The Pupil Response as an Indicator of User Experience in a Digital Exercise Game

Michael O. Gutjahr[1], Wolfgang Ellermeier[1], Sandro Hardy[2], Stefan Göbel[2] and Josef Wiemeyer[3]

[1]Institute of Psychology, Technische Universität Darmstadt, Darmstadt, Germany

[2]Department of Electrical Engineering & Information Technology, Technische Universität Darmstadt, Darmstadt, Germany

[3]Institute of Sports Science, Technische Universität Darmstadt, Darmstadt, Germany

Running head: PUPIL RESPONSES INDICATING GAME EFFECTS

Corresponding authors:

Michael O. Gutjahr

Technische Universität Darmstadt, Institute of Psychology, Alexanderstr. 10, D – 64283 Darmstadt, Germany.

E-mail: michael.gutjahr@kom.tu-darmstadt.de

Wolfgang Ellermeier

Technische Universität Darmstadt, Institute of Psychology, Alexanderstr. 10, D – 64283 Darmstadt, Germany.

E-Mail: ellermeier@psychologie.tu-darmstadt.de
phone: ++49-6151-1624010

Abstract

To study whether psychophysiological indicators are suitable measures of user experience in a digital exercise game (exergame), a laboratory study employing both psychophysiological and self-report measures was conducted. 66 participants cycled for ten minutes on an ergometer while pupil diameter, skin conductance and heart rate were measured; afterwards they completed a user experience questionnaire. The participants performed under three experimental conditions varying between subjects: active gaming (participants controlled the altitude of a digital bird by varying their pedal rate in order to catch letters flying across the screen), observing a game (they observed a replay of another participant's game) and no game (blank screen). Only the gaming condition showed evidence for statistically significant pupil dilations - indicating emotional arousal - in response to game events (catching a letter) or corresponding points in time. The observational condition did not differ statistically from the no-game control condition. Self-reports also indicated that the gaming condition was rated most fun and least demanding. Other psychophysiological indicators (heart rate, skin conductance) showed no systematic effects in response to game events, rather they steadily increased during training. Thus pupil responses were shown to be suitable indicators of positive emotional reactions to game events and user experience in a (training) game.

Keywords: User experience, emotion, pupillometry, pupil reaction, exergame

1. Introduction

For more than two decades, numerous laboratory studies have evaluated the use of psychophysiological measurement to detect emotional reactions to standardized stimuli such as pictures (IAPS: Bradley & Lang 1999a; Lang, Bradley, & Cuthbert 2005), sounds (IADS: Bradley & Lang 2007), and words (ANEW: Bradley & Lang 1999b). These studies have successfully validated a number of psychophysiological indicators of emotion, e.g. skin conductance (SC; e.g. Braithwaite, Watson, Jones, & Rowe, 2013), heart rate (HR; e.g. Bradley, Miccoli, Escrig, & Lang, 2008), electromyography (EMG; e.g. Codispoti, Bradley, & Lang, 2001), or electroencephalography (EEG; e.g. Cuthbert, Schupp, Bradley, Birbaumer, & Lang, 2000; Müller, Keil, Gruber, & Elbert 1999), typically in response to previously rated pictures and sounds (e.g. Bradley & Lang, 2000; Baumgartner, Esslen, & Jäncke 2006; Partala & Surakka, 2003). During the last decade, however, the focus has shifted towards emotional and cognitive reactions to complex or dynamic stimuli, like those occurring in digital games (e.g. Allison & Polich, 2008; Barlett & Rodeheffer, 2009; Kivikangas, & Ravaja, 2013; Nacke, 2009; Mandryk, Inkpen, & Calvert, 2006; Slater et al., 2006; for reviews see Kirsh, 2006; Kivikangas, Chanel, Cowley, Ekman, Salminen, Jarvela, & Ravaja, 2013; Potter & Bolls, 2012; Wiemeyer, Kickmeier-Rust, & Steiner, 2016). Few of these studies have actually used phasic psychophysiological measurements in response to game events (e.g. Ravaja, Saari, Salminen, Laarni, & Kallinen, 2006; Salmine, & Ravaja, 2007). Therefore, the main purpose of the present study is to verify whether a *remotely measured psychophysiological indicator* (here: pupil dilation) can reveal *phasic emotional reactions* to predefined events in *a complex game*. In addition several self-report user experience measures are collected to examine to which extent emotional reactions to specific events determine the overall experience.

To be able to measure an emotional reaction, it needs to be known what ‘an emotion’ is, in fact “part of the complexity of studying emotion is defining it” (Bradley & Lang,

2009, p.581). “An emotion can be loosely defined as a reaction to personally significant events, where ‘reaction’ is taken to include biological, cognitive, and behavioral reactions, as well as subjective feelings of pleasure or displeasure” (Parrott, 2004, p. 6). However, factor analyses have shown two dominant dimensions to be important: pleasure and arousal (e.g. Russell, 1980). This may be caused by two underlying “motivational systems, one appetitive and one defensive” (Bradley, Codispoti, Cuthbert, & Lang, 2001, p. 276). This dichotomy, i.e. whether something is good and desirable or bad and to be avoided, is a core component of most cognitive theories of emotion (Arnold, 1960; Lazarus, 1968; Weiner, 1986; Ortony, Clore, & Collins, 1988). One such theory (Ortony et al., 1988) defines the emotion of ‘having fun’ to occur when an event (1) seems to be sure to happen (or has already happened), (2) will be of primary concern to oneself, (3) is desirable and (4) the result has not been expected for sure (Reisenzein, Mayer, & Schützwohl, 2003). So mastering a challenging task may qualify as such a positive emotion, in fact the one to be investigated in the present study. This is related to the concept of ‘flow’ (Nakamura & Csikszentmihalyi, 2002; Sweetser & Wyeth, 2005; Weber, Tamborini, Westcott-Baker, & Kantor, 2009; Schmierbach, Chung, Wu, & Kim, 2014) which refers to “a sense that one is engaging challenges at a level appropriate to one’s capacities” (Nakamura & Csikszentmihalyi, 2002, p. 90).

When attempting to measure psychophysiological reactions to discrete (positive) events in a game, clearly, such measuring instruments are to be preferred, which do not restrict the players’ action space or spoil their game experience - such as being ‘wired up’ to a large number of electrodes for EEG or EMG measurements. Analyzing pupil reactions by means of a camera is clearly less obtrusive. Pupil diameter (for neural pathways see: Aston-Jones, & Cohen, 2005; Mathôt, 2018) reflects the balance of the autonomic nervous system (see e.g. Steinhauer, Siegle, Condray, & Pless, 2004). Parasympathetic activation causes a constriction (miosis) of the pupil, whereas sympathetic activation causes adilation (mydriasis). Pupil diameter is known to respond

to emotionally arousing pictures (Bernhardt, Dabbs, & Riad, 1996; Bradley, Miccoli, Escrig, & Lang, 2008), sounds (Partala & Surakka, 2003) and words (Võ, Jacobs, Kuchinke, Hofmann, Conrad, Schacht, & Hutzler, 2008) and to painful stimulation (Ellermeier & Westphal, 1995). Since it was first discovered, that "increases in the size of the pupil of the eye have been found to accompany the viewing of emotionally toned or interesting visual stimuli" (Hess & Polt, 1960, p. 349), the use of pupil reactions for measuring emotions has evolved (Libby, Lacey, & Lacey, 1973; Janisse, 1974; Mudd, Conway, & Schindler, 1990; for an overview of early studies see: Goldwater, 1972). Today, measuring pupil size variations in response to standardized, and typically stationary stimuli (such as IAPS pictures) is an established psychophysiological method for measuring emotions (Bradley, Miccoli, Escrig, & Lang, 2008, Bradley & Lang, 2015; Henderson, Bradley, & Lang, 2014; Snowden, O'Farrell, Burley, Erichsen, Newton, & Gray, 2016; van Steenbergen, Band, & Hommel, 2011). The present study aims to explore, whether the pupil response is equally suited to remotely measure (phasic) emotional reactions occurring in a gaming context.

In doing so, two fundamental problems with the pupil response have to be considered: (1) its insensitivity to the quality of the affect: "pupil diameter increases when people process emotionally engaging stimuli, regardless of hedonic valence" (Bradley, Miccoli, Escrig, & Lang, 2008, p.606), and (2) its lack of specificity to the emotional domain: Note that pupil diameter variation can also indicate cognitive effects (for an overview see: Beatty,1982) such as increasing workload (Chen, Epps & Chen, 2011; Kahneman, & Beatty, 1966), conflict processing (Steenbergen & Band, 2013), or attention shifts (Laeng, Sirois & Gredebäck, 2012). As "pupil dilatation is likely to reflect more than one process at the same time" (Steenbergen & Band, 2013, p.1), workload and emotion may simultaneously affect the pupil diameter (Xu, Wang, Chen, Choi, Li, Chen & Hussain, 2011). Not at last physical effort can cause pupil changes, too (Zénon, Sidibé & Olivier, 2014). Thus, when pupil diameter is used to measure emotional reactions to a game, great care has to be taken to make sure that other variables affecting the pupil response such as

effects of illumination, physical effort and cognitive load are controlled by experimental design.

The game chosen for the present study was a ‘serious game’ (Breuer, 2010; Dörner, Göbel, Effelsberg, & Wiemeyer, 2016), more specifically an exercise game that met the theoretical requirements stated in the last paragraph and was developed in house (Göbel, Hardy, Wendel, Mehm, & Steinmetz, 2010; Hardy, Dutz, Wiemeyer, Göbel, & Steinmetz, 2015).

It required participants to pedal on an exercise bike for a 10-minute training period while being assigned to one of three treatments – each of which was presented via a screen placed in front of the bicycle. The “gaming group” played an exergame in which the participant had to regulate the altitude of a virtual bird by adjusting their pedal rate in order to catch letters flying by on the screen. The participants in the “observational group” were shown a replay of another participant’s game. Participants in the “no-game” control group completed the exercise while only the standard numerical performance indicators of the exercise bike were displayed. Pupil diameter, HR, and SC were measured during the exercise session and subsequently all participants were asked to fill out a questionnaire asking for their user experience and effort. This was done to validate the psychophysiological data by relating them to a specific quality of feeling (positive emotions).

The main goal of the present study was to show that active and successful interaction with a *dynamic* exergame triggers positive emotions to *previously neutral stimuli*, and that measuring pupil reactions to rewarding game events is suited to detect these emotional reactions, even though there might *be additional ‘noise’* (interfering arousal) *produced* by controlling the game by way of physical exercise.

2. Method

2.1 Participants

Data from 66 participants (22 per condition) with a mean age of *M* = 22.44 (*SD* = 7.76) years were collected, 36 of the participants were female (gaming = 8; observing = 16, no game = 12). On average, the participants stated they were playing digital games for 2 hours and 40 minutes per week; 25 of them claimed to be playing less than 1 h per week. They bicycled for an average of 2 hours and 18 minutes per week; 33 participants did so for less than 1 h. The participants were primarily recruited from the student body at Technische Universität Darmstadt. Psychology students participated for course credit. One participant was excluded for erroneously receiving the wrong set of instructions. The protocol of the study was approved by the ethics commission of Technische Universität Darmstadt (EK 06/2012).

2.2 Design and Experimental Conditions

The independent variable was the gaming condition consisting of three levels (active gaming / observing a game / no game), with the second group constituting a 'yoked control' linked to the first one by coupling participants in pairs (receiving the same stimulation as the experimental group, but lacking active control; Seligman & Maier, 1967). To avoid transfer effects, a between-subjects design was employed (Greenwald, 1976). HR, SC, pupil diameter, user experience, and subjectively perceived exertion were the dependent variables. For the data collection phase, the participants were randomly assigned to one of three experimental conditions, differing in the opportunity to interact with the presentation on the screen. In all three conditions participants were instructed to reach a performance level of 70 revolutions per minute (rpm) to start the training (at a power of 75 Watts).

For each of the experimental conditions participants received written instructions, and their informed consent was obtained prior to the actual experiment.

2.2.1 Control group.

In the no-game group a black screen was presented throughout, containing only a standardized footer at the bottom of the screen, showing the participant's score indicator, fixed at zero points for the entire 10 minutes, the remaining time to cycle, current speed (in rpm) and current HR.

2.2.2 Experimental group.

In the active gaming group participants were instructed to control the flight of a digital dove on the screen by varying the pedal rate of the ergometer (Göbel, Hardy, Wendel, Mehm, & Steinmetz, 2010; Hardy, Dutz, Wiemeyer, Göbel, & Steinmetz, 2015). The bird kept flying in the middle of the screen when a pedal rate of 70 rpm was maintained. The higher the pedal rate was, the higher the bird flew. After 10 seconds, letters started to appear at different altitudes at the right side of the screen and moved along the screen to eventually disappear on its left edge. Every time a participant successfully matched the height of the bird to the height of the letter at the time it passed the bird, the letter was "caught", disappeared, and the points earned for capturing the letter were displayed in a small box at the position of the letter. This message disappeared after a few seconds by slowly floating outside the picture frame (Figure 1).

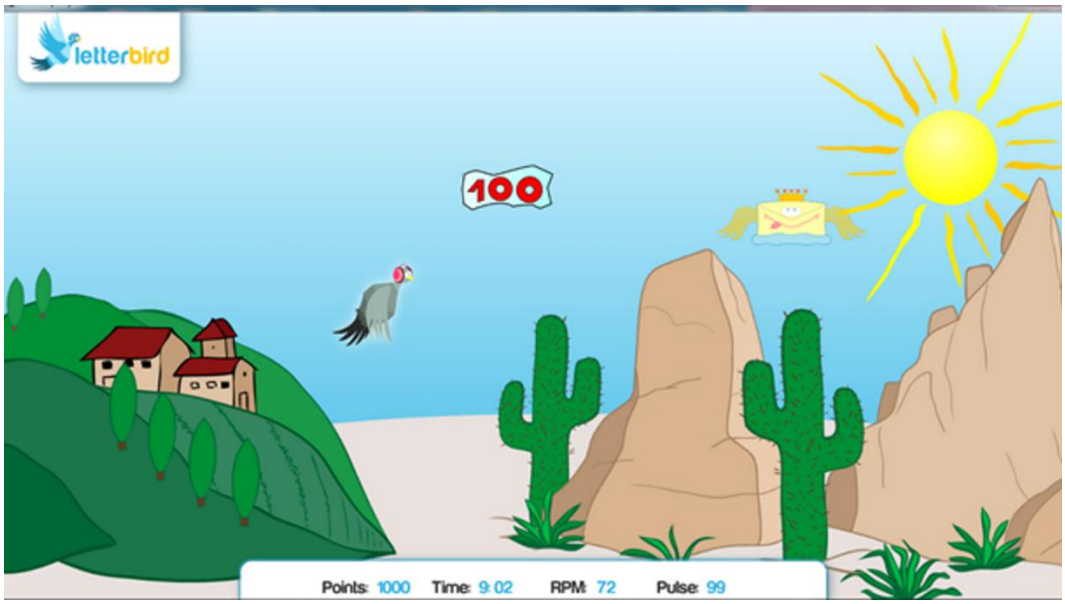

Figure 1. Screen shot of the game (gaming and observational condition). Participants were to steer the bird to catch letters flying at different altitudes from right to left. Once a letter was caught successfully, the points earned were shown on the screen.

To be able to measure the physiological reaction after successfully capturing a letter without interference from subsequent game events, the 10 minutes of game play were subdivided into "playing phases" with a high frequency of letters appearing at the edge of the screen at randomized altitude, and "measurement phases" with a low frequency of letters to appear. In the playing phase a letter appeared every 3.25 seconds on average with random deviations in an interval spanning ± 1.5 seconds (Table 1). In the measurement phase a letter appeared every 6.5 seconds on average with random deviations of up to ± 1 second. Participants were not informed about the different game phases (when asked in the post-experimental debriefing, no participant stated to be aware of different game phases). To further increase the challenge of the game, 30 percent of the letters appearing in the playing and the measurement phases moved faster and provided an opportunity to earn more points. For each participant the same number of letters was generated in each phase. Slow letters needed 6 seconds, fast letters 4 seconds to cross the screen to the point where the bird could intercept it. To avoid order effects, for each of the 22 participants participating in the gaming group a separate random sequence of letters

was generated. Furthermore, the order of the phases was balanced across participants (Table 1), half of them following schedule 1, the other half followed schedule 2. The standardized footer showed the points earned [points], remaining time [minutes:seconds], speed [rpm], and participant's HR [bpm] (see Figure 1 - from left to right).

Table 1

Organisation of the game phases

Schedule 1				*Schedule 2*			
Phase	Duration [in s]	interval betw. letters [in s]	Number of letters	Phase	Duration [in s]	interval betw. Letters [in s]	Number of letters
break	10	-	-	break	10	-	-
measurement	65	5.50 - 7.50	10	playing	125	1.75 - 4.75	39
playing	125	1.75 - 4.75	39	measurement	65	5.50 - 7.50	10
break	10	-	-	break	10	-	-
measurement	65	5.50 - 7.50	10	playing	125	1.75 - 4.75	39
playing	125	1.75 - 4.75	39	measurement	65	5.50 - 7.50	10
break	10	-	-	break	10	-	-
measurement	65	5.50 - 7.50	10	playing	125	1.75 - 4.75	39
playing	125	1.75 - 4.75	39	measurement	65	5.50 - 7.50	10

2.2.3 Yoked control group.

To implement a yoked-control condition in which the same game events were seen without active participation, in the observational group the participants completed the ergometer training while watching the entire 10-minute replay of another participant's performance who had earlier participated in the active gaming group. To match game

events in the gaming and observational groups, the gameplay of each participant in the gaming group was recorded and saved (by continuously recording rpm, altitude of the bird, letters appearing, etc.). So the game played by one person was shown as a replay to the next ('yoked') participant randomly assigned to the observational group. To assure that participants payed attention to the display, they were told that they were to observe a 'demo version' of a game which they could not control. The standardized footer showed the participant's remaining time, speed and HR, while the points displayed were taken from the replay.

2.2.4 Borg Scale.

The rating scale of perceived exertion (RPE; Borg & Borg, 2001) in its German version (Borg, 2004) was administered after the training. The scale ranges from 6 "not at all exhausting" to 20 "maximally exhausting", with 15 representing "exhausting".

2.2.5 Rating of the Training.

To evaluate their subjective perception of the 10-minute bicycling session, the participants were asked about the effort spent ("How exhausting was it to ride the bicycle?"), the distraction provided by the display ("Did the presentation on the screen distract from the exertion?") and the overall enjoyment of the training session ("How much fun has the whole session been?"). Participants rated these items on bipolar 5-point scales (- 2 = "not at all"; + 2 = "very much").

2.2.6 User Experience Questionnaire.

To measure the user experience caused by the three different game versions (gaming, observed game, no-game) a questionnaire developed by the authors (Göbel, Gutjahr, & Hardy, 2013) was used (see Appendix). The questionnaire consists of 21 items building an overall user experience score and is subdivided into 7 subscales, each consisting of 3 of the 21 items. The subscales are "avoiding negative emotion", "cognitive load", "positive emotion", "motivation", "immersion", "flow" and "arousal". This questionnaire

reflects the concepts of several user experience and flow theories and is quite similar (positive emotion, negative emotion, flow, immersion) to the game experience questionnaire proposed by Nacke (2009; see also Wiemeyer, Nacke, Moser, & Mueller, 2016), but it may be somewhat more general (including cognitive load, arousal, motivation) than the latter. The "immersion" subscale was removed for the present study. For each of the remaining 18 items the participants were asked to state how much they agreed with a given statement (e.g. "it has been fun") on a 10-point scale (1 = "not at all"; 10 = "very much"). The cognitive-load subscale inquired whether participants were "pleasantly stimulated" and "not overloaded" and therefore implies the concept of a challenging but manageable task (see Nakamura & Csikszentmihalyi, 2002). To obtain the values of the subscales as well as the overall user experience score the associated items were averaged. All scales were coded the same way, with higher scores reflecting greater user experience.

2.3 Apparatus

The game or the respective control conditions were presented on a 42-inch monitor (Samsung, 1920 x 1080 pixels), placed on a desk, adjustable in height. To measure pupil diameter, the two infrared cameras of an eye tracking system (Facelab 5, Seeing Machines, Canberra, Australia), were placed at the bottom to the left and right of the monitor. Pupil size was sampled at 60 Hz. To measure skin conductance (SC), a psychophysiological system (Captive L3000, TEA, Nancy, France) was used, the SC sensors (Procomp infiniti, Thought Technology Ltd., Montreal, Canada) were attached to the index and middle fingers of the left hand; SC was sampled at 256 Hz. To measure heart rate (HR), a pulse belt (Polar t31, Polar Electro GmbH, Büttelborn, Germany) was used. For bicycling, an ergometer (Ergo-Fit Cycle 4000 MED, Ergo-Fit, Pirmasens, Germany) was placed in front of the desk supporting the monitor and cameras. The signals recorded by the pulse belt were transferred to the computer of the ergometer. Here, peak to peak time (the RR-time-interval) was used to calculate HR values. These

HR values were stored with 60 Hz resolution. The height of the screen, the saddle of the exercise bike, and the distance to the cameras were adjusted individually for each participant to ensure optimal pupil measurement.

2.4 Procedure

Upon arrival at the laboratory the participants were informed about the purpose of the experiment and asked to fill out an informed consent form. After a short introduction on "how to attach the sensors correctly", the participants attached the sensors themselves. Subsequently, they were randomly assigned to one of the experimental conditions, and instructed regarding their task. SC, HR, and pupil diameter were continuously recorded. After the 10-min game/exercise, the participants removed the sensors and completed the Borg scale, the user experience questionnaire, and the ratings, and provided socio-demographical data. Finally the participants were debriefed.

2.5. Data reduction

To analyze the psychophysiological data, reactions to game events had to be defined. For the active gaming group, the moment a letter was successfully caught by a participant during the measuring phase constituted the onset of the emotional event to be analyzed (participants in this group successfully caught M = 99.13% letters during the measuring phase and M = 94.49% letters during the entire game). For the observed-game (i.e. yoked-control) group, the events generated by the player they observed were used. For the no-game control group looking at a black screen, the corresponding points in time of a matched participant in the gaming group were used. A 6-s interval was analyzed for each event, beginning 600 ms before and ending 5.400 ms after the onset of the event. The average values computed for the 600 ms before the onset were used as a baseline. The mean of the measurements from 2 seconds to 5 seconds after onset, i.e. a 3-sec interval, defined the pupil-dilation response (compare Bradley et al. 2008). To avoid accumulation of responses (which could still occur, since the letters appeared with at least 5.5 s separation, see Table 1, but crossed the screen at different speeds) only letters separated

by at least 5 s at the line of interception were evaluated. To eliminate measuring errors, the raw data were filtered for implausible pupil size values. Pupil (raw data) measurements of less than 2.5 mm diameter (averaging 0.27 percent per participant) and greater than 7.5 mm (1.65 percent per participant) were defined as artefacts and excluded.

2.6 Statistical Analysis

Since SC and HR continuously rose during the ten-minute training, and since the pupil diameter is bound to increase while watching a black screen, relative changes of the psychophysiological measurements were used, rather than the raw data. To achieve this, all measurements were referenced to the 600-ms baseline interval before the onset of the emotional event [(test interval – baseline) / baseline] for further analysis. To statistically analyze if there is a difference between the three levels of the independent variable, nonparametric tests (Kruskal-Wallis H tests) were performed, since the pupil-dilation measure was considerably skewed. For additional nonparametric testing within subjects, Friedman tests were used. The questionnaire-based user experience data were subjected to analyses of variance (ANOVAs). Testing was done two-sided, assuming a type-I error of $\alpha = 0.05$.

3. Results

3.1 Scale of perceived exertion.

The Borg scale was completed by 64 participants. The active gaming group (M = 11.85; SD = 2.65; n = 20) perceived the ergometer training to be slightly less demanding than the observational (M = 13.23; SD = 2.53; n = 22) and the no-game control groups (M = 12.36; SD = 2.19; n = 22). These differences were not statistically significant, however, $F(2,61) = 1.701$; $p = 0.19$.

3.2 Rating of the Training.

The training was rated by 62 participants. The active gaming group (M = 0.00; SD = 1.12; n = 20), the observational group (M = 0.33; SD = 1.02; n = 21) and the no-game group (M = -0.24; SD = 1.00; n = 21) did not differ significantly in perceived effort during the training, $F(2,59) = 1.583$; $p = 0.21$. In the active gaming group, however, the visualization on the screen (M = 0.95; SD = 0.83) was perceived to distract more from the effortful training than in the observational (M = -0.10; SD = 1.30) and the no-gaming groups (M = -0.81; SD = 1.29), these differences being statistically significant, $F(2,59) = 11.801$; $p < .001$; $\eta^2_p = 0.29$. In addition, for the gaming group (M = 1.40; SD = 0.60) the training session appeared to be more fun than for the observational (M = 0.62; SD = 1.07) and the no-game groups (M = 0.38; SD = 1.12). This effect was statistically significant as well, $F(2,59) = 6.239$; $p = 0.003$; $\eta^2_p = 0.18$. Pairwise testing using two-tailed t-tests shows: While there is no significant difference between the observational and the no-game group ($p > 0.05$), the active gaming group significantly differed from the observational and the no-game group in perceived distraction and enjoying the training session ($p < 0.05$).

3.3 User Experience Questionnaire.

The user experience questionnaire was completed by 64 participants. The data of 58 to 64 participants were used to calculate Cronbach's alpha for the overall user experience score (0.93) and for the subscales "avoiding negative emotion" (0.42), "cognitive load" (0.49), "positive emotion" (0.78), "motivation" (0.80), "flow" (0.91), and "arousal" (0.75). Overall user experience was higher in the gaming group (M = 6.08; SD = 1.32; n = 21), than in the observational (M = 4.51; SD = 1.65; n = 22) and no-game groups (M = 3.33; SD = 1.29; n = 21). This effect was statistically significant, $F(2,62) = 19.16$; $p < 0.001$; $\eta^2_p = 0.39$. Being in the active gaming group resulted in the highest score for each of the subscales as well (all $p < 0.01$); "avoiding negative emotion" ($\eta^2_p = 0.35$), "cognitive load" ($\eta^2_p = 0.30$), "positive emotion" ($\eta^2_p = 0.38$), "motivation" ($\eta^2_p = 0.16$),

"flow" ($\eta^2_p = 0.19$), and "arousal" ($\eta^2_p = 0.29$). Post hoc testing between subjects with two-sided t-test shows: The active gaming group differed significantly from the no-game group in overall user experience score and on all subscales ($p < 0.05$). Furthermore, the active gaming group showed greater overall user experience and had higher scores on 4 of the 6 subscales (avoiding negative emotion, cognitive load, positive emotion, motivation) than the observational group ($p < 0.05$). The observational group reported greater overall user experience and higher scores on 2 of the 6 subscales (positive emotion, arousal) than the no-game group ($p < 0.05$).

3.4.1 Pupil Reactions.

The data of 11 participants had to be excluded, because their pupils were not reliably detected by the tracking system. The data of the remaining 55 participants (gaming: n = 18 (6 female); observation: n = 17 (13 female); no-game: n = 20 (11 female)) were analysed, yielding a median number of 21 usable trials (letters succesfully caught in the measuring phase, see Method section) per participant. These traces were averaged for each participant in successive 200-ms intervals. The change in pupil diameter triggered by the critical game events (capturing letters, or respective control conditions) can be seen in Figure 2. The largest pupil response was observed in the active gaming group (M = 1.53; SD = 2.66; n = 18), changes in the observational (M = 0.40; SD = 1.34; n = 17) and no-game (M = -0.20; SD = 1.46; n = 20) group were smaller and more erratic (see Figure 2). A nonparametric, between-subjects analysis of the event-triggered pupil diameter changes (in percent) comparing all three groups was statistically significant, Kruskal-Wallis H test: $\chi^2 = 7.25$; $p = 0.027$, Cohen's $w = 0.13$. Post hoc testing further showed: In the active gaming group the relative pupil dilation was significantly greater than in the no-game ($\chi^2 = 6.18$; $p = 0.013$) and marginally greater than in the observational group ($\chi^2 = 3.66$; $p = 0.056$). The difference between the observational and the no-game group was not statistically significant ($\chi^2 = 0.95$; $p = 0.329$).

In order to check whether the analysis interval for evaluating pupil responses (2 to 5 s) was actually well chosen, we also analyzed earlier responses (0 to 1 s, 0 to 2 s, and 1 to 3 s), none of which yielded significant differences between the experimental conditions.

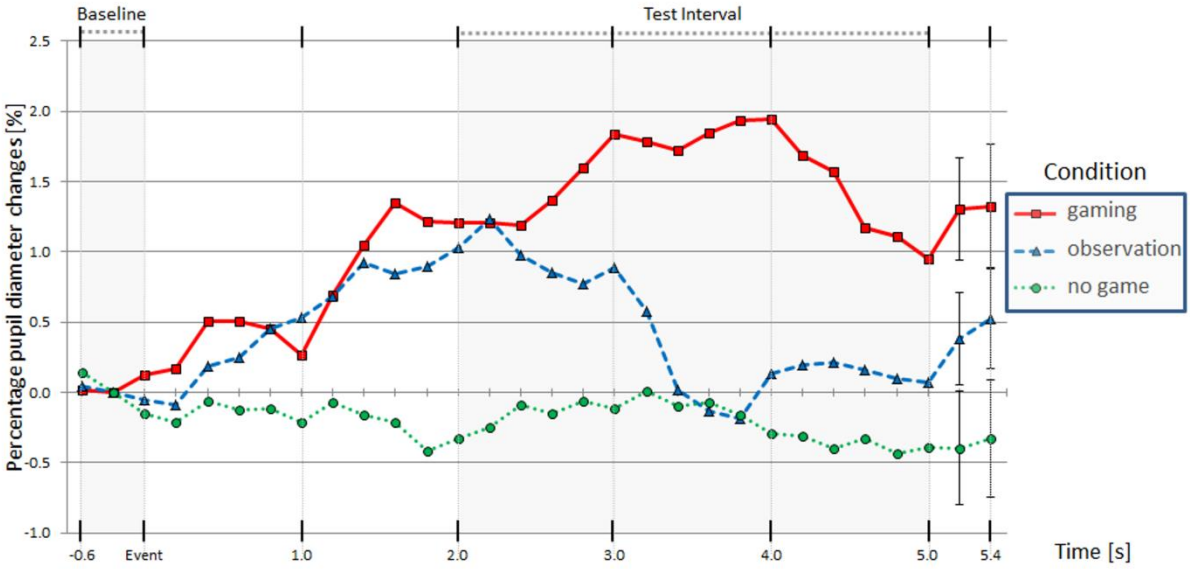

*Figure 2.*Percentage change in pupil diameter right before and after a letter is caught for each of the three conditions (gaming, observation, no game). Each data point shown in the figure is a moving average across three consecutive 200-ms intervals, relative to the interval 600 ms before the event. The graph is calculated by using all reliably measured participants in a given condition (n = 18, n = 17, and n = 20, respectively). Each data point in the figure is based on 294 to 412 measurements (averages per 200-msec interval). To indicate dispersion, the standard errors of the means in the last two intervals are shown.

3.4.2 Pupil Reactions – additional analyses

An additional analysis of the pupil responses was performed in order to assess whether the effort spent to catch a letter might have contributed to the pupil dilation via sympathetic arousal. To that effect, the pupil data in the gaming group (which actually controlled effort by pedaling harder or slowing down to catch a letter at its incoming altitude) were sorted into three kinds of cases: (1) those in which the current letter appeared roughly at the same altitude (+- 15% of the maximal range) as the previous

letter (“same”, requiring little effort), (2) those in which it appeared at a higher screen position than the previous one (“higher”, requiring extra pedaling effort), and finally (3) those cases in which it appeared at a lower screen position (“lower”, leading to a reduction in pedaling effort). Individual baseline-referenced average pupil dilations across a 3-s interval were calculated as before. For all three cases into which the gaming group data had been divided, the pupil dilations remained significantly greater - as indicated by Kruskal-Wallis H tests - than in the no-game control group: That was true whether letter appeared at the “same” altitude (M = 2.32; SD = 2.18; χ^2 = 13.14; p < 0.001), a “higher” (M = 2.47; SD = 3.11; χ^2 = 7.71; p = 0.005) , or a “lower” altitude (M = 1.67; SD = 2.39; χ^2 = 6.62; p = 0.010). The differences between the three altitude groups were not statistically significant (Friedman test: χ^2 = 0.70; p = 0.703).

To assess whether the magnitude of the pupil reaction is correlated with overall user experience, the mean pupil reaction in percent of all successfully caught letters was calculated for each member of the active gaming group (one person skipping questions and showing stereotypical response bias was excluded). Those participants exhibiting a mean pupil dilation greater than one percent were classified as “strong responders” (n = 6), those showing a pupil dilation of less than one percent were classified as “weak responders” (n = 11). It turned out that strong responders rated their motivation (t(15) = 3,725, p = 0.002), flow (t(15) = 2,772, p = 0.014), and overall user experience (t(15) = 2,387, p = 0.031) as significantly higher than weak responders did.

To check, whether the differing cognitive demands imposed by the three experimental conditions might have produced differences in tonic pupillary diameter potentially interacting with the observed phasic responses, mean pupil diameter across the entire session was calculated for each participant and condition (game [M = 4.11 mm / SD = 0.62 mm], ‘observing’ yoked control [M = 4.08 mm / SD = 0.77 mm], no game [M = 5.24 mm / SD = 0.95 mm]). While the pupil diameter is slightly larger in the ‘no-game’ control condition, very likely due to the reduced luminance of the display, across-session tonic

pupil diameter did neither correlate with ratings on the Borg scale, nor with cognitive load or the 'exhaustion' question (all $p > 0.10$).

3.5 Heart Rate and Skin Conductance.

Heart rate (HR) and skin conductance (SC) changes had been recorded as well. When phasic (HR or SC) responses were analysed as triggered by the game events, neither HR nor SC showed statistically significant patterns distinguishing the gaming and no-game control groups. Rather, both (tonic) HR and SC monotonically increased during the ten-minute training (by as much as 44 percent [HR] and 207 percent [SC], respectively), suggesting that they simply reflect the participants' physical exertion increasing over the course of the exergame. Furthermore, tonic levels or slopes of neither HR nor SC showed any systematic differences between experimental conditions. Therefore, these physiological indicators were not further analysed.

3.6 Interrelations between self-report measures.

Borg's scale of perceived exertion and the single-item rating of perceived effort were significantly correlated, $r(61) = .69$, $p < 0.001$. Furthermore, Borg's Scale was correlated with the mean HR during cycling, $r(62) = 0.45$, $p < 0.01$, and the percentage change of the HR between the first and the last 10 seconds of cycling, $r(56) = 0.31$, $p < 0.05$.

The overall user experience score and all subscales of the user experience questionnaire as well as the retrospective questions about "how much fun" the game was and to what extent it "distracted from the exercise" were all significantly correlated with each other (all $p < 0.01$). Correlations between the 6 user experience scales and the two single-item ratings (of 'fun' and 'distraction') can be seen in Table 2. That is, all scales reflecting positive perceptions were significantly correlated. The correlation between the "cognitive load" subscale, that was used to measure the challenge of the task, and the user experience score was $r(64) = .76$, $p < 0.001$. Most important for a high overall user

experience score were the subscales "flow", $r(64) = 0.86$, $p < 0.001$, and "positive emotion", $r(64) = 0.89$, $p < 0.001$.

Table 2
Pearson product-moment correlations between the subjective scales used

	Scales								
Scales	User Experience	Avoid negative emotion	Cognitive load	Positive emotion	Motivation	Flow	Arousal	"how much fun"	"distracted from the exercise"
User Experience	-								
Avoid negative emotion	.72**	-							
Cognitive load	.76**	.53**	-						
Positive emotion	.89**	.69**	.74**	-					
Motivation	.81**	.37**	.47**	.60**	-				
Flow	.86**	.45**	.52**	.67**	.81**	-			
Arousal	.82**	.52**	.51**	.66**	.65**	.67**	-		
"how much fun"	.59**	.46**	.40**	.61**	.41**	.49**	.50**	-	
"distracted from the exercise"	.62**	.52**	.59**	.48**	.43**	.53**	.54**	.40**	-

** $p < .01$

There were no significant correlations between the self-report measures, and the individual average increase in pupil diameter due to the game events, except that the latter was weakly correlated with the cognitive load subscale, $r(54) = 0.33$, $p = 0.014$: The more the participant agreed to be "pleasantly stimulated" and "not excessively challenged" the larger was his/her pupil reaction to game events.

3.7 Demographics.

To control for potential effects of age, gender, time spent gaming per week, time working with a computer and time to use a bicycle per week, these variables - taken from the questionnaire - were correlated with the pupil reaction to the game events. No (significant) correlation with any of these demographic variables was found. Time working with a computer per week was correlated with the perceived distraction by the presentation on the screen ($r(62) = 0.37$, $p = 0.003$) and the experience of flow ($r(62) = 0.34$, $p = 0.007$.

4. Discussion

4.1 General results

The main goal of the present study was to investigate whether pupil reactions constitute a suitable measure to detect emotional responses to game events. A significantly ($p < 0.05$) larger pupil diameter change in response to positive game events was observed in the active gaming group (percent pupil dilation: $M = 1.53$) than in the no-game ($M = -0.20$) control group. The comparison with the yoked control group ($M = 0.40$) merely observing a game was marginally significant ($p = 0.056$). While the active gaming group showed a mean pupil reaction more than three times as strong than the observational control group, the no-game control group exhibited no systematic pupil diameter changes. Controlling for spontaneous pupil diameter variation (no-game group) and illumination (the observational control group watched the same kinds of visual events on the screen) as well as the physical effort spent cycling, suggests that the pupil diameter variation observed may indeed reflect an emotional reaction to the game event (successfully capturing the letter) to which it was referenced. This conjecture is supported by the finding that the (retrospective) user-experience scale employed (and all of its subscales) discriminate the three gaming modes (active gaming, observation, no-game control) in much the same way as the pupil reactions do. Furthermore, a positive

correlation between the average individual pupil diameter change in response to the game events and the user experience subscale reflecting the challenging nature of the task was found. Finally, participants showing more pronounced pupil reactions to game events ('strong responders') also had a higher overall user experience, as well as elevated scores on the user experience subscales of 'motivation' and 'flow'.

4.2 Comparison with previous work

Previous studies have measured pupil reactions to pictures, brief sounds, or words, i.e. to stimuli of a relatively stationary kind. By contrast, the present study investigated pupil reactions to dynamic game events. Furthermore, in contrast to the earlier studies, in the present work a stimulus with no (a priori) emotional content was used (a letter), and the participants' emotional reactions to the stimuli were based on subjective significance, i.e. on their function in the game. Based on the classical cognitive theories sketched out in the introduction (Arnold, 1960; Lazarus, 1968; Weiner, 1986; Ortony et al., 1988), there is no reason, why seeing a letter should result in an emotional reaction at the outset of the game. But while the game proceeds, linking the letter to the points granted by catching it, the letter should become a cue to a rewarding event, and thus elicit positive emotions.

The magnitude of the pupil diameter responses obtained in the present study may be compared to earlier work: When participants in the gaming group reacted to an event, pupil diameter slowly increased after stimulus onset to its maximum value of a 2-percent change with respect to baseline and decreased after some 4 s. This time course is quite similar to what has been observed in earlier studies (Bradley, Miccoli, Escrig, & Lang, 2008; Henderson, Bradley, & Lang, 2014; Partala, Jokiniemi, & Surakka, 2000), but the magnitude of the response appears to be somewhat smaller than previously obtained with emotional sounds (Partala, et al., 2000: 5.6% to 6.8%), or pictures (Bernhardt et al., 1996: 1% to 2%; Bradley at el., 2008: 5.4%). Note, however, that the duration of the events may

also be a factor, since the game events are quite short, while stationary pictures and sounds are typically presented for 6 or 7 seconds.

4.3 Validity

To justify the conclusion that the pupil diameter reaction reflects an emotional response, several alternative explanations have to be ruled out. For example, illumination, physiologically based arousal, and effects of cognition have to be controlled. First, a (yoked) control group was used to control for effects of illumination by having an observational group see exactly the same game events as the gaming group. Second, to control for physiological arousal a task was used that demanded moderate effort (increasing pedal rate to catch a letter) as often as relaxation (decreasing pedal rate to catch a letter). Third, given that participants are mastering a challenging task, the cognitive evaluation of "being successful" is part of the generation of the emotion, not a confound, in our view.

A more serious concern might be that what distinguishes the 'gaming' group from the two control groups (observing and no-game condition) is not only the latter lacking the rewarding experience of succeeding in the game, but also not having to regulate the physical effort spent (to catch a letter) simultaneously. Therefore, it might be suspected, that the pupil reaction – in addition to whatever emotional response it might reflect - also contains some cognitive (load) component due to regulating physical activity. Four arguments can be put forward against this potential confound: First, the pupil reaction well matches the kind of curves reported in the literature for emotional reactions to visual stimuli (e.g. Bradley, Miccoli, Escrig, & Lang, 2008; Henderson, Bradley, & Lang, 2014). Second, pupil reactions based on cognitive processes studied in the literature (e.g. Laeng, Sirois, & Gredebäck, 2012; Privitera, Renninger, Carny Klein, & Aguilar, 2008; Steenbergen & Band, 2013; Verney, Granholm, & Marshall, 2004) typically peak between one and two seconds after stimulus onset and therefore do not fall into the interval of two to five seconds used in the present study. Third, the subjective post-

experimental ratings, and the questionnaire-based measures of user experience, all concur in identifying positive emotional responses in the experimental more than in the control groups. Fourth, a more fine-grained analysis of the pupil responses showed that catching letters requiring greater physical and cognitive tracking effort resulted in essentially the same response magnitude.

4.4 Limitations

Comparing the pupil diameter changes in the gaming and control groups shows, that there is a significant pupil response to the emotional game events (successfully catching letters). Questionnaire data (the user experience score and its subscales) are consistent in showing that the gaming group enjoyed the exercise most. The magnitude of the pupil dilation and "cognitive load" were also correlated. However, no statistically significant correlation – with the exception of a weak link shown when 'strong' vs. 'weak' pupil responders were contrasted – between the individual event-based pupil diameter changes and a given participant's overall user experience was found. This may be due to three reasons. First, arousal needs to be attributed to trigger an emotion (Schachter & Singer, 1962) and the most obvious cause for arousal in the present experiment is the physiological exercise. So failing to attribute the arousal to an emotion may in fact attenuate the expected correlation. Second, the repeated experience of a specific type of game event may just constitute a small portion of the overall user experience. Since the correlations of the subscales of the user experience questionnaire with the overall user experience score indicate "cognitive load" (challenge) to be just a portion of the overall experience ($r = 0.76$), diffuse experiences like "flow" ($r = 0.86$) and "positive emotion" ($r = 0.89$) may be more important for good overall user experience, than the emotion triggered by a single type of event in the game. Finally, while pupil reactions were measured repeatedly throughout the 10-minute game play, the overall user experience was only assessed once, retrospectively after the game, thereby potentially weakening the statistical link. This is to say that physiological measurement may be suitable to register

an emotional reaction to a discrete event (e.g. successfully mastering a task or watching an IAPS picture), but it is doubtful whether it can predict the overall experience in a complex game (Ravaja & Kivikangas, 2008) – especially when the game presents more reasonable alternatives (physical exercise) to attribute arousal, as it is common for exergames.

Clearly, since some 95% of the target letters were actually caught, the present investigation studied reinforcing rather than frustrating game events. It might still be interesting to check, whether pupil reactions to negative versus positive emotional game events can actually be distinguished. To that effect, a study systematically and more evenly varying success and failure in a game may be called for.

Furthermore, in contrast to presenting pictures or sounds one at a time, measuring arousal as an indicator of emotion in a digital (exer-) game is complicated for several reasons. First, the positive emotion triggered by the event is not based on the properties of the stimulus itself (here: the picture of a letter), but on the interaction with the stimulus (the experience of successfully capturing the letter), therefore the player's movement is part of the generation of the emotional event, so individually different states of motivation can cause variance in the perception of the event. Second, movement caused by the exerciser can cause data loss and artifacts in the psychophysiological measurement, as well as interfering arousal. Third, the positive emotion is due to the success of mastering a challenge (here: capturing the letters), therefore aspects of gameplay have to be considered. It is not possible to extend the time between events in an arbitrary way (e.g. presenting letters at a much slower rate will avoid accumulation of arousal, but simultaneously it will reduce the challenge and thereby the elicitation of the emotion itself).

Finally, the measurement of heart rate (HR) using a chest belt is a limitation of the present study, both in terms of sensitivity and reliability. Future studies might use more

fine-grained ECG recordings to assess the usefulness of measuring cardiovascular responses for detecting emotional reactions to dynamic events during an exercise game.

4.5 Potential applications

Two applications of the present research are envisioned: Scientific and health-related. In studying emotion elicitation in games, for example, the present approach may help to identify emotional events in dynamic scenarios with pupil responses constituting one of their psychophysiological correlates. By studying the influence of specific events on overall user experience, a better understanding of complex psychological constructs like flow, immersion or motivation might be achieved. For health-related applications like gamified exercise programs, the present study suggests to use pupillometry as yet another sensor to assess user experience, and thereby to better allocate motivational incentives in rehabilitation / exercise programs, or to optimize parameters (e.g. adequate challenge) to maximize the health impact of these applications.

4.6 Conclusion

The main outcome of the present study is that measuring changes in pupil diameter constitutes a suitable method to detect players' (positive) emotional responses to discrete events in a digital game. In particular, successful interaction with the game (mastering a challenging task) was studied. The pupil dilation, recorded some 2-5 s after the crucial game event, was shown to distinguish the type of game played (active, observational or a task devoid of gaming challenges), and correlated with the cognitive load subscale of the user experience questionnaire. This suggests that the pupil reaction may indeed reflect the positive feeling of mastering a challenging task. So the pupil reaction is not only suited to measure the emotional content of comparatively 'static' stimuli like pictures or short sounds, but it can also be used to detect (positive) emotions triggered by dynamic events, thus making it a unobtrusive candidate to evaluate the emotional impact of health/exercise applications or (serious) games in general.

5. References

Allison, B. Z., & Polich, J. (2008). Workload assessment of computer gaming using a single-stimulus event-related potential paradigm. *Biological Psychology, 77*, 277-283. https://doi.org/10.1016/j.biopsycho.2007.10.014

Arnold, M. B. (1960). *Emotion and personality.* (Vol I). New York, NY: Columbia University Press.

Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: Adaptive gain and optimal performance. *Annual Review of Neuroscience 28*, 403-450. https://doi.org/10.1146/annurev.neuro.28.061604.135709

Barlett, C. P., & Rodeheffer, C. (2009). Effects of realism on extended violent and nonviolent video game play on aggressive thoughts, feelings, and physiological arousal. *Aggressive Behavior, 35*, 213-224. https://doi.org/10.1002/ab.20279

Baumgartner, T., Esslen, M., & Jäncke, L. (2006). From emotion perception to emotion experience: Emotions evoked by pictures and classical music. *International Journal of Psychology, 60*, 34-43. https://doi.org/10.1016/j.ijpsycho.2005.04.007

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin, 91*, 276–292. https://doi.org/10.1037/0033-2909.91.2.276

Bernhardt, P. C., Dabbs, J. M., & Riad, J. K. (1996). Pupillometry system for use in social psychology. *Behavior Research Methods, Instruments. & Computers, 28(1),* 61-66. https://doi.org/10.3758/BF03203637

Borg, G. (2004). Anstrengungsempfinden und körperliche Aktivität. *Deutsches Ärzteblatt, 15*, 16-21.

Borg, G., & Borg, E. (2001). A new generation of scaling methods: Level-anchored ratio scaling. *Psychologica, 28*, 15-45.

Bradley, M. M., & Lang, P. J. (1999a). *International affective digitized sounds (IADS): stimuli, instruction manual and affective ratings* (Tech. Rep. B-2). Gainesville, FL: University of Florida, the Center for Research in Psychophysiology.

Bradley, M. M., & Lang, P. J. (1999b). *Affective norms for English words (ANEW): Instruction manual and affective ratings* (Tech. Rep. No. C-1). Gainesville, FL: University of Florida, the Center for Research in Psychophysiology.

Bradley, M. M., & Lang, P. J. (2000). Affective reactions to acoustic stimuli. *Psychophysiology, 37,* 204–215. https://doi.org/10.1111/1469-8986.3720204

Bradley, M. M., & Lang, P. J. (2007). *The International Affective Digitized Sounds (2nd Edition; IADS-2): Affective ratings of sounds and instruction manual* (Tech. Rep. B-3). Gainesville, FL: University of Florida, the Center for Research in Psychophysiology.

Bradley M. M., & Lang P. J. (2009). Emotion and Motivation. In J.T. Cacioppo, L. G. Tassinary, and G. Berntson (Eds.), *Handbook of Psychophysiology* (3rd ed.) (pp. 581–607). New York, NY: Cambridge University Press.

Bradley, M. M., & Lang, P. J. (2015). Memory, emotion, and pupil diameter: Repetition of natural scenes. *Psychophysiology, 52*(9), 1186–1193. https://doi.org/10.1111/psyp.12442

Bradley, M. M., Codispoti, M., Cuthbert, B. N., & Lang, P. J. (2001). Emotion and motivation I: Defensive and appetitive reactions in picture processing. *Emotion, 1*(3), 276-298. https://doi.org/10.1037/1528-3542.1.3.276

Bradley, M. M., Miccoli, L., Escrig, M. A., & Lang, P. J. (2008). The pupil as a measure of emotional arousal and automatic activation. *Psychophysiology, 45,* 602-607. https://doi.org/10.1111/j.1469-8986.2008.00654.x

Braithwaite, J. J., Watson, D. G., Jones, R., & Rowe, M. (2013). A guide for analysing electrodermal activity (EDA) & skin conductance responses (SCRs) for psychological experiments. *Psychophysiology, 49,* 1017-1034.

Breuer, J., & Bente, G. (2010). *Why so serious?* On the relation of serious games and learning. *Journal for Computer Game Culture, 4(1),* 7-24.

Chen, S., Epps, E. & Chen, F. (2011). A comparison of four methods for cognitive load measurement. *Proceedings of the OzCHI11*, Australia, 76-79. http://doi.org/10.1145/2071536.2071547

Codispoti, M., Bradley, M. M., & Lang, P. J. (2001). Affective reaction to briefly presented pictures. *Psychophysiology, 38,* 474-478. https://doi.org/10.1017/S004857720198028X

Cuthbert, B. N., Schupp, H. T., Bradley, M. M., Birbaumer, N., & Lang, P., L. (2000). Brain potential in affective picture processing: Covariation with autonomic arousal and affective report. *Biological Psychology, 52,* 95-111. https://doi.org/10.1016/S0301-0511(99)00044-7

Dörner, R., Göbel, S., Effelsberg, W., & Wiemeyer, J. (Eds.)(2016). *Serious Games - Foundations, Concepts and Practice*. Cham, Switzerland: Springer International Publishing

Ellermeier, W., & Westphal, W. (1995). Gender differences in pain ratings and pupil reactions to painful pressure stimuli. *Pain, 61,* 435-39. https://doi.org/10.1016/0304-3959(94)00203-Q

Goldwater, B. C. (1972). Psychological significance of pupillary movements. *Psychological Bulletin, 77*, 340–355. https://doi.org/10.1037/h0032456

Göbel, S., Gutjahr, M. O. & Hardy, S. (2013). Evaluation of serious games. In: Bredl, K. & Bösche, W. (Eds.), *Serious Digital Games, MUVE and MMORPG in Adult Education and Health Care: Research, Reviews, Case Studies, and Lessons Learned* (pp. 105-116).Hershey, PA: IGI Global.

Göbel, S., Hardy, S., Wendel, V., Mehm, F. & Steinmetz, R. (2010). Serious Games for Health - Personalized Exergames. In: *Proceedings of ACM Multimedia 2010*, 1663-1666.

Greenwald, A. G. (1976). Within-subjects designs: To use or not to use?. *Psychological Bulletin, 83(2),* 314. https://doi.org/10.1037/0033-2909.83.2.314

Hardy, S., Dutz, T., Wiemeyer, J., Göbel, S., & Steinmetz, R. (2015). Framework for personalized and adaptive game-based training programs in health sport. *Multimedia Tools and Applications*, *74* (14), 5289-5311. https://doi.org/10.1007/s11042-014-2009-z

Henderson, R. Bradley, M. M., & Lang, P. J. (2014). Modulation of the initial light reflex during affective picture viewing. *Psychophysiology*, *51*, 815–818. https://doi.org/10.1111/psyp.12236

Hess, E. H., & Polt, J. M. (1960). Pupil size as related to interest value of visual stimuli. *Science*, 132, 349–350. https://doi.org/10.1126/science.132.3423.349

Janisse, M. P. (1974). Pupil size, affect and exposure frequency. *Social Behavior and Personality, 2(*2), 125-146. https://doi.org/10.2224/sbp.1974.2.2.125

Kahneman, D., & Beatty, J. (1966). Pupil diameter and load on memory. *Science, 154(3756),* 1583-1585. https://doi.org/10.1126/science.154.3756.1583

Kirsh, S. J. (2006). *Children, adolescents and media violence: A critical look at the research.* Thousand Oaks, CA: Sage.

Kivikangas, J. M., & Ravaja, N. (2013). Emotional responses to victory and defeat as a function of opponent. *Affective Computing, IEEE Transactions on affective computing, 4*(2), 173-182. https://doi.org/10.1109/T-AFFC.2013.12

Kivikangas, J. M., Chanel, G., Cowley, B., Ekman, I., Salminen, M. Jarvela S., & Ravaja, N. (2013). A review of the use of psychophysiological methods in game research. *Journal of Gaming and Virtual Worlds, 3*(3), 181-199. https://doi.org/10.1386/jgvw.3.3.181_1

Laeng, B., Sirois, S., & Gredebäck, G. (2012). Pupillometry: A window to the preconscious? *Perspectives on Psychological Science, 7*(1), 18–27. https://doi.org/10.1177%2F1745691611427305

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (2005). *International affective picture system (IAPS): Affective ratings of pictures and instruction manual* (Tech. Rep. No. A-6.). USA: University of Florida.

Lazarus, R. S. (1968). Emotion and adaption: Conceptual and empirical relations. In W. J. Arnold (Ed.), *Nebraska symposium on motivation* (Vol. 16, pp. 175-266). Lincoln, NE: University of Nebraska Press.

Libby, W. L., Lacey, B. C., & Lacey, J. I. (1973). Pupillary and cardiac activity during visual attention. *Psychophysiology, 10*, 270–294. https://doi.org/10.1111/j.1469-8986.1973.tb00526.x

Mandryk, R., Inkpen, K., & Calvert, T. W. (2006). Using psychophysiological techniques to measure user experience with entertainment technologies. *Behaviour and Information Technology, 25(2),* 1-16. https://doi.org/10.1080/01449290500331156

Mathôt, S. (2018). Pupillometry: Psychology, Physiology, and Function. *Journal of Cognition, 1(1)*: 16, 1–23. http://doi.org/10.5334/joc.18

Mudd, S., Conway, C. G., & Schindler, D. E. (1990). The eye as music critic: Pupil response and verbal preferences. *Studia Psychologica*, *32*, 23–30.

Müller, M. M., Keil, A., Gruber, T., & Elbert, T. (1999). Processing of affective pictures modulates right-hemispheric gamma band EEG activity. *Clinical Neurophysiology, 110*, 1913-1920. https://doi.org/10.1016/S1388-2457(99)00151-0

Nacke, L. E. (2009). *Affective ludology: scientific measurement of user experience in interactive entertainment*. Unpublished doctoral dissertation, Blekinge Institute of Technology - Karlskrona, Sweden.

Nakamura, J., & Csikszentmihalyi, M. (2002). The concept of flow. In C. R. Snyder & S. J. Lopez (Eds.), *Handbook of positive psychology* (pp. 89–105). New York, NY: Oxford University Press.

Ortony, A., Clore, G. L., & Collins, A. (1998). *The cognitive structure of emotions.* Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511571299

Parrott, G. W. (2004). The nature of emotion. In M. B. Brewer & M. Hewstone (Eds.), *Emotion and Motivation* (pp. 5-20). Oxford, UK: Blackwell.

Partala, T., & Surakka, V. (2003). Pupil size variation as an indication of affective processing. *International Journal of Human-Computer Studies, 59,* 185-198. https://doi.org/10.1016/S1071-5819(03)00017-X

Partala, T., Jokiniemi, M., & Surakka, V. (2000). Pupillary responses to emotionally provocative stimuli. *Proceedings of the 2000 symposium on Eye tracking research & applications, New York, NY (pp.* 123-129). http://doi.org/10.1145/355017.355042

Potter, R. F., & Bolls, P. D. (Eds.). (2012). *Psychophysiological measurement and meaning: Cognitive and emotional processing of media.* New York, NY: Routledge. http://doi.org/10.4324/9780203181027

Privitera, C. M., Renninger, L. W., Carney, T., Klein, S., & Aguilar, M. (2010). Pupil dilation during visual target detection. *Journal of Vision, 10*(10), 1-14. https://doi.org/10.1167/10.10.3

Ravaja, N., & Kivikangas, J. M. (2008). Psychophysiology of digital game playing: The relationship of self-reported emotions with phasic physiological responses. In A.J. Spink, M.R. Ballintijn, N.D. Bogers, F. Grieco, L.W.S. Loijens, L.P.J.J. Noldus, G. Smit, and P.H. Zimmerman (Eds.), *Proceedings of Measuring Behavior, Maastricht, The Netherlands*, (pp. 89-99).

Ravaja, N., Saari, T., Salminen, M., Laarni, J., & Kallinen, K. (2006). Phasic emotional reactions to video game events: A psychophysiological investigation. *Media Psychology, 8*, 343-367. https://doi.org/10.1207/s1532785xmep0804_2

Reisenzein, R. Meyer, W.-U., & Schützwohl, A. (2003). Einführung in die Emotionspsychologie. Kognitive Emotionspsychologie. Bern, Switzerland: Verlag Hans Huber.

Salmine, M., & Ravaja (2007). Oscillatory brain responses evoked by video game events. The case of super monkey ball 2**.** *CyberPsychology & Behavior, 10*(3), 1-9. https://doi.org/10.1089/cpb.2006.9947

Schachter, S., & Singer, J. E. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological Review, 69*(5), 379-399. https://doi.org/10.1037/h0046234

Schmierbach, M., Chung, M. Y., Wu, M., & Kim, K. (2014*).* No one likes to lose. The effect of game difficulty on competency, flow, and enjoyment. *Journal of Media Psychology, 26,*105-110. https://doi.org/10.1027/1864-1105/a000120

Seligman, M. E., & Maier, S. F. (1967). Failure to escape traumatic shock. *Journal of experimental psychology*, *74*(1), 1. https://doi.org/10.1037/h0024514

Slater, M., Angus, A., Davison, A., Swapp, D., Guger, C., Barker, C., Pistang, N., & Sanchez-Vives, M. (2006). A virtual reprise of the Stanley Milgram obedience experiments. *PloS ONE 1(1):* e39. https://doi.org/10.1371/journal.pone.0000039

Snowden, R. J., O'Farrell, K. R., Burley, D., Erichsen, J. T., Newton, N. V., & Gray, N. S. (2016). The pupil's response to affective pictures: Role of image duration, habituation, and viewing mode. *Psychophysiology, 53,* 1217-1223. https://doi.org/10.1111/psyp.12668

Steinhauer, S. R., Siegle, G. J., Condray, J., & Pless, M. (2004). Sympathetic and parasympathetic innervation of pupillary dilation during sustained processing. *International Journal of Psychophysiology*, *53,* 77–86. https://doi.org/10.1016/j.ijpsycho.2003.12.005

Sweetser, P., & Wyeth, P. (2005). GameFlow: A model for evaluating player enjoyment in games. *Computers in Entertainment, 3* (3), 1-24.

van Steenbergen, H., & Band, G., P., H. (2013). Pupil dilation in the Simon task as a marker of conflict processing. *Frontiers in Human Neuroscience, 7,* 1-11. https://doi.org/10.3389/fnhum.2013.00215

van Steenbergen, H., Band, G. P. H., & Hommel, B. (2011). Threat but notarousal narrows attention: Evidence from pupil dilation and saccade control. *Frontiers in Psychology, 2,* 1-5. https://doi.org/10.3389/fpsyg.2011.00281

Verney, S. P., Granholm, E., & Marshall, S. P. (2004). Pupillary responses on the visual backward masking task reflect general cognitive ability. *International Journal of Psychophysiology, 52*(1), 23-36. https://doi.org/10.1016/j.ijpsycho.2003.12.003

Võ, M. L.-H., Jacobs, A. M., Kuchinke, L., Hofmann, M., Conrad, M., Schacht, A., & Hutzler (2008). The coupling of emotion and cognition in the eye: Introducing the pupil old/new effect. *Psychophysiology, 45*, 130-140. https://doi.org/10.1111/j.1469-8986.2008.00745.x

Weber, Tamborini, Westcott-Baker, & Kantor (2009). Theorizing flow and media enjoyment as a cognitive synchronization of attentional and reward networks. *Communication Theory, 19*, 397-422. https://doi.org/10.1111/j.1468-2885.2009.01352.x

Weiner, B. (1986). *An attributional theory of motivation and emotion*. New York, YN: Springer. http://doi.org/10.1007/978-1-4612-4948-1

Wiemeyer, J., Kickmeier-Rust, M., & Steiner, C.M. (2016). Performance assessment in serious games. In R. Dörner, S. Göbel, W. Effelsberg & J. Wiemeyer (Eds.), *Serious Games - Foundations, Concepts and Practice* (pp.273-302). Cham, Switzerland: Springer International. https://doi.org/10.1007/978-3-319-40612-1_10

Wiemeyer, J., Nacke, N., Moser, C, & Mueller, F. (2016). Player experience. In R. Dörner, S. Göbel, W. Effelsberg & J. Wiemeyer (Eds.), *Serious Games - Foundations, Concepts and Practice* (pp.243-271). Cham, Switzerland: Springer International. https://doi.org/10.1007/978-3-319-40612-1_9

Xu, J., Wang, Y., Chen, F., Choi, H., Li, G., Chen, S., & Hussain, S. (2011). Pupillary response based cognitive workload index under luminance and emotional changes. *Proceedings of the ACM CHI2011*, Vancouver, Canada (pp. 1627-1632). https://doi.org/10.1007/978-3-642-23771-3_14

Zénon, A., Sidibé, M., & Olivier, E. (2014). Pupil size variations correlate with physical effort perception. *Frontiers in behavioral Neuroscience*, *8*, 286. https://doi.org/10.3389/fnbeh.2014.00286

Author Note

This study was supported by an internal start-up grant for interdisciplinary research from Technische Universität Darmstadt (FiF project ‘SG4 health’).

Appendix

a. User Experience Questionnaire as used in the study (German version) and its published English version (Göbel, Gutjahr, & Hardy, 2013).

[NE] 1. Die Bildschirmdarstellung hat Langeweile vermieden.

[NE] 1. The game avoided boredom.

[NE] 2. Die Bildschirmdarstellung hat Frustration vermieden.

[NE] 2. The game avoided frustration.

[NE] 3. Ich habe mich nicht über die Bildschirmdarstellung geärgert.

[NE] 3. The game only sometimes made me angry.

[CL] 4. Die Bildschirmdarstellung hat mich angenehm gefordert.

[CL] 4. The game challenged me in a pleasant way.

[CL] 5. Die Bildschirmdarstellung hat meine Fantasie angeregt.

[CL] 5. The story engaged my fantasy.

[CL] 6. Ich war durch die Aufgaben und Möglichkeiten nicht überfordert.

[CL] 6. I was able to keep track of tasks, impressions, information and possibilities of the game and was neither overstrained nor overloaded.

[PE] 7. Die Bildschirmdarstellung hat Spaß gemacht.

[PE] 7. The game was fun.

[PE] 8. Die Bildschirmdarstellung gab mir das Gefühl eigenbestimmt und kompetent zu sein.

[PE] 8. The game made me feel self-determined and competent.

[PE] 9. Ich fand die Bildschirmdarstellung ästhetisch ansprechend gestaltet.

[PE] 9. I found the game's design to be aesthetically pleasing.

[MO] 10. Die Bildschirmdarstellung war mitunter so einnehmend, dass ich unbedingt wissen wollte, wie es weiter geht.

[MO] 10. The game was at times so engaging that I had the need to know how it continued.

[MO] 11. Einen Entwicklungsprozess festzustellen motivierte mich stark weiter zu machen.

[MO] 11. Realizing a process of progression strongly motivated me to continue playing.

[MO] 12. Teilweise spielte ich nur noch um des Spieles willen.

[MO] 12. At times I played only for the sake of playing.

[FL] 13. Die Bildschirmdarstellung war so spannend, dass sie meine ganze Aufmerksamkeit beim.Spielen auf sich zog.

[FL] 13. The game was so exciting that it captured my whole attention during play.

[FL] 14. Die Bildschirmdarstellung war so interessant, dass ich gar nicht merkte, wie schnell die Zeit vergeht.

[FL] 14. The game was so interesting that I lost all track of time.

[FL] 15. An manchen Stellen war die Bildschirmanzeige so fesselnd, dass ich vollkommen vom ihr eingenommen wurde.

[FL] 15. At times the game was so enthralling that I was completely engaged in the game.

[AR] 16. Manchmal war ich im Nachhinein sehr erleichtert, da ich ein Scheitern befürchtete.

[AR] 16. After some points of the game I was very relieved since I had expected a failure.

[AR] 17. Ich merkte, dass ich teilweise stark emotional beteiligt war (Spannung, Trauer, Erleichterung, Freude, Wut).

[AR] 17. I noticed that I was at times strongly emotionally involved (excitement, sadness, relief, joy, anger).

[AR] 18. Ich fühlte mich durch die Bildschirmdarstellung in einen angenehmen Zustand versetzt.

[AR] 18. I was in a pleasant state due to playing the game.

Scales: avoiding negative emotion [NE], cognitive load [CL], positive emotion [PE], motivation [MO], flow [FL], Arousal [AR].

Range: From 1 (not at all) to 10 (extremely).

Table 1

Organisation of the game phases

Schedule 1				*Schedule 2*			
Phase	Duration [in s]	interval betw. letters [in s]	Number of letters	Phase	Duration [in s]	interval betw. Letters [in s]	Number of letters
break	10	-	-	break	10	-	-
measurement	65	5.50 - 7.50	10	playing	125	1.75 - 4.75	39
playing	125	1.75 - 4.75	39	measurement	65	5.50 - 7.50	10
break	10	-	-	break	10	-	-
measurement	65	5.50 - 7.50	10	playing	125	1.75 - 4.75	39
playing	125	1.75 - 4.75	39	measurement	65	5.50 - 7.50	10
break	10	-	-	break	10	-	-
measurement	65	5.50 - 7.50	10	playing	125	1.75 - 4.75	39
playing	125	1.75 - 4.75	39	measurement	65	5.50 - 7.50	10