Pupil Dilation Fulfills the Requirements for Dynamic Difficulty Adjustment in Gaming on the Example of Pong

Christoph Strauch
christoph.strauch@uni-ulm.de
General Psychology, Ulm University
Ulm, Germany

Michael Barthelmäs
michael.barthelmaes@uni-ulm.de
Social Psychology, Ulm University
Ulm, Germany

Elisa Altgassen
elisa.altgassen@uni-ulm.de
General Psychology, Ulm University
Ulm, Germany

Anke Huckauf
anke.huckauf@uni-ulm.de
General Psychology, Ulm University
Albert-Einstein-Allee 47, Ulm, Germany

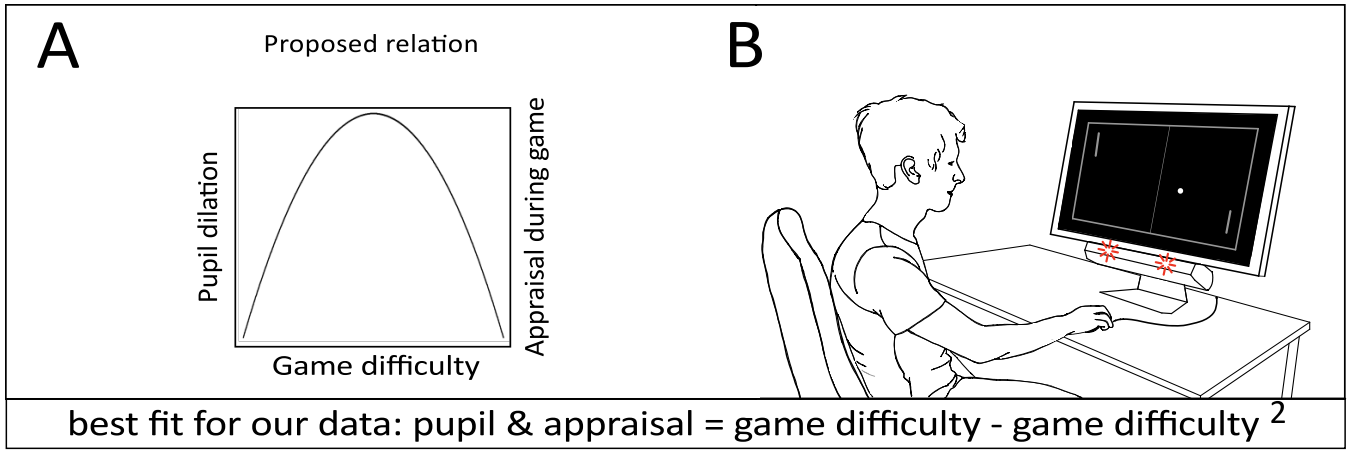

Figure 1: Assessing the player's level of challenge and appraisal during gaming is the prerequisite for dynamic game difficulty adjustment. A) We assumed that pupil dilation and appraisal would show an inverted u-shaped function to increasing game difficulty. Challenge and appraisal while B) playing 'Pong' could reliably be traced from pupil diameter. In two experiments pupil diameter followed the proposed relation: Pupil size scales with difficulty and appraisal until best appraisal and maximum pupil dilation is reached. Further increasing difficulty decreases both appraisal and pupil dilation.

ABSTRACT

When games are too easy or too difficult, they are likely to be experienced as unpleasant. Therefore, identifying the ideal level of game difficulty is crucial for providing players with a positive experience during gaming. Performance data is typically used to determine how challenged a player is; however, this information is not always available. Pupil diameter has recently been suggested as a continuous option for tracking gaming appraisal. In this paper, we describe two experiments with in total 55 participants playing 'Pong' under four levels of difficulty. Difficulty was manipulated via ball-speed (Experiment 1) and racket-size (Experiment 2) ranging from under- to overload. Pupil dilation and appraisal were maximal under medium difficulty (compared to easy and hard levels). These findings demonstrate the usefulness of pupil diameter as basis for psychophysiologically dynamic difficulty adjustment as it is sensitive both to under- and overload, hence underlining pupil dilation's potential value for user-adaptive interfaces in general.

ETRA '20 Adjunct, June 2–5, 2020, Stuttgart, Germany

ACM ISBN 978-1-4503-7135-3/20/06…$15.00
https://doi.org/10.1145/3379157.3388934

CCS CONCEPTS

• **Applied computing** → **Computer games**; Psychology; • **Software and its engineering** → *Interactive Games.*

KEYWORDS

Pupil dilation, dynamic difficulty adjustment, video games, adaptive systems

ACM Reference Format:
Christoph Strauch, Michael Barthelmäs, Elisa Altgassen, and Anke Huckauf. 2020. Pupil Dilation Fulfills the Requirements for Dynamic Difficulty Adjustment in Gaming on the Example of Pong. In *Symposium on Eye Tracking*

Research and Applications (ETRA '20 Adjunct), June 2–5, 2020, Stuttgart, Germany. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3379157.3388934

1 INTRODUCTION

Shaping games neither too easy nor too hard is still a difficult task for developers, as players can hold differing levels of skills, ranging from pure novices to experienced players. One option to establish an optimal level of difficulty is to monitor players' performance and to adjust the level of difficulty based on performance indices (e.g., reaction-times or scores) [Tan et al. 2011, e.g.]. However, this approach to *dynamic difficulty adjustment* was only feasible if performance-indices are available. For example during strategic considerations or before so-called quick-time events, players typically do not provide sufficient input to determine an optimal level of difficulty. In this paper, we investigate and discuss pupil dilation as an alternative psychophysiological measure to determine how challenged a player is. Pupil dilation is especially promising, since it provides valuable information when no performance data is available while its assessment is affordable, non-intrusive, and delivers a continuous signal during gaming.

Beyond gaming, investigating this rather specific research question may give fruitful insights into the suitedness of pupil dilation for user-adaptive interfaces on a more general level. Hence, findings from this investigation might contribute to the optimization of different types of interaction. For example in the context of gaze-based interaction, pupil dilation has been proposed to serve as a source of information for adaptively adjusting dwell-times for specific selections [Strauch et al. 2017, e.g.]. In addition, pupil dilation (together with other measures) could be used for adapting dwell-times/interfaces for several selections/commands as lower frequent information, if pupil dilation may indeed help identify an ideal level of challenge. Hence, pupil dilation might as well contribute to assistive technology, as learning environments or other applications that might benefit from an online adjustment to detected user appraisal.

2 RELATED WORK

2.1 Dynamic difficulty adjustment

Players often have the possibility to set a - usually static - difficulty in the beginning of a game to avoid the frustration of underload or overload. This strategy is functional when starting a game, however, the player's skill will improve over time, eventually creating a mismatch of skills and demands [Hunicke and Chapman 2004]. The ideal game difficulty should thus adapt dynamically to keep players in a match of ability and demands, facilitating performance and enhancing appraisal during gaming. This idea, known as dynamic difficulty adjustment [Hunicke 2005], is reflected in several theoretical concepts, ranging from Yerkes-Dodson law [Yerkes and Dodson 1908] over flow theory [Nakamura and Csikszentmihalyi 2014] to immersion [Nacke and Lindley 2008].

Following Hunicke [Hunicke and Chapman 2004], such adaptively adjusting systems need to (1) assess the need for adjustment (2) determine which changes are needed and (3) perform changes smoothly. Usually, behavioural measures (i.e., input by the player) are used to assess the need for difficulty adjustment, however, input revealing the current level of perceived challenge is not always available. For example, strategic long-term considerations can hardly be traced from user input with effects of a chosen strategy become manifest not until minutes or hours later. This implies the need for an additional way of assessing gaming experience. In the following paragraphs we discuss the promising role of physiological measures for dynamic difficulty adjustment.

2.2 Psychophysiologically adaptive gameplay

Psychophysiological measurements are potentially relevant in the context of dynamic difficulty adjustment, as they can be assessed non-intrusively while providing a continuous source of information. Most striking, this continuous signal is available independent from the availability of performance indices. The theoretical basis behind the idea to use physiological signals for dynamic difficulty adjustment may be traced back to Yerkes and Dodson [Yerkes and Dodson 1908]. Since the very beginning of the 20th century, it is well established that a medium level of physiological arousal is beneficial for task performance, whereas too high or low levels of arousal impede performance (as long as the task is not extremely simple). Conversely, the highest degree of task performance is typically associated with a medium level of arousal. Hence, a measurement that can reliably detect the level of arousal during an ongoing activity could serve as a proxy for the (mis-)match of ability and demands. A state of fit is typically associated with an optimal subjective experience [Nacke and Lindley 2008; Nakamura and Csikszentmihalyi 2014].

To qualify for an information source in dynamic difficulty adjustment, a potential signal must be able to detect both, under- and overload. The identification of both states of mismatch is necessary to keep the player involved by suitable means (i.e., by increasing or decreasing difficulty) [Hunicke and Chapman 2004]. In the following paragraph, we shortly introduce existing research on physiologically altered gameplay by example.

2.2.1 Brain based dynamic difficulty adjustment. Most existing research focuses on measures of the central nervous system, mainly EEG, that are for example employed in treating attention related disorders by using neurofeedback steered games ([Pope and Bogart 1994], see [Lofthouse et al. 2012] for a review). For dynamic difficulty adjustment, the usefulness of EEG measures for determining load induced by three different levels of Tetris has been demonstrated [Ewing et al. 2016]. However, EEG systems also come with practical disadvantages: First, direct contact of sensors with the scalp is required, implying a time-consuming preparation of each user by an experienced EEG-examiner. Second, due to this preparation not only acquisition, but also operating costs are relatively high. Third, even though real-time data processing is possible, processing of EEG data is complex as the raw signal is prone to various types of artifacts [Urigüen and Garcia-Zapirain 2015].

Functional near-infrared spectroscopy (fnirs) represents a more user-friendly central nervous measure, as no direct scalp contact is required. Fnirs has been demonstrated to yield useful information for dynamic difficulty adjustment. E.g., [Yuksel et al. 2016] demonstrate a system where players were practicing piano pieces by Bach and were tracked with fnirs. Whenever the estimated workload

fell below a threshold, difficulty was increased dynamically. Participants rated the employed system as favourable and were satisfied with the timing of level changes [Yuksel et al. 2016].

Besides central nervous measures, also peripheral physiological variables may provide useful information. In what follows, we give an overview over existing candidates and research using this idea.

2.2.2 Peripheral physiology based dynamic difficulty adjustment. So far, most research focuses on indicators such as heart rate, skin conductance or electromyography, whereas investigations using pupillometry are yet rare, despite being considered promising [Kivikangas et al. 2011]. Self-reported game experience correlates with peripheral measures, such as electrodermal activity or heart rate [Drachen et al. 2010, e.g.] (see [Kivikangas et al. 2011] for a comprehensive review).

As an example, when playing Tetris, heart rate variability was demonstrated to show an inverted u-shape with variability being lowest when demands and abilities matched, still, overload could only be distinguished on a trend level [Keller et al. 2011]. [Chanel et al. 2008] measured different peripheral signals, including skin conductance, heart rate, blood pressure, respiration, as well as temperature while participants were playing Tetris with changing difficulty. Self-reported valence was found to show an inverted u-shape, indicating lower ratings for boring or too challenging conditions. For the psychophysiological indicators however, only a linear relation to game difficulty was found.

Among those measures investigated more scarcely, pupil dilation might prove useful and usable for several reasons: First, the latency of the signal is comparably low [Mathôt 2018], allowing a near to real-time adaption of gaming difficulty. Second, processing and interpretation of the measured data is comparably easy. Third, pupil dilation can be tracked remotely using low-cost camera systems, already today [Hansen et al. 2018, e.g]. Further, eye trackers are becoming more and more ubiquitous, e.g., in VR devices. Hence, no extra hardware would be needed for pupil adaptive gameplay. First empirical evidence underlines the potential of pupil dilation in this context: When playing a serious game during physical activity, pupil diameter was found to be a better indicator of emotional events than heart rate and skin conductance [Gutjahr and Wiemeyer 2019]. Comparing EEG, heart rate, ECG, measures all needing direct sensor contact and remote pupil dilation, [Köles et al. 2015] found pupil dilation to be most indicative of changes in task load while playing Tetris in different levels of difficulty.

Summing up, in the few existing investigations comparing pupil diameter with other peripheral indicators of game experience, pupil dilation seems to be a promising, but understudied, candidate. In the next section, we will focus on the properties of this measure.

2.3 Pupillometry

Besides changes due to variations in brightness or accommodation, pupil diameter is sensible to changes in arousal. It has been shown that variations in Locus Coeruleus and the associated norepinephrine system tightly covary with pupil diameter in a fast temporal manner [Murphy et al. 2014, e.g.]. A variety of processes is investigated using pupillometry, spanning topics such as emotional activation [Ehlers et al. 2016, e.g.], decision making [Strauch et al. 2018, 2020, e.g.], and cognitive load [Granholm et al. 1996; Hess and Polt 1964, e.g.].

2.3.1 Pupil dilation as indicator of cognitive load. Among the first to relate pupil diameter with task difficulty in English speaking literature, [Hess and Polt 1964] demonstrated increasing pupil diameter with increasing task difficulty in multiplication tasks. It has later been demonstrated that this increase in pupil diameter is not linearly connected to task difficulty, as overly difficult tasks will overload and cause participants to give up [Granholm et al. 1996].

2.3.2 Pupil dilation as indicator of game difficulty. Investigating pupil dilation during gaming, [Gutjahr and Wiemeyer 2019] tested participants on an ergometer: Participants were playing a game actively versus participants observing the same game events passively or cycling only (control conditions). Pupil dilated more while playing compared to the control conditions. While this investigation focused on emotional responses to game events [Gutjahr and Wiemeyer 2019], one may assume that the effects of mental load and emotional activation are very often intertwined in pupil dilation, even more so in video games. In [Köles et al. 2015], pupil dilation was found to scale linearly with game difficulty. However, contrary to our assumption based on [Nakamura and Csikszentmihalyi 2014; Yerkes and Dodson 1908, e.g.], no inverted u-shape was reported. Instead, pupil diameter varied considerably between difficulty conditions and suggested a linear increase in pupil diameter going along with an increase in gaming difficulty [Köles et al. 2015].

This linear rather than quadratic relation of pupil dilation to gaming difficulty might result from suboptimally chosen levels of difficulty, i.e., the full spectrum from under- to overload might not have been tested. Therefore, the question arises whether pupil diameter may indeed be useful for future usage in physiologically adaptive games, i.e., whether it may provide information not only about under- but also optimal- and overload.

2.4 Research questions

Investigating whether pupil diameter indicates subjective gaming difficulty also in a quadratic fashion, two experiments were run consecutively. The optimally chosen levels of gaming difficulty should include too easy, medium, and too difficult levels. Furthermore, we expected pupil dilation to be related to self-reported gaming appraisal.

In this work, we tested whether pupil dilation could serve as a physiological proxy to detect (sub-) optimal arousal while playing.

3 RESEARCH METHODS

Two experiments were performed and analyzed using a within subject design. Ensuring that game difficulty was the main driver of results, game difficulty was operationalized in two different ways, while the general setting of the game remained the same across the experiments. Experiment 2 can thus be considered as conceptual replication, due to the iterative process and lessons learned from Experiment 1, additional measures were included.

All participants gave written informed consent prior to their participation in the experiments. Furthermore, participants gave demographic information (gender and age). Subsequently, the experimenter checked that pupil diameter could be tracked correctly

(as indicated by the tracker software). Participants were told that they would be playing 'Pong' against a computer four times. After each phase of playing, participants were asked to fill in items regarding their appraisal of the foregoing phase of playing. Participants were either rewarded with course credit or 5 Euro. The prototypical sequence of each experiment is depicted in Figure 2 (lower half).

3.1 Materials and Apparatus

Experiments were presented on a DELL P2210 monitor (1650*1050 pixels, 60 Hz). Pupil diameter was tracked using a SMI-RED 120 eye tracker (Sensomotoric Instruments GmbH). Participants were seated 60 cm from the monitor on a fixed chair, using a chinrest. Brightness was constantly 51.2 lx at eye position. All participants had normal or corrected to normal vision using contact lenses.

An implementation of Pong was used with four levels of difficulty that were presented in a random order to players, each lasting 400 s. The implementation was based on a blog post [Appleton 2014]. Pong is a simple two dimensional video game based on ping-pong: two rackets are on both sides of a field, with a ball moving in this field back and forth. For research, Pong is of particular interest, as it is (1) well researched, (2) easy to implement and replicate, and (3) foremost allows for the highest possible experimental standardization, which allows highest internal validity. The player's task is to hit the ball with the left racket. The classical implementation of this Atari game usually included two human players competing with each other. In this experiment, the player competed against the computer moving the right racket. Further, the implementation deviated from the original game by not including the current score to prevent possible non-difficulty related confounds to pupil diameter. The player controlled the racket's position by moving a mouse back and forth. The field always spanned 14.06 degree visual angle in horizontal and 8.86 degree visual angle in vertical dimension on screen. The background was presented in black, whereas lines and rackets were implemented in gray while the ball was white.

In Experiment 1, game difficulty was manipulated by adapting the speed of the ball, the classical way of manipulating difficulty in this game. Further ensuring that results were not only driven by more motion or eye-movements, in Experiment 2, an alternative way of manipulating difficulty was chosen: changing the players' rackets in size. The implementation as well as the differential manipulations of game difficulty are visualized in Figure 2. In Experiment 1, the racket was set to one sixth of the height of the field (i.e. 1.48 degree visual angle). The different levels were manipulated in the following way: in level 1, the ball needed 4 s to fly from side to side, 1.5 s in level 2, 0.4 s in level 3 and 0.21 s in level 4. We assumed that the game would be harder, the faster the ball was moving (Figure 2 upper left).

In Experiment 2, speed was set by default to 2 s from side to side. Racket sizes were 7.70 degree visual angle in level 1, 3.55 degree visual angle in level 2, 1.48 degree visual angle in level 3 (as in Experiment 1) and 0.59 degree visual angle in level 4. We assumed the game to be the harder, the smaller the player's racket (Figure 2 upper right). Parameters were set after piloting based on subjective difficulty perception of authors 1-3.

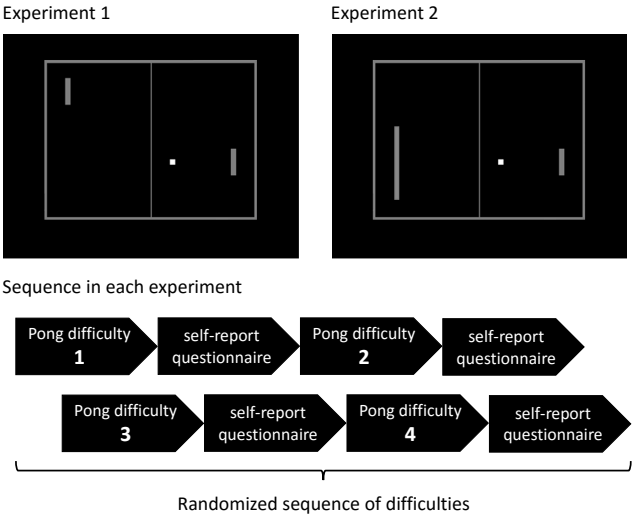

Figure 2: Representative depictions of the Pong implementation in both experiments. Upper left: experiment 1, speed was manipulated. Upper right: experiment 2, the size of the player's racket was manipulated. Lower half: prototypical experimental sequence in both experiments.

Assessing multiple aspects of gaming appraisal, players were asked to rate whether the game was fun to play, they enjoyed playing, they considered the game to be boring, the game held their attention. Additional measures captured participants' extent of focus on the game, and the degree of effort put into playing. The first three of these items were taken from the Intrinsic Motivation Inventory (IMI) [Deci and Ryan 2018], the later three from the Gaming Experience Questionnaire (GEQ) [IJsselsteijn et al. 2013]. For Experiment 2, participants were furthermore asked whether they were good at the task and whether they felt competent performing it as further manipulation check. Items were given in German based on the preference of the participant and were answered using a pen on paper. Items were answered with 5-point Likert scales (1 agree not at all - 5 completely agree). For the precise wording of each item, please see https://osf.io/56dkf/.

3.2 Participants

$n = 30$ students of anonymous university participated in Experiment 1 ($M_{\text{Age}} = 24.80$ years, 16 female). In Experiment 2, $n = 22$ students took part ($M_{\text{Age}} = 24{,}05$ years, 18 female).

In Experiment 2, a score increasing by one every time the player hit the ball was calculated, to have an objective indication of difficulty as manipulation check. Further ensuring that gaming difficulty and not movement was principally causing the effects, mouse movement was tracked.

3.3 Data processing

Pupil diameter was blink filtered using an interpolation approach for blinks, built on the physiologically possible boundaries of changes in pupil diameter as threshold for blink detection [Georgi et al. 2014]. Pupil diameter was then averaged for each experimental condition.

Subsequently, pupil diameter was z-standardized within participants across the difficulty levels for further analyses. For statistical analyses and data visualization, R was used, including lme, MuMin and ggplot packages [Barton and Barton 2013; Bates et al. 2014; Wickham 2016].

4 RESULTS

As difficulty is usually varied in Pong by changing the ball's speed, points scored were generated to evaluate the difficulty of each level for Experiment 2. Points scored per level were compared to assess the subjective difficulty of each level using t-tests for connected samples. In Experiment 2, the easiest and second easiest difficulty level did not differ significantly in points scored, which is why pupil diameter, as well as all other measures were averaged for those two levels. For the now three remaining difficulties, points scored differed significantly between each level (all beyond $p < 0.0001$; see Table 1 for the number of balls hit in each level).

Table 1: Manipulation checks for Experiment 2. Average with standard deviation in brackets. Upper row: number the ball was hit in each level. Lower row: mouse movement in arbitrary units.

	1	2	3	4
Points	27.00 (0)	26.00 (2.85)	12.05 (4.14)	6.09 (0.63)
Movement	4.56 (4.07)	12.51 (8.51)	17.07 (7.63)	16.16 (6.74)

4.1 Gaming difficulty and pupil dilation

Testing how pupil size relates to increasing difficulty, a generalized mixed model was calculated for each experiment. Here, pupil diameter was predicted by level and level^2. This fitted model was then compared to a model with only level as predictor. For both experiments, AIC based model comparison revealed the best trade off of parsimonity and explained variance for the model combining level and level^2 as predictors. Predictors, as well as their level of statistical significance and the resulting R^2 are given in Table 2. Predictors, statistical significance and R^2 were descriptively smaller for Experiment 2 than for Experiment 1.

Table 2: Multilevel models for all experiments with predictors for intercept, level of difficulty and the square of the level of difficulty. Furthermore, model fit is given by R^2. The written form of the fitted models is given in the teaser Figure 1. Statistical significance is indicated by asterisks ($p < 0.01$: $^+$; $p < 0.05$: *; $p < 0.01$: **; $p < 0.001$: ***).

	Intercept	Level	Level^2	R^2
Experiment 1	-6.85***	3.12***	-0.54***	0.67
Experiment 2	-4.27**	2.83*	-0.42*	0.17

In Figure 3, average pupil diameter per level is visualized for both experiments (left: Experiment 1, right: Experiment 2). In the upper row, average pupil diameter is given for each level. Colored lines indicate individual participants. In Experiment 1, the relation of pupil diameter to difficulty was more clear than in Experiment 2 for each individual participant. In the lower row, the same data points are given in conjunction with the regression line resulting from the models fitted given in Table 2 together with 95% confidence intervals. Both, the linearly increasing and the inverted u-shaped component of the regression are visible (Figure 3).

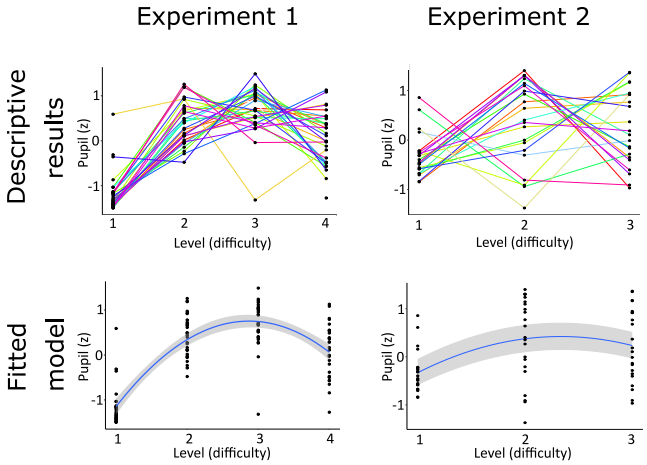

Figure 3: z-standardized pupil diameter across all difficulty levels for both experiments. In the upper row, each individual participant is given in a different color with points of measurement connected by a line. In the lower row, individual z-values are given in conjunction with the regression line as calculated from the general mixed models given in Table 2. The shaded gray area depicts a 95% confidence interval.

4.2 Questionnaire data and pupil dilation

Further assessing whether pupil diameter relates to gaming appraisal, a questionnaire was answered by participants after playing in each level. In Figure 4, correlations between items are given for both experiments. As some items were strongly correlated (Fun & Enjoyment; Holding attention, Focus & Effort; Boredom; Good & Competence), we opted to average these for visualization (Figure 5).

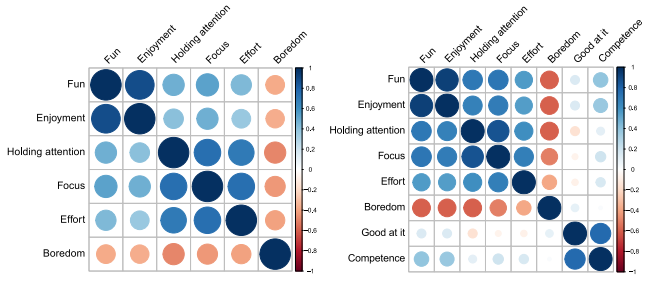

Figure 4: Correlations between questionnaire items, with dark blue depicting a strong positive and dark red depicting a strong negative correlation. Left: correlations for Experiment 1. Right: correlations for Experiment 2.

Table 3: Predictors for respective multilevel models. In each model, pupil size was the single predictor for the item indicated by the column title. Statistical significance is indicated by asterisks ($p < 0.1$: $^{+}$; $p < 0.05$: *; $p < 0.01$: **; $p < 0.001$: ***).

	Fun	Enjoyment	Holding attention	Focus	Effort	Boredom	Good at it	Competence
Experiment 1	0.29*	0.20+	0.47***	0.30**	0.38***	-0.43***		
Experiment 2	0.10	0.03	0.31	0.28	0.06	-0.20	-0.29	-0.40*

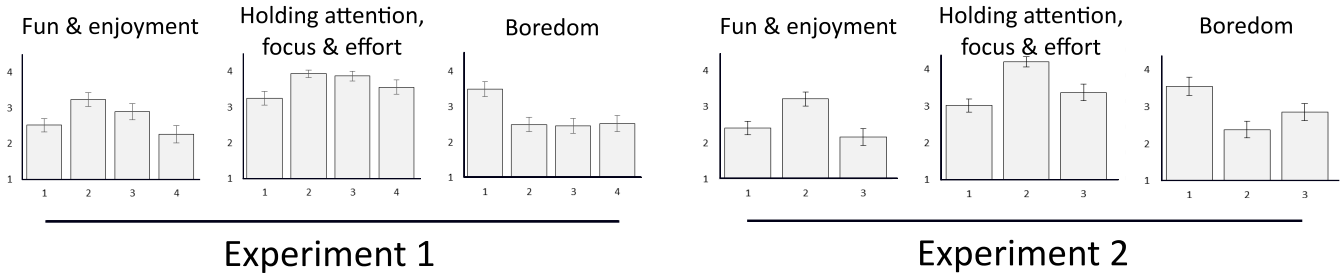

Figure 5: Summarized questionnaire data for both experiments. Highly correlated items were averaged. Left: Experiment 1, right: Experiment 2. Error bars represent +- one standard error of the mean. All items were presented on a Likert rating scale between 1 and 5.

In Experiment 1, all items revealed small to moderate correlations with pupil diameter. Except for Enjoyment, which showed a trend level significance in its correlation with pupil diameter, all these correlations were statistically significant. As expected, rated Fun, Enjoyment, Holding attention, Focus, and Effort were all positively correlated to pupil diameter. The reverse item Boredom was negatively correlated with pupil diameter. For Experiment 2, correlations were descriptively in the same direction for all items as for Experiment 1, however, the correlations were consistently smaller and not significant. The newly included items Good at it and Competence showed negative correlations with pupil diameter, which was significant for Competence. Table 3 depicts the correlations for pupil diameter with each questionnaire item separately for Experiments 1 and 2.

In Figure 5, questionnaire data are visualized for the categories derived from the between item correlations visualized in Figure 3. Similar to pupil diameter, an u-shaped pattern is visible in questionnaire data: levels 2 and 3 (Experiment 1) and level 2 (Experiment 2) were consistently rated better than the hardest and the easiest level. As an exception, players rated their subjective competence and how good they were in Experiment 2 the lower, the higher the level. An overview of all individual questionnaire items and the resulting descriptive patterns is given in (https://osf.io/56dkf/).

4.3 Further analyses

While participants were very consistent in showing a relatively small pupil in the easiest condition (Experiment 1), variance between participants was much larger in the hardest level. Aiming to disentangle this variance for the hardest level in Experiment 1 and 2, relatively rated Fun and Enjoyment, Holding attention, Focus and Effort and Boredom were correlated with the relative pupil size in the hardest level in both experiments. No correlation reached statistical significance. Still, Holding attention, Focus and Effort put into playing the most difficult level revealed a trend-level significant correlation: $r = 0.27$, $p = 0.068$. Fun and Enjoyment were clearly not related to the variation in pupil sizes in the hardest levels ($r = 0.015$, $p = 0.922$), neither was Boredom ($r = 0.017$, $p = 0.908$).

In Experiment 2, mouse movement was tracked to rule out potential artifacts of motor execution [Strauch et al. 2018, e.g.] and included in the regression model predicting pupil diameter: AIC based model comparison showed that the best model includes level and level^2 as predictors, but not mouse movement, arguing against an influence of movement on pupil dilation (see Table 1 for mouse movement and points scored in Experiment 2).

5 DISCUSSION

Different levels of difficulty of the video game 'Pong' could be distinguished using pupil dilation in two experiments. The classical manipulation of difficulty, i.e., speed, in Experiment 1 revealed an inverted u-shape for the relation of pupil diameter to increasing difficulty. A high fit between abilities and demands was accompanied by a large pupil dilation, whereas both under- and overload were associated with a relatively small pupil. Like pupil size, rated gaming appraisal showed an inverted u-shape from under- to overload. Even in Experiment 2, where difficulty was manipulated in a less common way, i.e., racket size, this relation was found. Effects were generally smaller for Experiment 2, possibly because, by default, difficulty may best be manipulated in Pong by adjusting the ball's speed.

Gaming appraisal correlated with pupil diameter, in medium (Experiment 1) to small sizes (Experiment 2). These correlations were especially strong for boredom and the degree of attention held by the player and were generally of larger size in Experiment 1 (up to $r = 0.47$). Compared to former investigations using other peripheral indicators of activation, correlations were stronger using pupillometry (vs. skin conductance and heart rate with up to $r = 0.288$) [Giakoumis et al. 2010].

The question arises whether behavioural measures may be employed for dynamic difficulty adjustment in Pong. In Experiment 2, levels 1, 2, and 3 could be distinguished based on the number of balls hit, showing a linear decline of balls hit with increasing difficulty. This score alone could however not dissociate whether players were under-/overloaded or were still putting effort in playing. Sensing a player's state beyond the behavioural input (here: moving the racket and hitting balls) is therefore most useful in order to perform the right adjustments in difficulty (easier, constant, harder).

For the hardest levels, pupil dilation varied between participants the most: Some participants having relatively large pupils in level 4 did not exhibit an inverted u-shape, but rather a linearly increasing pattern. They might just not have reached a state of overload, but kept being invested in the game and would only have given up completely at an even higher difficulty. Indeed, when correlating the individual relative pupil diameter in level 4 with the relative rated effort/focus/held attention, a trend-level effect ($p = 0.068$) with a small to moderate positive correlation was observed. This suggests that pupil was larger, the more engaged participants still were in playing level 4. In line with this assumption, also [Gutjahr and Wiemeyer 2019] report larger pupils for participants that rated their motivation put into playing higher.

Physiology based dynamic difficulty adjustment provides a context-free option for tracking user experience and could prove beneficial for a number of applications, not limited to gaming alone. Interactive films, such as 'Heavy rain' or 'Beyond: Two Souls', where the player only has to intervene the movie-like story-plot occasionally with speeded commands in so called quick-time events could benefit from this additional source of information about the player's state. The nature of such games makes it difficult to assess how challenged the player is, as typically employed information such as reaction times or other performance indices are not available. Another field of application may be seen in serious games, e.g., the level of detail might vary based on pupil diameter in online tutorials. Besides enabling adaptivity, pupil dilation might also reveal the current level of game appraisal to developers during the evaluation of game scenes. Moreover, the underlying physiology to the phenomenon investigated is beyond restricted user groups: In pupil size reflected fluctuations of the release of norepinephrine are so general, they can even be found in monkeys [Rajkowski et al. 2004]. For a u-shape, difficulty must be set differentially depending on skill, though. Beyond gaming, such information may be used for adapting user interfaces on a more general level, for example by dynamically adjusting dwell times during gaze input for several subsequent selections. For the context of gaze based interaction, it has to be noted that pupil dilation may also reflect other processes than mental effort and appraisal, such as deciding on a letter during text input [Strauch et al. 2018, 2020], however, effects related to mental effort usually are of substantially larger absolute size [Hess and Polt 1964, e.g.]. This should also allow for detecting workload in gaze based interaction, as demonstrated before with low-cost equipment [Hansen et al. 2018]. Of course, substantially different interfaces would require further research; this investigation may be a starting point towards such investigations.

Here, brightness was kept constant to preclude confounds in pupil data, however, applicability of pupillometric measurements is not restricted to strictly brightness controlled settings, as long as changes in brightness are sensed [Pfleging et al. 2016]. This is due to the linear additive nature of pupil diameter, which may be decomposed into brightness induced and activation induced changes [Pfleging et al. 2016; Reilly et al. 2019]. In this context, especially virtual reality games might prove attractive as first application, as brightness is controlled by the device. With 400 s, participants played comparably long in every condition. Likely, also shorter durations may provide data that can reliably indicate levels of (sub-) optimal fit, especially when going beyond a mere analysis of average pupil dilation, making use of more advanced signal processing [Duchowski et al. 2018; Wierda et al. 2012, e.g.]. As next step, investigations should therefore closely assess how long pupil dilation needs to be tracked for a reliable assessment of gaming appraisal and how this interval may become shorter. Furthermore, external validity needs to be ensured by testing present-day more complex games.

Additional eye-tracking metrics, such as saccades, could help identifying a player's level of gaming appraisal [Biswas et al. 2016; Stuyven et al. 2000]. For Pong, smooth pursuits when following the ball might also indicate gaming appraisal, as it has been demonstrated for cognitive load [Kosch et al. 2018].

Results showed no effect of mouse movement on pupil dilation beyond difficulty. Further ensuring that not a change in the perceptual input (e.g. a faster ball) caused pupils to dilate more during fit, but a better experience during playing, a yoked control condition could be tested: Here participants would have to monitor the same game scene without interacting as a baseline [Gutjahr and Wiemeyer 2019]. Results on mouse movement let assume that observed effects are independent of the specific input modality, be it by keyboard, joystick or gaze. However, this generalization will need to be checked in future studies.

Hunicke states three key steps for realising dynamic difficulty adjustments [Hunicke 2005]. We show that pupil dilation may serve for the assessment of the need to adjust (1), further studies will have to show how to use this information for determining which changes are needed (2) and how to perform them smoothly (3), best by considering existing frameworks for psychophysiologically adaptive games [Nacke et al. 2011]. It is important to note that dynamic difficulty adjustment has boundaries and may result in players felt being cheated by the adaptive system. However, when performed in accordance with the predefined steps [Hunicke 2005], appropriate changes were repeatedly shown to improve players' performance without stealing their sense of achievement [Hunicke 2005; Tan et al. 2011, e.g.]. This way, games could become more accessible and inclusive to players with all levels of individual abilities.

6 CONCLUSIONS

In this paper, we demonstrate that pupil diameter, a peripheral psychophysiological indicator, is a viable variable for assessing players' level of appraisal while playing video games. The inverted u-shape relation of pupil diameter to game difficulty underlines the potential of pupillometry for dynamically adjusting game difficulty: Pupil dilation, a remote indicator of load that is strongly associated to subjective gaming appraisal, may ensure a timely adjustment of gaming difficulty. Contributing to the endeavours of optimizing

user-adaptive interfaces in general, this investigation shows the principal feasibility and potential of pupil dilation as information source.

7 ONLINE RESOURCES

Further detailed analyses and data visualizations, as well as raw data are given in (https://osf.io/56dkf/).

ACKNOWLEDGMENTS

Christoph Strauch was funded by a PhD stipend by the Friedrich-Ebert-Stiftung. We are grateful to Seyma Nur's generous support with designing the teaser figure.

REFERENCES

Trevor Appleton. 2014. *Writing Pong using Python and Pygame.* http://trevorappleton.blogspot.com/2014/04/writing-pong-using-python-and-pygame.html

Kamil Barton and Maintainer Kamil Barton. 2013. Package MuMIn. *Version* 1 (2013), 18.

Douglas Bates, Martin Maechler, Ben Bolker, Steven Walker, et al. 2014. lme4: Linear mixed-effects models using Eigen and S4. *R package* 1, 7 (2014), 1–23.

Pradipta Biswas, Varun Dutt, and Pat Langdon. 2016. Comparing ocular parameters for cognitive load measurement in eye-gaze-controlled interfaces for automotive and desktop computing environments. *International Journal of Human-Computer Interaction* 32, 1 (2016), 23–38. https://doi.org/10.1080/10447318.2015.1084112

Guillaume Chanel, Cyril Rebetez, Mireille Bétrancourt, and Thierry Pun. 2008. Boredom, engagement and anxiety as indicators for adaptation to difficulty in games. In *Proceedings of the 12th international conference on Entertainment and media in the ubiquitous era*. ACM, 13–17. https://doi.org/10.1145/1457199.1457203

E.L. Deci and R.M. Ryan. 2018. Intrinsic motivation inventory (IMI). (2018).

Anders Drachen, Lennart E. Nacke, Georgios Yannakakis, and Anja Lee Pedersen. 2010. Correlation Between Heart Rate, Electrodermal Activity and Player Experience in First-person Shooter Games. In *Proceedings of the 5th ACM SIGGRAPH Symposium on Video Games* (Los Angeles, California) *(Sandbox '10)*. ACM, New York, NY, USA, 49–54. https://doi.org/10.1145/1836135.1836143

Andrew T Duchowski, Krzysztof Krejtz, Izabela Krejtz, Cezary Biele, Anna Niedzielska, Peter Kiefer, Martin Raubal, and Ioannis Giannopoulos. 2018. The index of pupillary activity: measuring cognitive load vis-à-vis task difficulty with pupil oscillation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 282. https://doi.org/10.1145/3173574.3173856

Jan Ehlers, Christoph Strauch, Juliane Georgi, and Anke Huckauf. 2016. Pupil size changes as an active information channel for biofeedback applications. *Applied Psychophysiology and Biofeedback* 41, 3 (2016), 331–339. https://doi.org/10.1007/s10484-016-9335-z

Kate C Ewing, Stephen H Fairclough, and Kiel Gilleade. 2016. Evaluation of an adaptive game that uses EEG measures validated during the design process as inputs to a biocybernetic loop. *Frontiers in Human Neuroscience* 10 (2016), 223. https://doi.org/10.3389/fnhum.2016.00223

Juliane Georgi, David Kowalski, Jan Ehlers, and Anke Huckauf. 2014. Real-time feedback towards voluntary pupil control in human-computer interaction: Enabling continuous pupillary feedback. In *International Workshop on ICTs for Improving Patients Rehabilitation Research Techniques*. Springer, 104–114. https://doi.org/10.1007/978-3-662-48645-0_10

Dimitrios Giakoumis, Athanasios Vogiannou, Ilkka Kosunen, Konstantinos Moustakas, Dimitrios Tzovaras, and George Hassapis. 2010. Identifying Psychophysiological Correlates of Boredom and Negative Mood Induced during HCI.. In *B-Interface*. 3–12. https://doi.org/10.13140/2.1.4997.7281

Eric Granholm, Robert F Asarnow, Andrew J Sarkin, and Karen L Dykes. 1996. Pupillary responses index cognitive resource limitations. *Psychophysiology* 33, 4 (1996), 457–461. https://doi.org/10.1111/j.1469-8986.1996.tb01071.x

Ellermeier W. Hardy S. Göbel S. Gutjahr, M. O. and J. Wiemeyer. 2019. The pupil response as an indicator of user experience in a digital exercise game. *Psychophysiology* (2019), e13418. https://doi.org/10.1111/psyp.13418

John Paulin Hansen, Diako Mardanbegi, Florian Biermann, and Per Bækgaard. 2018. A gaze interactive assembly instruction with pupillometric recording. *Behavior research methods* 50, 4 (2018), 1723–1733. https://doi.org/10.3758/s13428-018-1074-z

Eckhard H Hess and James M Polt. 1964. Pupil size in relation to mental activity during simple problem-solving. *Science* 143, 3611 (1964), 1190–1192. https://doi.org/10.1126/science.143.3611.1190

Robin Hunicke. 2005. The case for dynamic difficulty adjustment in games. In *Proceedings of the 2005 ACM SIGCHI International Conference on Advances in computer entertainment technology*. ACM, 429–433. https://doi.org/10.1145/1178477.1178573

R Hunicke and V Chapman. 2004. AI for Dynamic Difficulty Adjustment in Games, Challenges in Game Artificial Intelligence AAAIWorkshop.

WA IJsselsteijn, YAW De Kort, and Karolien Poels. 2013. The game experience questionnaire. *Eindhoven: Technische Universiteit Eindhoven* (2013).

Johannes Keller, Herbert Bless, Frederik Blomann, and Dieter Kleinböhl. 2011. Physiological aspects of flow experiences: Skills-demand-compatibility effects on heart rate variability and salivary cortisol. *Journal of Experimental Social Psychology* 47, 4 (2011), 849–852. https://doi.org//10.1016/j.jesp.2011.02.004

J Matias Kivikangas, Guillaume Chanel, Ben Cowley, Inger Ekman, Mikko Salminen, Simo Järvelä, and Niklas Ravaja. 2011. A review of the use of psychophysiological methods in game research. *Journal of Gaming & Virtual worlds* 3, 3 (2011), 181–199. https://doi.org//10.1386/jgvw.3.3.181_1

Máté Köles, Luca Szegletes, and Bertalan Forstner. 2015. Towards a physiology based difficulty control system for serious games. In *2015 6th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*. IEEE, 323–328. https://doi.org/10.1109/CogInfoCom.2015.7390612

Thomas Kosch, Mariam Hassib, Pawel W Wozniak, Daniel Buschek, and Florian Alt. 2018. Your Eyes Tell: Leveraging Smooth Pursuit for Assessing Cognitive Workload. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 436. https://doi.org/10.1145/3173574.3174010

Nicholas Lofthouse, L Eugene Arnold, Sarah Hersch, Elizabeth Hurt, and Roger DeBeus. 2012. A review of neurofeedback treatment for pediatric ADHD. *Journal of Attention Disorders* 16, 5 (2012), 351–372. https://doi.org//10.1177/1087054711427530

Sebastiaan Mathôt. 2018. Pupillometry: Psychology, physiology, and function. *Journal of Cognition* 1, 1 (2018). https://doi.org/10.5334/joc.18

Peter R Murphy, Redmond G O'connell, Michael O'sullivan, Ian H Robertson, and Joshua H Balsters. 2014. Pupil diameter covaries with BOLD activity in human locus coeruleus. *Human Brain Mapping* 35, 8 (2014), 4140–4154. https://doi.org//10.1002/hbm.22466

Lennart Nacke and Craig A Lindley. 2008. Flow and immersion in first-person shooters: measuring the player's gameplay experience. In *Proceedings of the 2008 Conference on Future Play: Research, Play, Share*. ACM, 81–88. https://doi.org/10.1145/1496984.1496998

Lennart Erik Nacke, Michael Kalyn, Calvin Lough, and Regan Lee Mandryk. 2011. Biofeedback game design: using direct and indirect physiological control to enhance game interaction. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 103–112. https://doi.org/10.1145/1978942.1978958

Jeanne Nakamura and Mihaly Csikszentmihalyi. 2014. The concept of flow. In *Flow and the foundations of positive psychology*. Springer, 239–263. https://doi.org//10.1007/978-94-017-9088-8_16

Bastian Pfleging, Drea K Fekety, Albrecht Schmidt, and Andrew L Kun. 2016. A model relating pupil diameter to mental workload and lighting conditions. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. ACM, 5776–5788. https://doi.org/10.1145/2858036.2858117

Alan T Pope and Edward H Bogart. 1994. Method of encouraging attention by correlating video game difficulty with attention level. US Patent 5,377,100.

Janusz Rajkowski, Henryk Majczynski, Edwin Clayton, and Gary Aston-Jones. 2004. Activation of monkey locus coeruleus neurons varies with difficulty and performance in a target detection task. *Journal of Neurophysiology* 92, 1 (2004), 361–371.

Jamie Reilly, Alexandra Kelly, Seung Hwan Kim, Savannah Jett, and Bonnie Zuckerman. 2019. The human task-evoked pupillary response function is linear: implications for baseline response scaling in pupillometry. *Behavior Research Methods* 51, 2 (2019), 865–878. https://doi.org//10.3758/s13428-018-1134-4

Christoph Strauch, Jan Ehlers, and Anke Huckauf. 2017. Pupil-assisted target selection (PATS). In *IFIP Conference on Human-Computer Interaction*. Springer, 297–312. https://doi.org/10.1007/978-3-319-67687-6_20

Christoph Strauch, Lukas Greiter, and Anke Huckauf. 2018. Pupil dilation but not microsaccade rate robustly reveals decision formation. *Scientific Reports* 8, 1 (2018), 13165. https://doi.org/10.1038/s41598-018-31551-x

Christoph Strauch, Ina Koniakowsky, and Anke Huckauf. 2020. Decision Making and Oddball Effects on Pupil Size: Evidence for a Sequential Process. *Journal of Cognition* (2020). https://doi.org/10.5334/joc.96

Els Stuyven, Koen Van der Goten, André Vandierendonck, Kristl Claeys, and Luc Crevits. 2000. The effect of cognitive load on saccadic eye movements. *Acta Psychologica* 104, 1 (2000), 69–85. https://doi.org/10.1016/S0001-6918(99)00054-2

Chin Hiong Tan, Kay Chen Tan, and Arthur Tay. 2011. Dynamic game difficulty scaling using adaptive behavior-based AI. *IEEE Transactions on Computational Intelligence and AI in Games* 3, 4 (2011), 289–301. https://doi.org/10.1109/TCIAIG.2011.2158434

Jose Antonio Urigüen and Begoña Garcia-Zapirain. 2015. EEG artifact removal—state-of-the-art and guidelines. *Journal of Neural Engineering* 12, 3 (2015), 031001. https://doi.org/10.1088/1741-2560/12/3/031001

Hadley Wickham. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Stefan M Wierda, Hedderik van Rijn, Niels A Taatgen, and Sander Martens. 2012. Pupil dilation deconvolution reveals the dynamics of attention at high temporal resolution. *Proceedings of the National Academy of Sciences* 109, 22 (2012), 8456–8460. https://doi.org//10.1073/pnas.1201858109

Robert M Yerkes and John D Dodson. 1908. The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology* 18, 5 (1908), 459–482. https://doi.org/10.1002/cne.920180503

Beste F Yuksel, Kurt B Oleson, Lane Harrison, Evan M Peck, Daniel Afergan, Remco Chang, and Robert JK Jacob. 2016. Learn piano with BACh: An adaptive learning interface that adjusts task difficulty based on brain state. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. ACM, 5372–5384. https://doi.org/10.1145/2858036.2858388