Predicting the Optimal Time for Interruption using Pupillary Data and Classification

Hagit Shaposhnik (h.shaposhnik@rug.nl)
Jelmer P. Borst (j.p.borst@rug.nl)
Niels A. Taatgen (n.a.taatgen@rug.nl)
Department of Artificial Intelligence, University of Groningen
Nijenborgh 9, 9747 AG, Groningen, the Netherlands

Abstract

In the current study we present an air traffic control (ATC) task in which we measured pupil dilation to automatically determine high and low workload periods. We manipulated working memory (WM) requirements across three conditions: a no WM condition, a passive WM condition in which information was accumulated, and an active WM condition in which information had to be added to and removed from WM. Results showed that no WM resulted in the least dilation, but that passive WM and active WM did not differ. Next, we used the pupil data to train a range of classifiers to differentiate between high and low workload periods with the ultimate goal to create an online task-independent interruption management system (IMS). The best predicting features were the median and a second-order polynomial fit, going back 12 seconds from the to-be-predicted moment. Using these features, our classifier was able to predict workload at high accuracy (77%). We conclude that pupil dilation can be used to create a reliable IMS.

Keywords: Working memory; Interruptions; Multitasking; Pupil dilation; Machine learning.

Introduction

Nowadays, we are interrupted continuously throughout the day. Especially interruptions in the middle of a task are known to have considerable costs. For example, during office work people are often interrupted by notifications on their smartphone, which disrupts their focus and can lead to large resumption costs. In certain work environments, interruptions are part of the normal work flow and cannot be avoided. For instance, air traffic controllers (ATC) follow aircraft traffic while at the same time instructing pilots and communicating with other controllers on the ground. In this case, mistakes due to interruptions may lead to fatal accidents. To reduce these potentially high costs of interruptions, the main goal of the current study is to develop a robust algorithm to automatically determine the best moment for interruptions. To this end, we employ pupillary data and machine learning techniques.

Previous studies have shown that the extent to which performance on a primary task is affected by an interrupting task depends on the degree of cognitive load in the primary task (Iqbal & Bailey, 2005). Moreover, interruptions during high workload increase the duration of the resumption process to return to the primary task (Altmann & Trafton, 2007; Altmann, Trafton, & Hambrick, 2014; Mark, Gonzalez & Harris, 2005). Multiple studies have shown that the less disruptive moment to present an interrupting task is between tasks rather than in the middle of a task and more specifically in low workload periods (Borst, Taatgen & van Rijn, 2015; Iqbal & Bailey, 2005, 2006; Katidioti & Taatgen, 2014; Monk, Boehm-Davis & Trafton, 2004; Salvucci & Bogunovich, 2010). Thus, interruptions during low workload moments have limited costs compared to high workload moments. Therefore, if we had an automatic way of determining workload, we could schedule interruptions at less disruptive moments.

One way of determining workload is by measuring pupil size. It has long been known that cognitive workload, and especially working memory load (WM), causes the pupil to dilate. For example, Kahneman and Beatty (1966) asked participants to report from a string of a memorized list of different digits. Results showed that pupil dilation increased for each additional digit, and after the last digit decreased again (for a similar recent study, see Karatekin, 2004).

As these results suggest, a number of studies have shown that interruptions during moments of high pupil dilation – and thus high workload – are more disruptive than interruptions when pupil size was small (Iqbal, Adamczyk, Zheng & Bailey, 2005; Katidioti, Borst, Bierens de Haan, et al., 2016). Based on these results, Katidioti, Borst, van Vugt & Taatgen (2016) designed a rudimentary interruption management system (IMS), and demonstrated that interruptions based on pupillary data resulted in better overall performance than self-interruptions.

Given that interruption can be disruptive and affects one's performance, Züger and Fritz (2015) were interested to measure the interruptibility of programmers. In their study, they did not use pupil dilation data, but other physiological sensors such as EEG, eye blinks, heart rate, BVP, EDA (measuring the activity of heart), and body temperature. In addition, they used machine-learning techniques to identify the programmer's interruptibility state. They found that based on these measurements their classifier identified the programmers' state of interruptability with high accuracy, which implies that this kind of classifier might be in used to automatically schedule interruptions during low workload periods. While this is promising, it seems infeasible to measure EEG and heart rate variability in a real environment, which is why we decided to concentrate on pupil dilation measurements.

Current Study

The main goal of the current study was to classify high and low workload periods based on pupillary data and improve the simple threshold technique of Katidioti and colleagues (2016). To this end, we designed an experiment that simulated a simplified ATC environment. We aimed to compare three conditions with different levels of workload: a condition with no WM requirements, one with a passive WM load, and one with an active WM load. In the no-WM condition no memorization was required and decisions wer based on a given rule. In the passive WM task information was accumulated and decision-making had to be based on previous information. In the active WM task information had to be updated several times throughout the task.

Our goal was to determine whether we can differentiate between periods in which participants had to make active decisions versus periods in which they had to wait for the next series of queries. To this end, we trained a classifier to make an online assessment of workload.

Method

Subjects

Twenty-five students from the University of Groningen participated in the experiment for monetary compensation of 14 euros. Data of one participant was not analyzed because of recording problems. Data of three other participants were excluded due to an error in the experimental code. Finally, one participant was excluded due to excessive eye blinks. This leaves 20 participants (12 females and 8 males, mean age 24.5 (range 20-30), *SD* = 2.7). All were right-handed, and had normal vision.

Design

In the experiment participants interacted with a simple air traffic control simulator (Figure 1). Each trial lasted 100 sec and was split into two on-task phases and one off-task phase. The trials started with a fixation cross in the middle of a centered circle (10 cm diameter) for 2 sec, which was followed by the appearance of aircraft entering the screen (i.e., airspace). Aircraft that were entering into the airspace flew at a constant speed. Some of the aircraft flew into the circle and others continued to fly outside of the circle, which represent the airspace the controller is responsible for. Aircraft that entered into the circle presented a request. The requests were either an altitude or a speed change, each which a specific new speed or altitude. Altitude requests consisted of 4 digits and speed requests of 2 digits. Requests were presented for 1.5 sec. The time between the offset of one request and the onset of the next request was 2.5 sec. The first on-task phase ended when aircraft started to leave the airspace, after which no aircraft entered for 26 sec (the off-task phase). After the off-task phase, a new on-task phase started in which aircraft entered the airspace again. After the second on-task phase the trial ended.

The experiment consisted of three conditions: no WM, passive WM, and active WM. In the no-WM condition, six aircraft entered into the circle one by one, and each aircraft presented a *speed request*. If the requested value was smaller than 35 it should be rejected, otherwise it should be

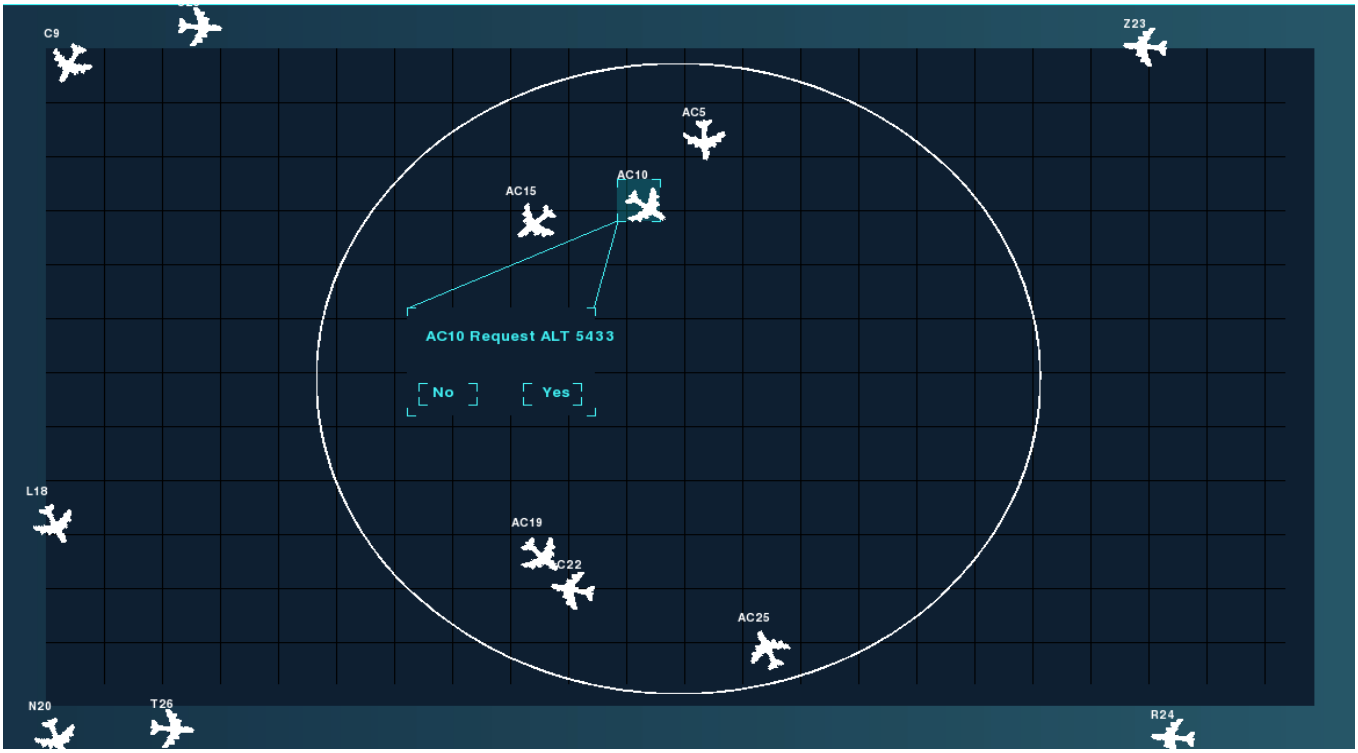

Figure 1: The ATC experiment. Participants had to respond to airplanes inside the circle. Here, the altitude change request should be allowed if none of the other 5 planes is at altitude 5433 – information participants had to maintain in their WM.

accepted. Thus, memorization of the requests was not required.

In the passive WM condition, six aircraft entered the airspace circle and presented *altitude requests* one at the time. Participants were required to compare each request to the previous requests by the other aircraft in the circle. If a requested altitude was already occupied by another aircraft, it should be rejected, otherwise it should be accepted. Thus, previous requests should be maintained in WM.

In the active WM condition 4 aircraft entered into the airspace circle and presented altitude requests one at the time, similar to the passive WM condition. After that, two aircraft would leave the airspace and two additional aircraft would enter the airspace circle, and present their request. Similar to the passive WM condition, no aircraft within the airspace could fly at the same altitude. However, altitudes can become available again if the aircraft at a particular altitude left the airspace. This meant that participants needed to actively update their WM when aircraft were leaving.

Participants had to press 'Z' to reject request and '/' to accept. Participants were required to respond while the request was still presented on the screen (1.5 sec). Following the response feedback was given for 1 sec. If the response was correct, the aircraft was colored by a green square, if it was incorrect the aircraft was colored by a red square. The study consisted of 36 trials grouped in 6 blocks. Within a block, two trials in each of the three conditions were administered in random order.

Procedure

Participants were tested individually in a windowless room containing a desk on which a monitor, eye-tracker camera and chinrest were placed. The seating distance from the 20-inch LCD monitor (1600×1200, 60 Hz) to the chinrest with forehead support was 59 cm.

The room was illuminated using a ceiling lamp, resulting in ambient light that provided a comfortable level of luminance to participants. Eye position and pupil dilation of either left or right eye, depending on each participant's dominant eye, was measured at the sampling rate of 500 Hz using an SR Eyelink 1000 eye tracker. Calibration and drift correction were performed before the experiment and after each break, using a randomized target order with 9 points.

Before staring the experiment, participants gave informed consent. After reading the experimental instruction a verbal instruction was provided to ensure that they understood the task. Participants started with a practice block that contained all three conditions. If, after three practice trials, participants did not understand the task, they were required to repeat it. Afterwards, the experiment started.

Analysis and Classifier

To create the optimal predictor for a potential IMS, we trained a classifier with pupillary data to identify different workload periods. Before implementing the classifier, the pupillary data were preprocessed. First, in order to reduce artifacts, saccades and blinks were detected and replaced by quadratic-interpolation after extending the rejection area with 50 msec before and after the saccades and 100 msec before and after the blinks. Further, pupil dilation was down-sampled from 500 Hz to 50 Hz, and normalized by a moving-average baseline of the last 20 seconds[1].

To analyze the data, we used linear mixed effects models; models were compared using chi-squared likelihood tests. Contrast testing was performed with Tukey Post-Hoc tests. Participants and phase were submitted as random effects.

To classify high and low workload moments, we used binomial logistic regression. Classifiers were trained and tested with 10-fold cross-validation within-subject. To predict workload, we used features from the pupillary data. Features were created by splitting the data into windows of 2, 3, or 4 seconds. From these windows, we calculated the median, SD, and a second order polynomial fit. We first thought to examine the slope of each curve but since our pupillary data is not linear, we decided that polynomial will fit better to our data. For the second order polynomial fit we used the following function: $p(x) = p1x^2 + p2x$,the fitting method attempt to fit the best possible coefficients (p1, p2) to the given curve at each window using the least squares polynomial fitting algorithm. Thus, for each window we had four features: the median, the SD, and the two coefficients of the polynomial fit.

Results

Behavioral Data

Figure 2 shows the average accuracy and average RTs for each condition. The results show that participants performed well on 3 conditions with a mean proportion of correct responses of 0.93 (*SD* = 0.1). RTs were measured from the moment of the request to the moment of response. Results show that mean RTs were 0.65 (*SD* = 0.07) for the 3 conditions

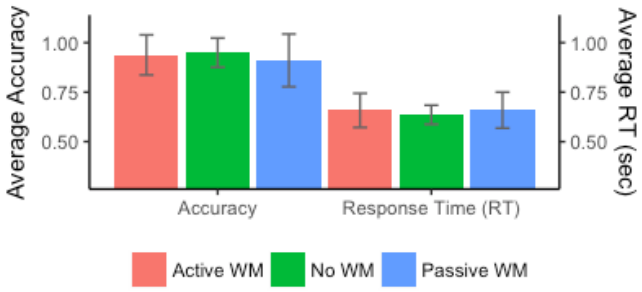

Figure 2: Average Accuracy and Average Response Time (RT) per condition.

[1] We used a moving-average baseline because the final goal is to design an IMS that does online interruption management.

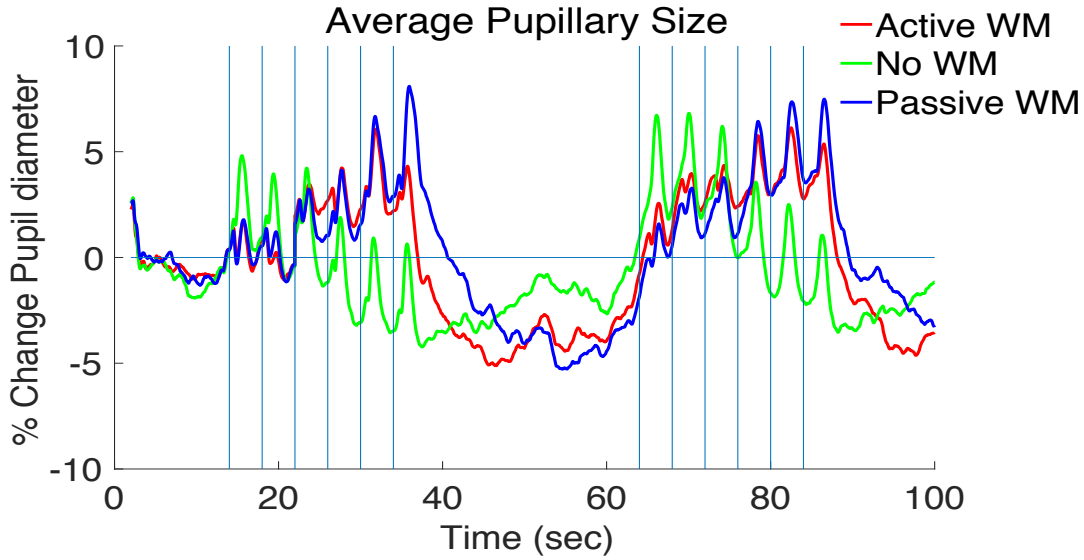

Figure 3: Percentage change in pupil size measured for each condition across time.

Pupillary Data

Figure 3 shows percentage change in pupil size measured for each condition across time. Each vertical line represents a request (1 to 6) in each phase, the white gap represents the off-task interval period between the two on-task phases. The red line represents the active WM task, the green line represents the no-WM task, and the blue line represents the passive WM task. The initial dilation of the pupil started after each request and reached a peak after 1.5 sec to 2 sec and reduced again afterwards. In both WM conditions, pupil response increased gradually as a function of time. A more pronounced increase followed the 5th and the 6th request, after which it decrease below the baseline during the off-task period. In the control task (no-WM) the pupil response was initially higher than in both other conditions until the 3rd request, but then reduced almost below the baseline.

To analyze the pupillary data, we calculated the average pupillary peak response for each condition across the two response phases (Figure 4; there was no difference in pupillary response between phases; $\beta = 0.5$, $\chi^2(1) = 0$, p=1). First we found a main effect of condition on pupillary response that showed a significant difference between both WM tasks and the no-WM task ($\beta = -0.170$, $\chi^2(1) = 35.3$, p<0.001). There was no difference between the active WM and passive WM tasks ($\beta = -0.185$, $\chi^2(1) = 0.586$, p=0.443). The effect of Request was found to be significant ($\beta = -0.471$, $\chi^2(1) = 34.15$, p<0.001). Tukey post-hoc testing indicated a different pupil response to requests in the active WM conditions and the no-WM condition ($\beta = 0.342$, SE = 0.626; z = -8.297, $p_{adjusted} < 0.001$; $\beta = 0.342$, SE = 0.626; z = -9.93, $p_{adjusted} < 0.001$).

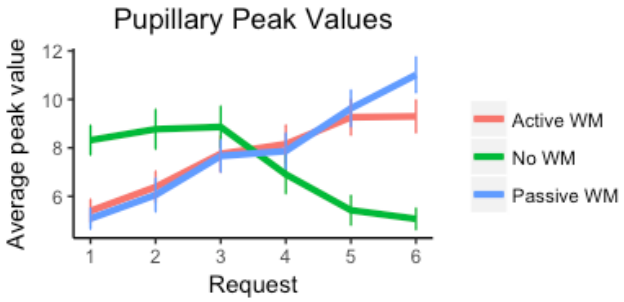

Figure 4: Average pupillary peak response for each condition as a function of request.

Classification

The classifier was trained to distinguish between two periods; the on-task period, which included the requests, and the off-task period in which participants did not perform a task for a period of approximately 26 sec. As explained above, 10-fold cross validation was performed for each subject.

Table 1. Results of the classifier (average accuracy + range) for four different feature combinations, for windows sizes of 2, 3, and 4 seconds.

	Median			**Median + SD**			**Polynomial**			**Median + Polynomial**		
Condition	4 s	3 s	2 s	4 s	3 s	2 s	4 s	3 s	2 s	4 s	3 s	2 s
Active WM	73% (49-90)	71% (50-86)	72% (50-87)	73% (50-90)	72% (47-86)	73% (49-88)	77% (52-91)	72% (47-87)	73% (55-88)	77% (50-90)	72% (49-85)	73% (56-88)
Passive WM	74% (54-92)	72% (52-88)	72% (52-89)	70% (48-84)	67% (47-82)	68% (52-83)	76% (53-91)	68% (54-83)	70% (54-83)	76% (55-91)	72% (51-83)	72% (53-83)
No WM	69% (47-83)	66% (43-81)	67% (52-84)	73% (50-90)	72% (47-86)	73% (49-88)	77% (52-91)	72% (47-87)	73% (55-88)	77% (50-90)	72% (49-85)	73% (56-88)
All	**70%** (49-84)	**69%** (50-84)	**69%** (50-84)	**70%** (48-84)	**67%** (47-82)	**68%** (52-83)	**76%** (53-91)	**68%** (54-83)	**70%** (54-83)	**77%** (55-91)	**72%** (51-83)	**72%** (53-83)

We explored which features yield the best classification results. Table 1 shows the overall mean and min-max rates for different features and window combinations. The first observation we can make is that larger windows on average gave the best predictions, that is, it was more effective to go 12 second back in time than less. Second, using the polynomial fit in each window improved performance over using either the median, or the median and the SD. Thus, not only average dilation and its variance are informative, but also the pattern of dilation inside each window. Third, the hardest condition to classify was the no-WM condition, presumably because the difference between on- and off-task phases was the smallest in this condition (Fig. 3).

The best feature combination were three four-second windows, for each of which we used the median and a two-coefficient polynomial fit. On average, this resulted in 77% correct classifications. Figure 5 shows that for a majority of

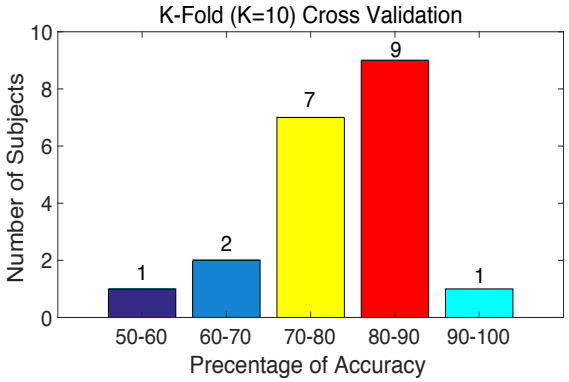

Figure 5: Distribution of classifier accuracy over subjects.

the participants' classification was even over 80% in this case, and only one participant scored below 60%.

Discussion

The main goal of the present study was to develop a classifier that will predict and differentiate between low and high workload periods using machine learning techniques based on pupillary data. In order to do so we designed a ATC task that simulated a real world scenario and included different levels of workload. Based on previous studies we hypothesized that pupillary size would increase with increasing memory load and decrease with decreasing load. The results confirmed the hypotheses; dilation increased with increased WM load and was smaller when WM load decreased. Figure 3 clearly shows the change in pupil dilation during the period of on-task and off-task. Additionally, it can be observed that during the on-task pupil size increase proximally 1.5 to 2 seconds after each request and decrees afterward.

Based on these results, we developed a classifier. In order to identify high and low workload periods, we applied binominal logistic regression to the pupillary data. Table 1 shows that it is possible to classify high and low workload periods based on pupillary data. Using large windows, going back 12 seconds in time, including the median value and a polynomial fit per window, gave the best predictions for our model. This indicates that not only the level of pupil dilation (median) is relevant to determine workload, but also the direction of the change in pupil dilation. We hypothesize that decreasing dilation indicates the start of a low-workload period, whereas increasing dilation signal the start of a high-workload period.

Figure 5 shows the accuracy of the model for each participant. These results show that the classifier could classify above chance for all subjects, and with a high accuracy for the large majority. Different from our study in which we focused only on pupillary data, Züger and Fritz (2015) used several physiological sensors. On the one hand, this provided more features for the classifier, but it also makes the classifier less practical to use in a real work environment, while pupil dilation might be measured with high-end webcams (Rafiqi, Wangwiwattana, Fernandez, Nair, & Larson, 2015). Performance of the two classifiers was comparable, where we reached 77% accuracy on average, Züger and Fritz (2015) classifier reached 75%.

A notable feature of the pupillary pattern observed in this study was the way pupil size was affected by the number of requests. Figure 4 indicates the difference in peak size between the no-WM condition and both WM conditions as a function of request. At the first 3 requests, the average peak size was larger during in the no-WM condition, but it then decreased sharply compared to both WM conditions. In

contrast, in the WM conditions the average peak response increased gradually as a function of the number of requests. This implies that the amount of information that is stored in memory evoked changes in pupillary size.

We might assume that the average of peak was larger in the no-WM condition during the first 3 requests because participants had to decide if a presented value was higher or lower than the given threshold. This may be confusing at the beginning but once participants got used to the drill it becomes an easy task. On the contrary, in the WM conditions the answer to the first request was automatically 'YES', since participants were not required to compare it to any previous value and no storage or decision-making was required. Yet, as the task continued more information was stored in WM and more decision-making was required, which affected cognitive workload and pupil size.

To conclude, in this study we have shown that it is possible to use pupil dilation to determine high and low workload periods. These results will form the basis an online task-independent IMS. As our next step, we aim to fit our model across participants. Additionally, we would like to identify the high and low load periods during the on-task phase, to enable more fine-grained interruption management.

Acknowledgements

This research was supported by AFOSR grant FA9550-17-1-0309 awarded to Niels Taatgen and Jelmer Borst.

References

Altmann, E.M., & Trafton, J.G. (2007). Timecourse of recovery from task interruption: data and a model. *Psychonomic Bulletin & Review 14* , 6, 1079–84.

Altmann, E.M., Trafton, J.G., & Hambrick, D.Z. (2014). Momentary interruptions can derail the train of thought. *JEP: General 143*, 1, 215–226.

Borst, J. P., Taatgen, N. A., & Van Rijn, H. (2015). What Makes Interruptions Disruptive? A Process-Model Account of the Effects of the Problem State Bottleneck on Task Interruption and Resumption. *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: CHI 2015*. New York: ACM.

Hess, E. H., & Polt, J. M. (1960). Pupil size as related to interest value of visual stimuli. *Science, 132,* 349–350.

Iqbal, S. T., & Bailey, B. P. (2005). Investigating the effectiveness of mental workload as a predictor of opportune moments for interruption. *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: CHI 2005* (pp. 1489–1492).

Iqbal, S.T. and Bailey, B.P. (2006). Leveraging characteristics of task structure to predict the cost of interruption. *In Proceedings of CHI 2006* (pp.741–750).

Iqbal, S.T., Adamczyk, P., Zheng, X., and Bailey, B.P. (2005). Towards an index of opportunity: understanding changes in mental workload during task execution. In *Proceedings of the SIGCHI. 2005*, 311–320.

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

Karatekin, C. (2004). Development of attentional allocation in the dual task paradigm. *International Journal of Psychophysiology 52*, 7–21.

Katidioti, I., Borst, J.P., Bierens de Haan, D.J., Pepping, T., Van Vugt, M.K., & Taatgen, N.A. (2016). Interrupted by your Pupil: An Interruption Management System based on Pupil Dilation. *International Journal of Human-Computer Interaction 32*(10), 791-801.

Katidioti, I., Borst, J. P., & Taatgen, N. A. (2014). What happens when we switch tasks: Pupil dilation in multitasking. *Journal of Experimental Psychology: Applied, 20(4)*, 380–396.

Katidioti, I., Borst, J. P., Van Vugt, M. K., & Taatgen, N. A. (2016). Interrupt me: External interruptions are less disruptive than self-interruptions. *Computers in Human Behavior*, *63*, 906–915.

Mark, G., Gonzalez, V., & Harris, J. (2005). No task left behind? Examining the nature of fragmented work. In *CHI 2005 Proceedings* (pp. 321–330), ACM Press.

Monk, C. A., Boehm-Davis, D. A., & Trafton, J. G. (2004). Recovering from interruptions: Implications for driver distraction research. *Human Factors*, *46*, 650–663.

Rafiqi, S., Wangwiwattana, C., Fernandez, E., Nair, S., & Larson, E. C. (2015). Work-in-progress, PupilWare-M: Cognitive Load Estimation Using Unmodified Smartphone Cameras. MASS 2015*,* SocialSens 2015.

Salvucci, D. D., & Bogunovich, P. (2010). Multitasking and monotasking: The effects of mental workload on deferred task interruptions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.*

Züger, M., & Fritz, T. (2015). Interruptibility of software developers and its prediction using psycho-physiological sensors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '15), Seoul, South Korea.*