as well. Additionally to those FA, 25% of the participants featured talking behavior during H_1 , externalizing game decisions.

For the Tetris game, the three most frequent FA for H_0 were frown (36 repetitions among 4 subjects), smile not showing teeth (14 repetitions, 4 subjects) and lip pressor (11 repetitions, 4 subjects). For H_1 , the FA were frown (42 repetitions among 4 subjects), lip pressor (28 repetitions, 6 subjects) and smile not showing teeth (16 repetitions, 5 subjects). By comparing those results to the most frequent FA in the Mushroom game, only frown is present in both; it is important to stress that frown was featured by less than 25% of the participants in both games, which highlights the difficulties in finding a pattern that can be applied to all subjects to identify a boring or a stressful situation, even when the most frequent FA are used. On the other hand, two FA presented a significant change from one period to another in the Tetris game: lip pressor (from 11 to 28 repetitions, 4 to 6 subjects) and talking (from 0 to 15 repetitions, 0 to 6 subjects). Both actions were featured by 30% of the participants, which could be further investigated in the pursue of FA that can help in the identification of emotional states. Regarding the talking FA, we observed from the recordings that some subjects tended to externalize in words any wrong decisions they

made in the game, such as how pieces were positioned, in a similar way observed during the Platformer game; in that sense, talking could be used as an indicator of activity in the game, since it is a clear facial manifestation that happened, in our case, when players were frustrated. For further FA analysis based on a group level, see [1] .

Finally we conducted a per-subject inspection of all annotated FA according to the procedure described in Section 4. We aimed to identify, for each subjects, which FA appeared in H_0 (or H_1) of *all* three games with a higher frequency than they did in H_1 (or H_0), if any. Table 5 shows the results of such inspection. Marked numbers represent the frequency of a FA that was present in all three games for the specified subject and period. In total 10 participants (50%) featured at least one FA that appeared in all three games, in the same period (boring or stressful part) with a frequency equal or greater than its appearance in the counter-period. Subject 2, for instance, featured one lip pressor during H_0 , while the total number of times the same FA appeared in H_1 for all three games combined was 18. We highlight that subject 16 was the only one who featured a FA more frequently in H_0 of all three games than he/she did during H_1 ; all other subjects featured FA more frequently in H_1 than in H_0 .

5.4. Discussion

Our results suggest that game-based emotion elicitation from our approach produce variations on HR and FA that are different at boring and stressful moments. It can potentially be used to train emotion detection models, particularly in a user-tailored approach. Information from HR and FA can complement one another and provide more information about emotional states, since the use of multiple signals for emotion detection is more likely to produce accurate results [56]. The following discussion presents more details on how these results have been interpreted.

A number of subjects presented a higher value for V , the relativized HR mean coefficient, towards the end of the Mushroom and the Platformer games when compared to the same period of the Tetris game, as shown by Table 1. Both the Mushroom and the Platformer game were completely new to the subjects, since they were developed exclusively

Table 5: Subject-based frequency of FA that appeared in the same period of all three games

		Frequency	
Subject	FA	H_0	H_1
2	Lip pressor	1	18 ^b
15	Lip pressor	2	9 ^b
10	Laughing	2	19 ^b
14	Laughing	3	9 ^b
12	Smile not showing teeth	2	8 ^b
13	Smile not showing teeth	0	6 ^b
18	Smile not showing teeth	4	10 ^b
11	Lips parted	1	10 ^b
17	Lip stretcher	0	8 ^b
16	Talking	7 ^b	1

^b FA was present in all three games for the specified subject and period

for the experiment. For the self-reported 5-point Likert scale regarding familiarity with the games/genres, the mean value was 2.75 for the Mushroom, 2.8 for the Platformer and 3.35 for the Tetris game (5 being extremely familiar). Such numbers could indicate that subjects were less likely to predict what was going to happen in the Mushroom and the Platformer when compared to the Tetris game. It could explain the greater number of subjects with higher V during the end (stressful) part of those two games when compared to the smaller number of subjects with higher V at the end of Tetris. The later is a popular game and subjects were more familiarized with it, so they might be more likely to guess what is about to happen in the game, reducing anxiety levels. This is specially true if the subject is trained

to deal with the inherent stress of the mechanic, for instance.

A significant number of subjects presented a negative value for V in some periods. In total 16 subjects (80%) in the Mushroom game, 11 (57.8%) in the Platformer and 12 (60%) in the Tetris game presented negative values, particularly in the first half of the game session. A negative value indicates that the subjects had a lower HR mean while playing the game at specific periods than while resting. After the experiment, some subjects reported discomfort during the resting period, mentioning that it was too long and boring. We believe the resting period might have been stressful for some subjects, as they were required to rest while being seated without any entertainment of the sort they may be used to when relaxing, e.g. mobile phones. Further investigation is required, however we could speculate that negative values could be attributed to HR deceleration from baseline caused by a negative condition, which was reported in previous work with interactions involving visual [57] and audio [58] stimuli varying in affective content. It is also possible that increased attention prompted by the initial contact with the game, i.e. learning its rules, is manifested by parasympathetically mediated HR deceleration [59]. Another explanation for such negative values is that our calculation of the subject's baseline B_s might be a weak approximation of the real HR mean of each subject during rest, since we only measured two 140-seconds long resting situations for each subject. We believe, however, that our baseline calculation is still a good parameter, since the average difference between the mean HR of the two resting periods was significantly low, as explained in Section 4.

Regarding the confirmation of our expectation regarding the differences of HR in boring and stressful moments of the games, the mean of the differences between V at the last ($t = n$) and the second ($t = 1$) minutes of gameplay, presented in Table 2, shows statistical significance in the difference for all games. It reinforces findings of previously mentioned works [39, 40, 10, 13, 60] which indicate that HR tends to be higher (above the subject's baseline) during stressful moments and lower (closer to subject's baseline) during boring moments in a gaming context. As previously described, the reason why we used $t = 1$ (second minute of gameplay) instead of $t = 0$ (first minute of gameplay) for our main comparison is because we believe the first minute of all games might not be ideal for a fair

comparison. During the first minute, subjects are less likely to be in their usual neutral emotional state. Such line of reasoning is supported by the exploratory analysis of the mean of the differences of V at periods, as presented in Table 3. In the beginning of the games, the HR mean during $t = 0$ (0:00 to 1:00) was higher than during $t = 1$ (1:01 to 2:00) for all three games. It could indicate that subjects were more stimulated at the very beginning, probably caused by the excitement of the initial contact with a game to be played. We believe that such difference during $t = 1$ and $t = 0$ could also be explained as a response to the fact that subjects probably understood the game mechanic. During the first minute of gameplay, subjects are probably still working to understand the game, so an opinion regarding boredom is still being formed. After the one minute mark, subjects are more likely to have fully understood the game, so they could judge whether it was boring. Additionally, a better understanding of the mechanic combined with the fact that subjects were not allowed to change the game pace, e.g. make the game more challenging, probably increased the feeling of boredom.

Concerning FA, even though further investigation is required, our calculations indicate that subjects featured a neutral face for a longer period of time during the first half (H_0) of all games when compared to the second half (H_1). Since FA annotations were made only when the subject's face featured anything different from her/his neutral face, more annotations indicate more facial activity. Additionally the results might indicate that subjects featured more FA (different from the neutral face) under stressful situations than they did under boring situations, where a neutral face/expression is probably dominant. Our games were designed to gradually increase the difficulty level until the subject was not able to handle it. As a consequence, we postulate that smiles and laughs during the second half could be connected to the subject's perception that the games were too difficult to continue playing properly. On the other hand, they could indicate genuine manifestations of enjoyment during the moments the subjects felt the game was properly balanced and engaging. Regarding the other FA, such as lip pressor and lips parted, further investigation is required to accurately connect or use such actions to predict/detect emotional states, however the results give a clue about how FA variations can be different on the individual level. As previously discussed, the

analysis and generalization of FA on a group level is less clear than an individual approach, since FA behavior might be specific to each person. Our per-subject analysis indicated that, for a portion of the participants, at least one FA was present in the three games, in the same period for the same person. Such information might be used as the starting point for further investigation regarding FA and an individual-tailored detection model for boredom/stress, for instance.

5.5. Limitations

One potential limitation of our work is the internal validity. As previously described, the experiment was based on a one-group posttest design, which does not use a control group to measure the effects of the treatment. Such design could be criticized for having low internal validity, since it is not possible to unambiguously attribute cause and effect [61]. A two-group approach could be suggested as having stronger internal validity, since it contains a control group and allows a less ambiguous conclusion. In the context of our research, however, any multiple group design implies the comparison of physiological signals and emotional perceptions among different people. Given the social and cultural background of the participants, it is virtually impossible to compare two groups of people regarding stress/boredom. People have different preferences, culture and expectations, which cause maturation and history threats to internal validity [62]. Additionally the process of comparing variations of physiological signals among different subjects is a complex task, even when subjects are similar, e.g. same age and sex. As a consequence, a subject in a control group might present a set of variations of signals and classify a game as boring, while a similar subject in another group might classify the same game as not boring at all, presenting a different set of variations of signals. In that light, our experiment relies on a one-group experimental design to increase internal validity, since subjects were compared with themselves, which removes inter-subject differences. It could also be argued that the laboratory setting of the experiment, i.e. room with special light and camera, could bias the results. The physical sensor used for HR monitoring could also increase subject's awareness of being monitored [63], which could affect the results.

Another limitation is the empirical approach used to annotate the FA, which was not based on a formal scheme and was conducted by a single person without validation by other researchers. We believe that the exploratory nature of our study regarding FA allows the use of such an approach. Our aim is not to standardize FA regarding stress/boredom, but to document the perceptions of naked-eye observations of FA in a context involving games, so that they can be used to guide further steps regarding the utilization of FA in a multifactorial analysis. A frame-by-frame annotation of our video recordings using a formal scheme, such as FACS, would be a significantly more laborious and time-consuming task, which is not motivated by our exploratory and empirical approach. Additionally a larger sample and the use of a more formal approach for FA annotations and analysis could produce more conclusive results regarding facial activity. The use of physical sensors and the manual approach used for the analysis of FA also limit the reach of our approach, e.g. for use in the games industry or in the game research community it would be essential that the two states explored could be identified in an automated way, ideally without the need of connecting players to physical monitors. Another limitation is the assumption used when dividing each game session in half, presuming that the middle point of the period indicates a transition from two distinct periods: H_0 , perceived as more boring, and H_1 , perceived as more stressing. It is not necessarily true. Even though our data indicate that subjects perceived the beginning of the games as being boring and the end as being stressful, our point of division or the periods themselves remain an assumption. There might be moments towards the end of the game, for instance, that could be perceived as more boring or joyful depending on the subject, since each participant has her/his own specific expectations and skill level regarding games. Finally the core mechanic of the Mushroom game is based on the color of the mushrooms (instead of patterns, for instance), which is not suitable for color blind subjects.

6. Conclusion

This paper presented the description and results of an experiment aimed at exploring the variations of heart rate (HR) and facial actions (FA) during gaming sessions with induced

boredom and stress. In total twenty adults of different ages and gaming experiences participated in the experiment, where they played three different games while being recorded by a video camera and monitored by a HR sensor. The games used in the experiment were carefully designed and implemented to have a difficulty level that linearly increases over time, from a boring to a stressful point. According to self-reported answers in post-games questionnaires, participants perceived the games as being boring at the beginning and stressful at the end. Such configuration gives our experiment a novel approach for the exploration of HR and FA regarding their connection to emotional states, since information can be categorized according to the induced (and theoretically known) emotional states.

The HR data related to the game session of each subject was divided in periods of 1 minute each, whose HR mean was calculated and compared to a baseline value obtained from the HR mean of the subject during rest. Based on the self-reported answers regarding stress and boredom, we analysed the HR mean at specific periods, such as the second minute of gameplay (perceived as boring) and the last minute of gameplay (perceived as stressful). Our results indicate that the average HR mean for all subjects during the last minute of gameplay was greater than the average HR mean during the second minute of gameplay, for all games with statistical significance ($p < 0.001$). Our findings are aligned with and reinforce previous research that indicate higher HR mean during stressful situations in a gaming context. The design of our games permitted a more elaborated analysis of boring and stressful periods, which contributes information regarding variations of HR mean during such conditions in gaming sessions. Additionally we performed an exploratory investigation regarding HR mean during other key periods in the games, e.g. first and penultimate minutes of gameplay. Further analysis is still required, however our numbers suggest that the average HR mean during the first minute of gameplay was greater than during the second minute of gameplay, probably as a consequence of unusual excitement during the first minute, e.g. the idea of playing a new game.

Regarding facial actions, the 6 hours of recordings of all subjects were manually analyzed and FA were annotated empirically. The annotations were categorized according to the period when they happened (the boring first part or the stressful second part of the games).

We conducted an analysis on such annotated FA on group and individual level, aiming to find patterns between the featured FA and the boring/stressful periods of the games. Our results show that more FA annotations were made during the stressful part of the games, which indicates that participants remained with a neutral face for longer periods of time during the boring part. The analysis on group level revealed that any FA pattern was related to 5 subjects (25% of the group) at most. In the analysis conducted on the individual level, we found particular patterns for 10 subjects (50% of the group).

Our findings suggest that changes in the HR during gaming sessions is a promising indicator of stress, which could be incorporated into a model aimed at emotion detection. As pointed out by previous work, a user-tailored model based on several signals, e.g. HR and FA, is more likely to detect emotional states of users. In the context where the measurement of physiological signals by physical and contact-based sensors is intrusive or not desired, e.g. remote estimation of HR, information from different channels is required. One of such additional channels of information might be facial expressions, such as the FA analysis performed in this paper. For the context of our experiment, FA analysis on an individual level produced more information to connect FA and stress/boredom emotional states. We believe that this paper contributes with information regarding HR and FA in the context of games, which can be combined to create user-tailored models for emotion detection based on different data sources.

Acknowledgements

The authors would like to thank the participants and all involved personnel for their valuable contributions. This work has been performed with support from: CNPq, Conselho Nacional de Desenvolvimento Científico e Tecnológico - Brasil; University of Skövde; EU Interreg ÖKS project Game Hub Scandinavia; UFFS, Federal University of Fronteira Sul.

References

- [1] F. Bevilacqua, P. Backlund, H. Engstrom, Variations of Facial Actions While Playing Games with Inducing Boredom and Stress, in: 2016 8th International Conference on Games and Virtual Worlds for

- Serious Applications (VS-GAMES), IEEE, Institute of Electrical and Electronics Engineers (IEEE), 1–8, doi:10.1109/vs-games.2016.7590374, URL <http://dx.doi.org/10.1109/vs-games.2016.7590374>, 2016.
- [2] P. J. Lang, The emotion probe: Studies of motivation and attention., *American psychologist* 50 (5) (1995) 372, doi:10.1037//0003-066x.50.5.372.
 - [3] P. Ekman, W. V. Friesen, Constants across cultures in the face and emotion., *Journal of Personality and Social Psychology* 17 (2) (1971) 124–129, doi:10.1037/h0030377.
 - [4] J. T. Cacioppo, G. G. Berntson, J. T. Larsen, K. M. Poehlmann, T. A. Ito, et al., The psychophysiology of emotion, *Handbook of emotions* 2 (2000) 173–191.
 - [5] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, J. G. Taylor, Emotion recognition in human-computer interaction, *Signal Processing Magazine, IEEE* 18 (1) (2001) 32–80.
 - [6] R. W. Picard, *Affective computing*, MIT press, 2000.
 - [7] J. M. Kivikangas, G. Chanel, B. Cowley, I. Ekman, M. Salminen, S. Järvelä, N. Ravaja, A review of the use of psychophysiological methods in game research, *Journal of Gaming & Virtual Worlds* 3 (3) (2011) 181–199, doi:10.1386/jgvw.3.3.181-1.
 - [8] P. Rani, C. Liu, N. Sarkar, E. Vanman, An empirical study of machine learning techniques for affect recognition in human–robot interaction, *Pattern Analysis and Applications* 9 (1) (2006) 58–69, doi:10.1007/s10044-006-0025-y.
 - [9] T. Saari, N. Ravaja, J. Laarni, K. Kallinen, M. Turpeinen, Towards emotionally adapted Games, *Proceedings of Presence* (2004) 182–189.
 - [10] F. Bousefsaf, C. Maaoui, A. Pruski, Remote assessment of the Heart Rate Variability to detect mental stress, in: *Proceedings of the ICTs for improving Patients Rehabilitation Research Techniques, IEEE, Institute for Computer Sciences, Social Informatics and Telecommunications Engineering (ICST)*, 348–351, doi:10.4108/icst.pervasivehealth.2013.252181, 2013.
 - [11] C. Yun, D. Shastri, I. Pavlidis, Z. Deng, O' game, can you feel my frustration?, in: *Proceedings of the 27th international conference on Human factors in computing systems - CHI 09, ACM, Association for Computing Machinery (ACM)*, 2195–2204, doi:10.1145/1518701.1519036, 2009.
 - [12] R. Sharma, S. Khera, A. Mohan, N. Gupta, R. B. Ray, Assessment of computer game as a psychological stressor, *Indian journal of physiology and pharmacology* 50 (4) (2006) 367.
 - [13] A. Rodriguez, B. Rey, M. D. Vara, M. Wrzesien, M. Alcaniz, R. M. Banos, D. Perez-Lopez, A VR-Based Serious Game for Studying Emotional Regulation in Adolescents, *IEEE Comput. Grap. Appl.* 35 (1) (2015) 65–73, doi:10.1109/mcg.2015.8, URL <http://dx.doi.org/10.1109/mcg.2015.8>.
 - [14] N. Ravaja, T. Saari, M. Salminen, J. Laarni, K. Kallinen, Phasic Emotional Reactions to Video Game Events: A Psychophysiological Investigation, *Media Psychology* 8 (4) (2006) 343–367, doi:

- 10.1207/s1532785xmep0804.2.
- [15] G. Chanel, C. Rebetez, M. Bétrancourt, T. Pun, Emotion Assessment From Physiological Signals for Adaptation of Game Difficulty, *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 41 (6) (2011) 1052–1063, doi:10.1109/tsmca.2011.2116000, URL <http://dx.doi.org/10.1109/TSMCA.2011.2116000>.
 - [16] T. J. W. Tijs, D. Brokken, W. A. IJsselsteijn, Dynamic Game Balancing by Recognizing Affect, in: *Fun and Games*, Springer Nature, 88–93, doi:10.1007/978-3-540-88322-7_9, 2008.
 - [17] R. L. Hazlett, Measuring emotional valence during interactive experiences: boys at video game play, in: *Proceedings of the SIGCHI conference on Human Factors in computing systems*, ACM, 1023–1026, doi:10.1145/1124772.1124925, 2006.
 - [18] R. L. Mandryk, M. S. Atkins, K. M. Inkpen, A continuous and objective evaluation of emotional experience with interactive play environments, in: *Proceedings of the SIGCHI conference on Human Factors in computing systems*, ACM, 1027–1036, 2006.
 - [19] S. Jerrieta, M. Murugappan, R. Nagarajan, K. Wan, Physiological signals based human emotion Recognition: a review, in: *2011 IEEE 7th International Colloquium on Signal Processing and its Applications*, IEEE, Institute of Electrical and Electronics Engineers (IEEE), 410–415, doi: 10.1109/cspa.2011.5759912, 2011.
 - [20] N. Ravaja, Effects of a small talking facial image on autonomic activity: The moderating influence of dispositional BIS and BAS sensitivities and emotions, *Biological Psychology* 65 (2) (2004) 163–183, doi:10.1016/s0301-0511(03)00078-4.
 - [21] L. R. Goldberg, The structure of phenotypic personality traits., *American psychologist* 48 (1) (1993) 26, doi:10.1037//0003-066x.48.1.26.
 - [22] J. van Aart, C. Bartneck, J. Hu, M. Rauterberg, B. Salem, How to behave as Alice in Wonderland—about boredom and curiosity, *Entertainment Computing* 1 (3) (2010) 125–137, doi: 10.1016/j.entcom.2010.09.003.
 - [23] R. Koster, *Theory of fun for game design*, ” O’Reilly Media, Inc.”, 2013.
 - [24] J. Schell, *The Art of Game Design: A book of lenses*, CRC Press, 2014.
 - [25] K. Salen, E. Zimmerman, *Rules of play: Game design fundamentals*, MIT press, 2004.
 - [26] J. Chen, Flow in games (and everything else), *Communications of the ACM* 50 (4) (2007) 31, doi: 10.1145/1232743.1232769.
 - [27] M. Csikszentmihalyi, K. Rathunde, The psychology of wisdom: An evolutionary interpretation, *Wisdom: Its nature, origins, and development* (1990) 25–51doi:10.1017/cbo9781139173704.004.
 - [28] E. Brown, P. Cairns, A grounded investigation of game immersion, in: *CHI’04 extended abstracts on Human factors in computing systems*, ACM, 1297–1300, 2004.

- [29] D. Weibel, B. Wissmath, Immersion in computer games: The role of spatial presence and flow, *International Journal of Computer Games Technology* 2011 (2011) 6.
- [30] C. A. Cruz, J. A. R. Uresti, Player-centered game AI from a flow perspective: Towards a better understanding of past trends and future directions, *Entertainment Computing* 20 (2017) 11–24, doi: 10.1016/j.entcom.2017.02.003.
- [31] P. Sweetser, P. Wyeth, GameFlow, *Computers in Entertainment (CIE)* 3 (3) (2005) 3–3, doi: 10.1145/1077246.1077253.
- [32] A. Landowska, Emotion Monitoring Verification of Physiological Characteristics Measurement Procedures, *Metrology and Measurement Systems* 21 (4) (2014) 719–732, ISSN 2300-1941, doi:10.2478/mms-2014-0049.
- [33] C. Stockhausen, J. Smyzek, D. Krömker, Beats down: Using heart rate for game interaction in mobile settings, in: *Human-Computer Interaction–INTERACT 2013*, Springer, 523–530, 2013.
- [34] X. Xiao, J. Wang, Towards Attentive, Bi-directional MOOC Learning on Mobile Devices, in: *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction - ICMI '15*, ACM, Association for Computing Machinery (ACM), 163–170, doi:10.1145/2818346.2820754, 2015.
- [35] P. A. Nogueira, V. Torres, R. Rodrigues, E. Oliveira, An annotation tool for automatically triangulating individuals’ psychophysiological emotional reactions to digital media stimuli, *Entertainment Computing* 9-10 (2015) 19–27, doi:10.1016/j.entcom.2015.06.003.
- [36] B. M. Appelhans, L. J. Luecken, Heart rate variability as an index of regulated emotional responding., *Review of general psychology* 10 (3) (2006) 229.
- [37] M. Fenton-O’Creevy, J. T. Lins, S. Vohra, D. W. Richards, G. Davies, K. Schaaff, Emotion regulation and trader expertise: Heart rate variability on the trading floor., *Journal of Neuroscience, Psychology, and Economics* 5 (4) (2012) 227.
- [38] C. Schubert, M. Lambertz, R. Nelesen, W. Bardwell, J.-B. Choi, J. Dimsdale, Effects of stress on heart rate complexity—a comparison between short-term and chronic stress, *Biological psychology* 80 (3) (2009) 325–332.
- [39] S. Vandeput, J. Taelman, A. Spaepen, S. Van Huffel, Heart rate variability as a tool to distinguish periods of physical and mental stress in a laboratory environment, in: *Proceedings of the 6th international workshop on biosignal interpretation (BSI)*, New Haven, CT, 187–190, 2009.
- [40] A. Garde, B. Laursen, A. Jørgensen, B. Jensen, Effects of mental and physical demands on heart rate variability during computer work, *European journal of applied physiology* 87 (4-5) (2002) 456–461.
- [41] B. Grundlehner, L. Brown, J. Penders, B. Gyselinckx, The Design and Analysis of a Real-Time, Continuous Arousal Monitor, in: *2009 Sixth International Workshop on Wearable and Implantable Body Sensor Networks*, IEEE, Institute of Electrical and Electronics Engineers (IEEE), 156–161, doi:

- 10.1109/bsn.2009.21, 2009.
- [42] D. McDuff, S. Gontarek, R. Picard, Remote measurement of cognitive stress via heart rate variability, in: 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Institute of Electrical and Electronics Engineers (IEEE), ISSN 1557-170X, 2957–2960, doi: 10.1109/embc.2014.6944243, URL <http://dx.doi.org/10.1109/EMBC.2014.6944243>, 2014.
 - [43] D. J. McDuff, J. Hernandez, S. Gontarek, R. W. Picard, COGCAM: Contact-free Measurement of Cognitive Stress During Computer Tasks with a Digital Camera, in: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16, Association for Computing Machinery (ACM), doi:10.1145/2858036.2858247, URL <http://dx.doi.org/10.1145/2858036.2858247>, 2016.
 - [44] G. Giannakakis, M. Pediaditis, D. Manousos, E. Kazantzaki, F. Chiarugi, P. Simos, K. Marias, M. Tsiknakis, Stress and anxiety detection using facial cues from videos, Biomedical Signal Processing and Control 31 (2017) 89–101, doi:10.1016/j.bspc.2016.06.020, URL <http://dx.doi.org/10.1016/j.bspc.2016.06.020>.
 - [45] P. Ekman, W. V. Friesen, Facial action coding system .
 - [46] J. F. Cohn, Z. Ambadar, P. Ekman, Observer-based measurement of facial expression with the Facial Action Coding System, The handbook of emotion elicitation and assessment (2007) 203–221.
 - [47] A. Samara, L. Galway, R. Bond, H. Wang, Sensing Affective States Using Facial Expression Analysis, in: Ubiquitous Computing and Ambient Intelligence, Springer Nature, 341–352, doi:10.1007/978-3-319-48746-5_35, 2016.
 - [48] D. McDuff, Discovering facial expressions for states of amused, persuaded, informed, sentimental and inspired, in: Proceedings of the 18th ACM International Conference on Multimodal Interaction - ICMI 2016, ACM, Association for Computing Machinery (ACM), 71–75, doi:10.1145/2993148.2993192, URL <http://dx.doi.org/10.1145/2993148.2993192>, 2016.
 - [49] Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang, J. F. Cohn, Q. Ji, L. Yin, Multimodal Spontaneous Emotion Corpus for Human Behavior Analysis, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 3438–3446, doi:10.1109/cvpr.2016.374, 2016.
 - [50] D. Heylen, M. Ghijsen, A. Nijholt, R. op den Akker, Facial signs of affect during tutoring sessions, in: Affective Computing and Intelligent Interaction, Springer, 24–31, 2005.
 - [51] J. F. Grafsgaard, J. B. Wiggins, K. E. Boyer, E. N. Wiebe, J. C. Lester, Automatically Recognizing Facial Expression: Predicting Engagement and Frustration, in: EDM, 43–50, 2013.
 - [52] J. N. Bailenson, E. D. Pontikakis, I. B. Mauss, J. J. Gross, M. E. Jabon, C. A. Hutcherson, C. Nass, O. John, Real-time classification of evoked emotions using facial feature tracking and physiological responses, International Journal of Human-Computer Studies 66 (5) (2008) 303–317, doi:

- 10.1016/j.ijhcs.2007.10.011, URL <http://dx.doi.org/10.1016/j.ijhcs.2007.10.011>.
- [53] N. Shaker, S. Asteriadis, G. N. Yannakakis, K. Karpouzis, A game-based corpus for analysing the interplay between game context and player experience, in: *Affective Computing and Intelligent Interaction*, Springer, 547–556, 2011.
 - [54] N. Unsworth, T. S. Redick, B. D. McMillan, D. Z. Hambrick, M. J. Kane, R. W. Engle, Is playing video games related to cognitive abilities?, *Psychological science* 26 (6) (2015) 759–774.
 - [55] M. S. Terlecki, N. S. Newcombe, How important is the digital divide? The relation of computer and videogame usage to gender differences in mental rotation ability, *Sex Roles* 53 (5) (2005) 433–441.
 - [56] D. Kukolja, S. Popović, M. Horvat, B. Kovač, K. Čosić, Comparative analysis of emotion estimation methods based on physiological measurements for real-time applications, *International Journal of Human-Computer Studies* 72 (10-11) (2014) 717–727, doi:10.1016/j.ijhcs.2014.05.006, URL <http://dx.doi.org/10.1016/j.ijhcs.2014.05.006>.
 - [57] B. N. Cuthbert, M. M. Bradley, P. J. Lang, Probing picture perception: Activation and emotion, *Psychophysiology* 33 (2) (1996) 103–111, doi:10.1111/j.1469-8986.1996.tb02114.x.
 - [58] C. V. O. Witvliet, S. R. Vrana, Play it again Sam: Repeated exposure to emotionally evocative music polarises liking and smiling responses, and influences other affective reports, facial EMG, and heart rate, *Cognition & Emotion* 21 (1) (2007) 3–25, doi:10.1080/02699930601000672.
 - [59] N. Ravaja, Contributions of psychophysiology to media research: Review and recommendations, *Media Psychology* 6 (2) (2004) 193–235.
 - [60] T. Yamakoshi, K. Yamakoshi, S. Tanaka, M. Nogawa, M. Shibata, Y. Sawada, P. Rolfe, Y. Hirose, A Preliminary Study on Driver's Stress Index Using a New Method Based on Differential Skin Temperature Measurement, in: *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, Institute of Electrical and Electronics Engineers (IEEE)*, 722–725, doi:10.1109/iembs.2007.4352392, 2007.
 - [61] R. E. Kirk, *Experimental design*, Wiley Online Library, 1982.
 - [62] W. M. Trochim, J. P. Donnelly, *Research methods knowledge base*.
 - [63] J. A. Healey, R. W. Picard, Detecting stress during real-world driving tasks using physiological sensors, *Intelligent Transportation Systems, IEEE Transactions on* 6 (2) (2005) 156–166.