Predicting User Engagement in Video Advertisement: Insights from Pupillary Response and Heart Rate

Gregor Strle[1,2], Andrej Košir[1], Evin Aslan Oğuz[1,3] and Urban Burnik[1]

[1] *University of Ljubljana, Faculty of Electrical Engineering, Tržaška 25, 1000 Ljubljana, Slovenia*
[2] *ZRC SAZU, Novi trg 2, 1000 Ljubljana, Slovenia*
[3] *Nielsen Lab d.o.o., Obrtniška ulica 15, 6000 Koper, Slovenia*

Abstract

The article presents the results of predicting user engagement with in-video ads using physiological sensor signals. Specifically, we examine pupil response and heart rate as possible predictors of user engagement. To this end, we conducted an experiment with 33 young participants (age M = 21.70, SD = 2.36; female = 68%) in which their psychometric (engagement score) and physiological responses (pupil dilation and heart rate) to four in-video ads were recorded. The ground truth for the ad engagement was collected using the User Engagement Scale Short Form (UES-SF), a standardized psychometric instrument for measuring user engagement. The UES-SF dimensions Aesthetic Appeal (AE) and Perceived Usability (PU) were used to calculate the combined User Engagement Score. Several machine learning classifiers were evaluated that used heart rate and pupil response as predictors of engagement. The best overall results were obtained by Random Forest Classifier ('weighted' F1 score = .76, Precision=.84, Recall=.95), Logistic Regression and Support Vector Classifier (the latter two with the same scores: 'weighted' F1 score = .74, Precision=.82, Recall=1)

Keywords

user engagement, perceived usability, aesthetic appeal, advertisement exposure, pupil dilation, heart rate

1. Introduction

Advances in technology and digital media advertising have enabled new approaches to measuring consumer engagement and exposure to online marketing [1]. These go beyond traditional frequency measurements (reports by recall, number of views, likes, etc.) and now include online consumer behavior, social media metrics and trends, and user engagement and attitude ratings [2, 1]. One area of advertising that is growing particularly strongly is ad-supported video streaming, which has overtaken video-on-demand streaming [3])

The goal of the research presented here is to assess heart rate and pupil dilation as predictors of user engagement with in-video advertising used by online streaming services. The ground truth for the ad engagement was collected using the User Engagement Scale-Short Form (UES-SF) [4], an established psychometric instrument for measuring user engagement. The focus group of the presented research was younger consumers who use streaming services extensively and are accustomed to in-video advertising [5].

Human-Computer Interaction Slovenia 2022, November 29, 2022, Ljubljana, Slovenia
gregor.strle@fe.uni-lj.si (G. Strle); andrej.Kosir@fe.uni-lj.si (A. Košir); evin.aslan-oguz@fe.uni-lj.si (E. A. Oğuz); urban.burnik@fe.uni-lj.si (U. Burnik)

The contribution of the presented study is to examine the potential of physiological cues as a measure of advertising exposure. It is well known that advertising evokes emotional arousal and triggers cognitive processes in consumers [6]. These, in turn, influence a person's measurable physiological responses and can give us new insights into consumer behavior. Due to the steady development of wearable devices with physiological sensors (e.g., smartwatches), physiological measurements may be used in the future as novel marketing strategies and technologies related to the impact of advertising and consumer behavior [1].

In the following, we briefly present related work. Next, we present the experimental design and procedure for collecting user responses and the selection of materials and tools. The statistical analysis and the evaluation of machine learning models are presented in the Results section. The article concludes with a summary of the main results and possible directions for future work.

2. Related Work

Contemporary advertising strategies and services require rapid and accurate insight into consumer engagement and exposure to media content. Fast and efficient measurement methods with little interference with the observed subject are preferred. We hypothesize that measuring physiological signals known to be associated with emotional arousal and cognitive processes could provide a good basis for unobtrusive measures of engagement and exposure to media advertising.

Richardson et al. examined engagement with video and audio narration using wrist sensors that measure heart rate variability, electrodermal activity, and body temperature [7]. They found a significant physiological response to all 3 observed measures. Ayres et al. provided a comprehensive review of physiological measures of intrinsic cognitive load, including heart rate, heart rate variability, respiratory measurements, pupil dilation, blink rate, fixation, electrodermal measurements, functional near-infrared spectroscopy, electroencephalography, and functional magnetic resonance imaging [8]. The most sensitive physiological measurements were blink rate, heart rate, pupil dilation, and alpha waves.

Recent research has shown that pupil dilation can be used to assess cognitive processes in participating subjects. P. van der Wel [9] reported that with respect to the cognitive control domains of updating, switching, and inhibition, an increase in task demands leads to increased pupil dilation. However, the study does not establish a clear model for the relationship between pupil dilation and performance. The use of pupil dilation in studies of advertising effectiveness has been known since the early 1970s Hensel1070. Using pupil dilation to measure emotional arousal during video consumption is challenging because pupil dilation is sensitive to changes in brightness. In [10], a linear model of the pupillary light reflex is proposed that predicts a viewer's pupil diameter based only on incident light intensity. The model can be used to subtract the effects of brightness to determine subjects' emotional arousal as a function of the observed scene. Jerčić et al. [11] examined pupil dilation and heart rate as measures of physiological arousal and identified them as possible indicators of cognitive ability in serious-gaming participants. In [12], the application of personalized advertising systems based on the measurement of heart rate variability on the go is proposed. Pham et al. [13] found that heart

rate variability (HRV) can be a stable and affordable source of information for neurophysiological and psychophysiological studies, provided that appropriate acquisition procedures and well-developed indices are available. Schaffer et al. [14] have addressed the complexity of heart oscillations and reviewed existing methods for monitoring HRV in the time and frequency domains using nonlinear metrics.

Much of the evidence cited above suggests that among physiological measures, pupil dilation and heart rate are good choices for detecting fluctuations in emotional arousal, mental activity, and cognitive processes. In the presented study, we focus on the putative effects of the two on advertiser engagement and media use.

3. Materials and Methods

A pre-selection of 12 videos and 12 ads was taken from the online streaming service YouTube. These materials were selected in collaboration with three marketing experts from The Nielsen Company to address different levels of engagement. All materials were in English and aired in the United States.

A crowdsourcing study with Clickworker[1] was conducted to determine engagement with YouTube's 12 videos and ads. The ads were inserted into the videos at random positions, with combinations of different engagement levels. The goal of these combinations was to simulate the experience of ad-supported video streaming. Engagement was measured on a 5-point scale (How engaging is the ad?: none-medium-strong-very strong). Engagement scores were collected from study participants (N=360, age=18-24).

Based on the results of the crowdsourcing study, different video and in-video ad combinations were created based on the engagement scores. The final selection for the experiment was made in collaboration with Nielsen media experts. For the experiment, four combinations of ads and videos were created based on the engagement level and brand awareness criteria for the ads (2x2: known vs. unknown brand with higher and lower engagement scores). The number of combinations was limited to make the experiment feasible (see 3.2). The following four ads were used: Dior Joy Perfume[2], Coca Cola[3], Little Baby's Ice Cream[4], Waring Ice Cream Maker[5].

3.1. Psychometric and Physiological Measures

Ground truth for user engagement was measured using the User Engagement Scale-Short Form (UES-SF) [4]. The UES-SF is a 12-item questionnaire covering four dimensions of engagement: Focused Attention (FA), Aesthetic Appeal (AE), Perceived Usability (PU), and Reward (RW). The dimensions are rated on a 5-point scale and a total score can be calculated as an average across the selected dimensions. For the purposes of this study, AE and PU were selected as the most relevant aspects of ad engagement. This is in line with the guidelines of UES-SF, where a subset of the dimensions from UES-SF (relevant to a particular case) can be used to calculate

[1] Clickworkerhttps://www.clickworker.com
[2] https://www.youtube.com/watch?v=vfOnEaaPaF4
[3] https://www.youtube.com/watch?v=vUMQeNw2QDA
[4] https://www.youtube.com/watch?v=erh2ngRZxs0
[5] https://youtu.be/GJ4P6ko_aLU

user engagement [4]. According to [4], PU is defined as ”negative affect experienced as a result of the interaction and the degree of control and effort expended”, while AE is defined as ”the attractiveness and visual appeal of the interface” (or, in our case, the ad). Example items for both dimensions, tailored to our case: ”PU.1: I felt frustrated while watching this Ad.” and ”AE.1 This Ad was attractive.” [4]. The combined user engagement score was then calculated by averaging the scores of both dimensions.

Several physiological sensor signals were collected from the participants: eye-tracking data with the Tobii Pro Glasses 2 eyetracker (pupil dilation, saccades, fixations) and heart rate and electrodermal activity (EDA) with Empatica 3. In this article, we report pupil dilation and heart rate.

3.2. Experimental Procedure

An experimental design was used in which all participants watched all four ads within their assigned shuffle set. This was done to control for potential carryover effects from the preceding combination to the next, as the engagement response elicited by the preceding sequence could influence the participant's response in the next sequence. Four sets were created, yielding four combinations (Set1: Ad1, Ad2, Ad3, Ad4; Set2: Ad1, Ad3, Ad2, Ad4; Set3: Ad4, Ad2, Ad3, Ad1; Set4: Ad4, Ad3, Ad2, Ad1). Note that the number of combinations was limited to four sets to keep the duration of the experiment per user still workable, but shuffle ads so that no ad has the same preceding ad more than once.

Next, the four sets were randomly and evenly assigned (taking into account age and gender) to participants (N=44, age=18-24), with each participant being assigned only one set. Within each set, the four combinations of video ads were interspersed with a 2-minute break to further isolate possible carryover effects of the preceding combination to the next and to give participants a break.

The experiment was conducted in a controlled environment (the Lucami laboratory at the Faculty of Electrical Engineering) to ensure a uniformly lit and quiet environment. First, informed consent and demographic data were collected from each participant. The participants were familiarized with the goal of the experiment (to collect physiological data and engagement scores for the ads) and given time to familiarize themselves with the procedure. They then sat down on a sofa and watched each video sequence on TV. After viewing each ad, they were asked to rate their engagement with the ad (on a laptop) using UES-SF. Participants also recorded other aspects related to advertising exposure, including affective state (valence and arousal), brand familiarity and recall, and purchase intention. The average duration of the experiment was 45 minutes. The experiment was completed by 44 participants. Due to incomplete responses and errors in sensor measurement, a final sample of 33 participants (age M=21.70, SD =2.36; female=68%) was used for further analysis.

3.3. Data Preprocessing

Within-subject normalization of pupil response data was performed to account for individual differences in pupil dilation between the participants. Pupil dilation and heart rate data were analyzed directly; no additional features were extracted. For each individual participant, the

average pupil dilation across all four ads was calculated and then subtracted from the mean values of pupil size for each ad. In this way, only the relative differences in pupil dilation per ad were recorded. Heart rate data were averaged, and median heart rates per ad per participant were calculated. Outliers (outside the threshold SD =2.5) for both pupil dilation and heart rate were imputed with the medians per participant and per ad. Next, both heart rate and pupil dilation were normalized using MinMaxScaler (sklearn), heart rate data to a range [0, 1] and pupil dilation to a range [-1,1].

Statistical analyses (ANOVA and pairwise t-tests) and classification using machine learning were performed in Python with the Scipy, Pinguoin, open-cv, and sklearn libraries. Visualization of the data was done using the matplotlib and seaborn libraries. The boxplots in Figures 1, 2, and 3 represent the minimum, first quartile, median, third quartile, and maximum of the visualized data.

4. Results

4.1. User Engagement

The dimensions Aesthetic Appeal (AE) and Perceived Usability (PU) were selected for modeling user engagement. For each participant and for each dimension, the average scores were first calculated from the average scores of the respective items of the dimensions. Figure 1 shows the distribution of AE and PU across the ads. We can observe a strong negative correlation between the two dimensions ($r(117)=-.67$, $p < .001$).

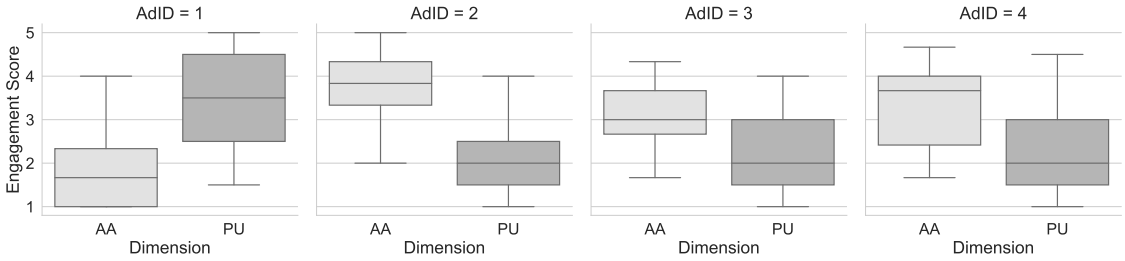


Figure 1: The average scores of Aesthetic Appeal (AE) and Perceived Usability (PU) for each ad. We can observe a strong negative correlation between AE and PU across all four ads.

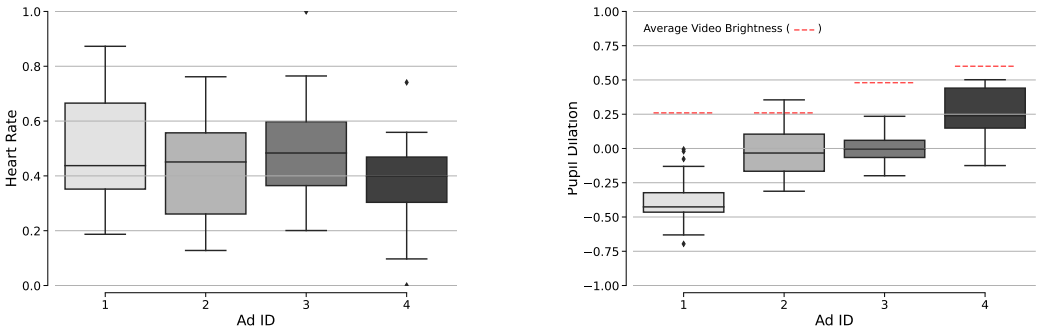
Next, both dimensions were summed and averaged to produce a combined engagement score, which was later used in machine learning. Statistical analysis examined differences in participant engagement between the four ads. Summary statistics are provided in Table 1. The Kruskal-Wallis test revealed no significant differences in engagement scores between the four ads ($H=6.83$, $p=.078$).

4.2. Pupil Dilation and Heart Rate

In our earlier work [15] we had reported preliminary summary statistics for each physiological dimension. The Shapiro-Wilk test showed that the data for both dimensions were normally distributed. A Pearson correlation coefficient was calculated to assess the relationship between

	N	Mean	SD	SE	95% Conf. Interval
Ad1	29	2.61	0.49	0.09	[2.43 – 2.80]
Ad2	30	2.89	0.39	0.07	[2.75 – 3.04]
Ad3	28	2.66	0.41	0.08	[2.50 – 2.82]
Ad4	30	2.85	0.43	0.08	[2.69 – 3.00]

Table 1
Summary statistics for the combined User Engagement Score. No significant differences were found between the ads.



(a) Heart Rate (b) Pupil Dilation

Figure 2: (a) The differences in heart rate among the ads. (b) The differences in pupil dilation between the ads. The red line represents the average video brightness of each video ad.

pupil dilation and heart rate, and a weak negative correlation was found between the two sensor signals (r(117)=-.17, p=.08).

One-way ANOVAs were performed to determine whether the median of pupil dilation and heart rate (Figure 2) was the same between the four ads (Ad1-Ad4). For the heart rate (F=2.80, p=.043; M=.4429, SD =.1736, SE =.02, 95%CI=[.41, .47]), multiple comparison t-tests (Bonferroni-corrected) revealed no significant difference between the ads. For pupil dilation (F=76.44, p <.001; M=-.037, SD =.28, SE =.026, 95%CI=[-.089, .014]), several significant differences were found, as also visible in Figure 2(b). Multiple pairwise t- tests (with Bonferroni correction) showed significant differences in the increase in pupil dilation between the following ads: Ad1 and Ad2 (T=-7.74, p <.001), Ad1 and Ad3 (T=-9.69, p <.001), Ad1 and Ad4 (T=-13.70, p <.001), Ad2 and Ad4 (T=-6.32, p <.001), and Ad3 and Ad4 (T=-6.88, p <.001). No significant difference in pupil dilation was observed between Ad2 and Ad3.

A moderate negative correlation was found between the engagement dimensions AE and PU and pupil dilation (AE and pupil dilation: r(117)=-.39, 95%CI=[.23, .53], p <.001, power=.99; PU and pupil dilation: r=-.36, 95%CI=[-.51, -.19], p <.001, Hedges' g=.98). No significant correlations were found between heart rate and the dimensions AE and PU.

4.3. The Effect of Ad Brightness on Pupil Dilation

Some studies have reported that pupil dilation can be affected by both light and contrast [16]. Also in our case, the differences in pupil dilation between the ads could be due to the differences in brightness and contrast between the ads and not to the content of the ads.

To this end, we analyzed the brightness and contrast characteristics of the ads. The overall average brightness (median) and contrast (standard deviation) of an ad were calculated from the averages of the 1-second frames (first converted to grayscale) for each in-video ad. A strong positive correlation was found between brightness and pupil dilation, r(117)=-.67, p <.001. Although the results show a significant difference in overall pupil dilation, the brightness of the ad may not be the most important predictor of pupil size. As Figure 2(b) shows, the average brightness is different for all but Ad1 and Ad2. The connected red line shows the average brightness value for each ad, which is given here along with the standard deviation (average contrast). For example, while there is a significant difference in pupil dilation between Ad1 and Ad2, there is no difference between the brightness levels of the two ads (Ad1: M=.26, SD =.37; Ad2: M=.26, SD =.26). On the other hand, while there is not much difference in pupil dilation between Ad2 and Ad3, there is a significant difference in brightness (Ad2: M=.26, SD =.26; Ad3: M=.48, SD =1.8). While the difference in brightness between Ad4 (M=.6, SD =1.9) and Ad3 (M=.48, SD =1.8) is smaller than between Ad2 and Ad3, the pupil dilation is significantly larger for Ad4 compared to all the other ads. From these results, it appears that there is no independent effect of brightness on pupil dilation.

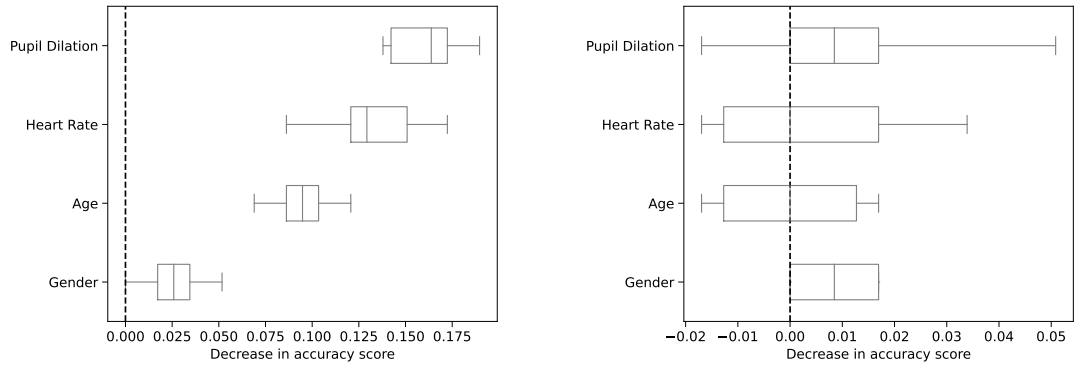
4.4. Heart Rate and Pupil Dilation as Predictors of User Engagement

Several classifiers were used without optimizing the parameters of the models: Logistic Regression, Support Vector Classifier, Decision Tree, Random Forest, Gaussian Naive Bayes, and AdaBoost. Features 'Age', 'Gender, 'Heart Rate, and 'Pupil Dilation' were used as predictors, while 'Engagement Score' was used as a binary target. The decision to model engagement as a binary score was due to the bimodal distribution of engagement scores. The midpoint of the 5-point scale was set as the threshold for the binary engagement classes: lower engagement class < 2.5 vs. higher engagement class > =2.5. Because the target distribution has a relatively large class imbalance (only 22% of the target represents the 'lower engagement' class) and the data set is small, repeated stratified cross-validation (folds=10, repeats=5, fixed number of seeds) was used to preserve the distribution of samples for each target class. Consequently, model performance was evaluated primarily using F1 scores and "weighted" F1 to account for class imbalance, as these provide more reliable evaluation metrics when data are imbalanced. Table 2 shows the evaluation metrics for each model. Logistic Regression, Support Vector Classifier, and Random Forest Classifier achieved the highest overall scores.

For the feature evaluation, the permutation importance metric was used because of the imbalaned class distribution of the target variable. Another advantage of the permutation-based estimation is the ability to assess whether individual feature is useful as a predictor for the test set. Permutation importance was calculated based on the Random Forest Classifier. Figure 3 shows the permutation importance of the predictors for the training set (a) and the test set (b). We can see how the importance of each predictor changes.

Table 2
Classifier Performance Evaluation. Classifiers: LogReg: Logistic Regression, SVC: Support Vector Classifier, DT: Decision Tree, RFC: Random Forest, GNB: Gaussian Naive Bayes,Ada: AdaBoost

	LogReg	SVC	DT	RFC	GNB	Ada	Best Score
Accuracy	0.82	0.82	0.68	0.8	0.81	0.73	Logistic Regression
Precision	0.82	0.82	0.83	0.84	0.82	0.83	Random Forest
Recall	1	1	0.78	0.95	0.99	0.86	Logistic Regression
F1 Score	0.9	0.9	0.8	0.89	0.89	0.84	Logistic Regression
F1_weighted	0.74	0.74	0.69	0.76	0.73	0.71	Random Forest



(a) Permutation importance: Train set

(b) Permutation importance: Test set

Figure 3: Feature estimation with permutation importance by Random Forest Classifier. (a) The permutation importance of features on the training set. (b) The permutation importance of features on the test set.

5. Discussion and Conclusion

The evaluation metrics for several classifiers provide good results and show potential for predicting ad engagement based on pupil dilation and heart rate. However, the presented models still need to be evaluated on a larger number of ads to reliably assess their robustness and predictive power. An interesting finding of the presented study is the strong negative correlation between the dimensions AE and PU, which is consistent throughout all the ads. These differences are partially lost in the combined engagement score. In our future work, we will evaluate both dimensions separately.

The results on the effects of brightness on pupil dilation are inconclusive, as no independent effect of brightness on pupil dilation was found across the ads. Further research is needed to evaluate the effects of brightness, again with a larger sample of ads and specifically in the context of the effects of ad exposure.

References

[1] T. Araujo, J. R. Copulsky, J. L. Hayes, S. J. Kim, J. Srivastava, From purchasing exposure to fostering engagement: Brand–consumer experiences in the emerging computational advertising landscape, Journal of Advertising 49 (2020) 428–445. doi:10.1080/00913367.2020.1795756.

[2] B. J. Calder, M. S. Isaac, E. C. Malthouse, How to capture consumer experiences: A context-specific approach to measuring engagement, Journal of Advertising Research 56 (2016) 39–52. doi:10.2501/JAR-2015-028.

[3] The Nielsen Company, Beyond SVOD, The Nielsen Company, Technical Report, The Nielsen Company, 2020. URL: https://www.nielsen.com/us/en/insights/report/2020/ad-supported-streaming-is-starting-to-stand-out-as-video-options-multiply/.

[4] H. L. O'Brien, P. Cairns, M. Hall, A practical approach to measuring user engagement with the refined user engagement scale (ues) and new ues short form, International Journal of Human-Computer Studies 112 (2018) 28–39. URL: https://www.sciencedirect.com/science/article/pii/S1071581918300041. doi:https://doi.org/10.1016/j.ijhcs.2018.01.004.

[5] A. Munsch, Millennial and generation z digital marketing communication and advertising effectiveness: A qualitative exploration, Journal of Global Scholars of Marketing Science 31 (2021) 10–29. doi:10.1080/21639159.2020.1808812.

[6] E. Eijlers, M. A. S. Boksem, A. Smidts, Measuring neural arousal for advertisements and its relationship with advertising success, Frontiers in Neuroscience 14 (2020). URL: https://www.frontiersin.org/articles/10.3389/fnins.2020.00736. doi:10.3389/fnins.2020.00736.

[7] D. C. Richardson, N. K. Griffin, L. Zaki, A. Stephenson, J. Yan, T. Curry, R. Noble, J. Hogan, J. I. Skipper, J. T. Devlin, Engagement in video and audio narratives: contrasting self-report and physiological measures, Scientific Reports 10 (2020). doi:https://doi.org/10.1038/s41598-020-68253-2.

[8] P. Ayres, J. Y. Lee, F. Paas, J. J. G. van Merriënboer, The validity of physiological measures to identify differences in intrinsic cognitive load, Frontiers in Psychology 12 (2021). doi:https://doi.org/10.3389/fpsyg.2021.702538.

[9] P. van der Wel, H. van Steenbergen, Pupil dilation as an index of effort in cognitive control tasks: A review, Psychonomic Bulletin & Review 25 (2018) 2005–2015. URL: https://doi.org/10.3758/s13423-018-1432-y. doi:10.3758/s13423-018-1432-y.

[10] P. Raiturkar, A. Kleinsmith, A. Keil, A. Banerjee, E. Jain, Decoupling light reflex from pupillary dilation to measure emotional arousal in videos (2016). doi:10.1145/2931002.2931009.

[11] P. Jerčić, C. Sennersten, C. Lindley, Modeling cognitive load and physiological arousal through pupil diameter and heart rate, Multimedia Tools and Applications 79 (2018) 3145–3159. doi:10.1007/s11042-018-6518-z.

[12] D. C. Orazi, G. Nyilasy, Straight to the heart of your target audience, Journal of Advertising Research 59 (2019) 137–141. doi:10.2501/JAR-2019-020.

[13] T. Pham, Z. J. Lau, S. H. A. Chen, D. Makowski, Heart rate variability in psychology: A review of hrv indices and an analysis tutorial, Sensors 21 (2021). URL: https://www.mdpi.com/1424-8220/21/12/3998. doi:10.3390/s21123998.

[14] F. Shaffer, J. P. Ginsberg, An overview of heart rate variability metrics and norms, Frontiers

in Public Health 5 (2017). URL: https://www.frontiersin.org/articles/10.3389/fpubh.2017.00258. doi:10.3389/fpubh.2017.00258.

[15] T. Dobovšek, Š. Bernik, G. Strle, Pupil dilation and heart rate as responses to ad exposure, in: Proceedings of the MEi: CogSci Conference, volume 16, 2022. URL: https://journals.phl.univie.ac.at/meicogsci/article/view/400.

[16] J. Beatty, B. Lucero-Wagoner, The pupillary system, 2000.