INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION
http://dx.doi.org/10.1080/10447318.2016.1198525

Taylor & Francis
Taylor & Francis Group

Interrupted by Your Pupil: An Interruption Management System Based on Pupil Dilation

Ioanna Katidioti, Jelmer P. Borst, Douwe J. Bierens de Haan, Tamara Pepping, Marieke K. van Vugt, and Niels A. Taatgen

Department of Artificial Intelligence, University of Groningen, Groningen, The Netherlands

ABSTRACT

Interruptions are prevalent in everyday life and can be very disruptive. An important factor that affects the level of disruptiveness is the timing of the interruption: Interruptions at low-workload moments are known to be less disruptive than interruptions at high-workload moments. In this study, we developed a task-independent interruption management system (IMS) that interrupts users at low-workload moments in order to minimize the disruptiveness of interruptions. The IMS identifies low-workload moments in real time by measuring users' pupil dilation, which is a well-known indicator of workload. Using an experimental setup we showed that the IMS succeeded in finding the optimal moments for interruptions and marginally improved performance. Because our IMS is task-independent—it does not require a task analysis—it can be broadly applied.

1. Introduction

Nowadays it is nearly impossible for a work environment to be free of interruptions. Interruptions are often part of the job itself: it is hard to imagine a profession in which one focuses only on a single task for an extended amount of time. A pilot has to talk to the air traffic controller while operating a plane, a professor has to answer a student's question while giving a lecture, and a receptionist has to answer the phone in the middle of providing information to a visitor.

The prevalence of interruptions has been quantified by several observational studies that show how often people are interrupted in their workplace (e.g., Czerwinski, Horvitz, & Wilhite, 2004; Eyrolle & Cellier, 2000; Gonzalez & Mark, 2004). For example, Gonzalez and Mark's (2004) study revealed that office workers switched tasks every 3 minutes. In addition—and perhaps more worrisome—many other studies have shown that interruptions are very disruptive for the main task: users make more errors and become slower when interrupted (e.g., Edward & Gronlund, 1998; Gould, Brumby, & Cox, 2013; Hodgetts & Jones, 2006; Iqbal & Bailey, 2007; Jin & Dabbish, 2009). As it seems impossible to ban interruptions from the workplace, it is crucial to find a way of managing interruptions that minimizes their negative effects.

With this goal in mind, we developed a task-independent interruption management system (IMS) that uses real-time pupil dilation measurements to interrupt users at the least disruptive moments of a task. We evaluated this system in a lab study, and showed that our IMS is able to find the optimal moments for interruptions. Before we describe our IMS in detail, we will first discuss what aspects of interruptions affect their disruptiveness and what kind of IMS's have been developed previously. We will then show how our IMS is able to identify the optimal interruption moments on the basis of pupil size independent of the current task.

1.1. Background

One of the main theories on interruptions is Memory for Goals (Altmann & Trafton, 2002). In this theory, each task is characterized by a goal, which has a certain activation level. When a task is interrupted, its goal is stored in memory and starts decaying. In the meantime, the goal of the interrupting task is activated. Returning to the main task entails resumption of its goal. The longer the interruption lasts, the more the goal has decayed in declarative memory, and the harder it is to resume the main task.

In order to cover more of the factors that can affect the disruptive effect of interruptions, Borst, Taatgen, and van Rijn (2015) extended Memory for Goals theory to Memory for Problem States. Instead of goals, this model focuses on problem states. The problem state contains the information that is necessary to complete the next steps in a task, e.g., when trying to find the value for x in an equation such as $2x + 4 = 14$, the information $2x - 10$ is stored in the problem state before proceeding to $10/2 = 5$. The main idea in Memory for Problem States is the same as in Memory for Goals: when the main task is interrupted, its problem state is stored and starts decaying. However, if the main or the interrupting task does not require a problem state, the main task will not be hard to resume even if a considerable amount of time has passed.

CONTACT Ioanna Katidioti i.katidioti@rug.nl Department of Artificial Intelligence, University of Groningen, 9747 AG Groningen, The Netherlands.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/hihc.

Memory for Problem States accounts for most of the factors that can affect the disruptive effect of an interruption. It is well known that interruptions disrupt the main task and affect performance (see for instance the seminal work by Gillie & Broadbent, 1989). However, there are multiple factors that affect the level of disruptiveness of interruptions. Several studies showed that a long interruption is more disruptive than a short one (e.g., Borst et al., 2015; Hodgetts & Jones, 2006; Monk, Boehm Davis, & Trafton, 2004), which is something that both Memory for Goals (Altmann & Trafton, 2002) and Memory of Problem States account for, since the longer the goal or state of a task has to be stored, the harder it is to resume it. Other studies suggest that a more complex interruption is more disruptive than a simpler one (e.g., Borst et al., 2015; Cades, Boehm Davis, Trafton, & Monk, 2007; Monk et al., 2004). In addition, an interruption different from the main task is more disruptive than an interruption more similar to the main task (e.g., Gould et al., 2013; Iqbal & Bailey, 2008). The timing of the interruption during the main task also plays an important role. Interruptions at high-workload moments (typically in the middle of a (sub) task) are more disruptive than interruptions during low-workload moments (between (sub)tasks; Gould et al., 2013; Iqbal & Bailey, 2005, 2006; Katidioti & Taatgen, 2014; Monk et al., 2004). In this article, we will focus on minimizing the negative effects of interruptions by adjusting their timing.

In one of the studies focusing on the timing of interruptions, Gould and colleagues (2013) interrupted participants during a data-entry task either mid-subtask or between subtasks, with the former interruptions being more disruptive than the latter. Similar results were found by Monk and colleagues (2004) in a VCR programming task. Iqbal and Bailey (2006) used GOMS modeling to find high- and low-workload moments in three different tasks. They interrupted users at these moments, and showed that the cost of interruptions during low-workload moments was smaller than during high-workload moments.

In most of these studies, high workload was defined as participants' working memory being occupied and low workload as their working memory being unoccupied with task information. To test whether working memory requirements indeed play an important role in the disruptiveness of interruptions, Salvucci and Bogunovich (2010) created a real-life setup with clear high- and low-workload moments determined by working memory requirements (the current experiment is based on their study). In Salvucci and Bogunovich's study, participants simulated a client service employee for an electronics company, by answering emails from fictional clients asking them product prices. Participants had to read the email, look up the price in a browser, and write a response to the client. From time to time, a chat message arrived in the background. Participants were free to choose when to answer these messages. Results showed that participants preferred not to interrupt themselves during high-workload moments, which were the moments they had to remember the product name or the product price, causing their working memory to be occupied.

In a follow-up study by Katidioti and Taatgen (2014) that used the same email-and-chat task, participants were encouraged implicitly through an artificial delay in the main task to switch during high-workload moments in half the experiment. As a result, participants were slower to complete an email than when they switched at low-workload moments. Thus, interruptions at high-workload moments are more disruptive, and workload seems to be strongly dependent on working memory load.

Memory for Goals (Altmann & Trafton, 2002) does not account for the effect the moment of interruption can have on the level of disruptiveness of an interruption, nor for the complexity of the tasks since the main focus of this theory is the effect of the length of the interruption. Memory for Problem States (Borst et al., 2015) can explain the effects of the moment of interruption on its disruptiveness (e.g., Gould et al., 2013; Iqbal & Bailey, 2005, 2006; Katidioti & Taatgen, 2014; Monk et al., 2004). According to Memory for Problem States, if the main task does not require a problem state (and therefore the working memory is not occupied with task information), it will be resumed more easily after an interruption than a main task that requires a problem state, as no problem state has to be retrieved from memory in the former case.

1.2. Managing Interruptions

The fact that interruptions can be more or less disruptive based on the circumstances can be exploited by IMSs, which aim to find optimal—least disruptive—points for interruptions. McFarlane (2002) was one of the first to exploit this concept, and created an IMS that calculated the workload of the specific task he used and interrupted participants when the workload of the task was low. He then conducted an experiment in which he used a collection of performance and personal preference indices to compare four different kinds of interruptions: immediate, negotiated, mediated and scheduled. His results showed that mediated interruptions (determined by the IMS) were less damaging to performance than scheduled (occurring every 25 s) and immediate (random occurrence) interruptions. Negotiated interruptions (in which the user determined when to be interrupted) were comparable with the mediated interruptions on most indices, but required the user to make the decision when to switch. This suggested that a combination of these two systems would be beneficial for managing disruptions: a mediated system as the default, with the possibility to override the mediator and choose your own moment of interruption.

Arroyo and Selker (2011) developed an IMS that focused on the similarity between the interrupting task and the main task. Their IMS allowed relevant interruptions (defined as those with similar content as the main task) to pass, while it held back the irrelevant ones. This system led to a performance benefit for important/urgent tasks. Züger and Fritz (2015) used psycho-physiological measures (EEG data, eye blinks and electrodermal activity) to measure interruptibility of programmers. Although they did not create an IMS, they were able to identify a programmer's state of interruptibility by means of machine-learning classifiers with high accuracy. This suggests that such classifiers could be potentially used in real time to interrupt users at low-workload moments.

Tanaka, Abe, Aoki, and Fujita (2015) and Kobayashi, Tanaka, Aoki, and Fujita (2015) have developed an IMS that uses head motion and computer operations (typing, mouse clicks, opening and closing windows, etc.) in order to identify a user's low-workload moments. Their system already shows very promising results, although it is limited to specific work environments and restricted to tasks that involve clicking, typing and window usage. This means, for example, that their system cannot find an interruption moment if one is reading a paper on a computer screen.

The goal of the current study is to find a way to interrupt people at low-workload moments. Therefore, we need a non-invasive way to measure workload. The studies reviewed above suggest that the best way to create a task-independent IMS is to use a physiological measure. The physiological measure we decided to focus on is pupil dilation.

1.3. Pupil Dilation

Since the 1960s, studies have established that the size of the pupil is not only affected by changes in light, but also by other stimuli. Hess and Polt (1960) were the first to show that pupil size also depends on covert cognitive variables. Nowadays, pupil dilation is known to react to a wide variety of cognitive processes such as task difficulty and working memory load (see Beatty & Lucero-Wagoner, 2000; Laeng, Sirois, & Gredebäck, 2012 for reviews). In general, it is clear that mental workload has a considerable impact on pupil size (e.g., Beatty, 1982; Hoeks & Levelt, 1993). A more difficult task evokes a larger pupil dilation than an easier task (e.g., Beatty & Lucero-Wagoner, 2000; Jennings & van der Molen, 2005). For instance, in a study by Kahneman, Tursk, Shapiro, and Crider (1969), pupil dilation increased as the difficulty of mathematical equations increased.

There are many studies linking pupil dilation to working memory load—an important factor in mental workload. Kahneman and Beatty (1966) report that participants' pupil dilation increased as the number of digits they had to remember increased from 3 to 7. Pupil dilation increased again to the same size when participants had to repeat the digits for the second time. Peavler (1974) measured changes in pupil dilation while participants had to keep strings of 5, 9, and 13 digits in working memory. Results showed that pupil dilation kept increasing until the 7th or 8th digit and then reached an asymptote, reflecting the limits of working memory.

Besides working memory, pupil dilation is used in the study of many different forms of cognitive effort, operationalized as task complexity (e.g., Moresi et al., 2008; Prehn, Heekeren, & Van der Meer, 2011), Stroop interference effects (Laeng, Ørbo, Holmlund, & Miozzo, 2011), or difficulty of retrieving information from memory (van Rijn, Dalenberg, Borst, & Sprenger, 2012). Despite the fact that pupil dilation is widely used in cognitive science, there is one drawback of using it as a real-time measure: there is a delay of approximately 1 s before the pupil reaches its maximum dilation after the onset of a stimulus (e.g., Hoeks & Levelt, 1993).

Pupil dilation has also frequently been used in interruption research (e.g., Iqbal, Adamczyk, Zheng, & Bailey, 2005; Katidioti, Borst, & Taatgen, 2014). For example, Iqbal and colleagues (2005) found that pupil dilation decreased between subtasks, which are low-workload moments. In a follow-up study, Iqbal and Bailey (2005) found that there were smaller time costs when participants were interrupted at those low-workload moments, compared to high-workload or random moments. Combining all their findings, Iqbal and Bailey (2010) created the OASIS IMS, which delays interruptions until there is a natural breakpoint in the task. However, those breakpoints were decided by statistical models, based on behavioral data from previous studies (Iqbal & Bailey, 2007, 2008) and not by real-time changes in pupil dilation. Thus, this system, as does the Arroyo and Selker (2011) IMS, requires a task analysis before it can be used.

The IMS we describe here chooses the optimal moments for interruption based only on changes in pupil dilation, independent of the specific task. When the user's pupil size drops below a certain threshold (which is constantly updated), it is considered a low-workload moment, and the user is interrupted. Since our IMS does not require a task analysis, it is task-independent.

2. Interruption Management System

We developed an IMS that uses real-time changes in pupil dilation to identify the low-workload moments of a task. We then tested it on the email task that is interrupted by chat messages (Katidioti & Taatgen, 2014; Salvucci & Bogunovich, 2010) with minor adjustments to fit the current setup. As the task progresses, the IMS calculates a workload identifier value (WIV). When pupil size is below the WIV it is considered to be a low-workload moment. If there are consecutive low-workload moments for 200 ms,[1] the IMS interrupts the participant (Figure 1). The WIV is calculated by using the following values: the baseline pupil size, the percentage change in pupil size (PCPS; Iqbal et al., 2005), the live average, and a threshold adapter.

The baseline pupil size is measured at the beginning of the study. During the experiment, pupil size is measured continuously and then transformed into PCPS values by subtracting the baseline pupil size from each measurement and dividing the result by the baseline pupil size. In order to avoid multiplication with negative numbers, 1,000 was added to each PCPS value. Thus, PCPS is given as follows:

$$\text{PCPS} = \frac{\text{current pupil dilation} - \text{baseline}}{\text{baseline}} + 1000. \qquad (1)$$

The live average is defined as the average PCPS over the last minute and is used to account for possible changes in pupil dilation due to familiarity with the task, changing head position or changes in light (the latter two did not occur during the experiment).

[1]The 200 ms interval was chosen after pilot studies. Because of blinks, saccades, and noise, a smaller interval might not have provided enough information. A bigger interval might have indicated wrong moments in the current task, which has quick changes from low to high-workload moments.

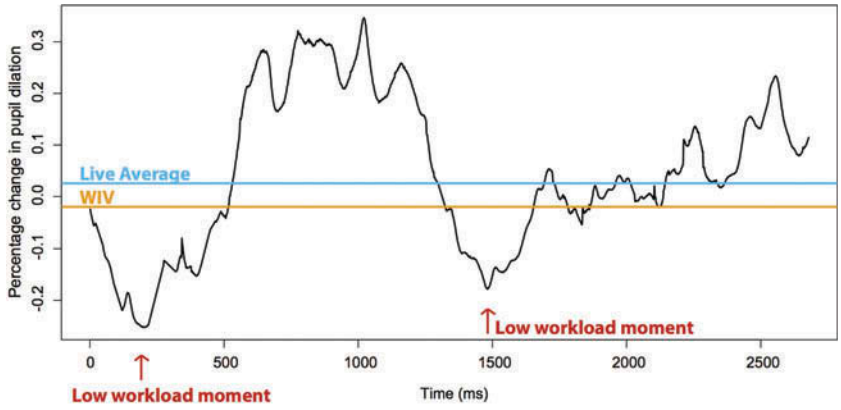

Figure 1. Percentage change in pupil size (PCPS) during a random email sequence. The top line is the value of the live average, the bottom line is the value of the WIV (workload identifier value) and the low workload moments of this email are indicated with arrows. The IMS added 1000 to the PCPS values for calculation reasons, which is not shown in the figure.

The threshold adapter is set to 0.997 at the beginning of the task (the value was chosen after a pilot study). The threshold adapter is multiplied by the live average to calculate the WIV:

$$\text{WIV} = \text{live average} \times \text{threshold adapter}. \quad (2)$$

The IMS allows for an interruption if pupil size measurements are below the WIV for 200 ms consecutively. In order to find the optimal WIV for each participant, the threshold adapter increases or decreases by 0.001 when there are more than one or no interruptions, respectively, during a specific time interval.

During the interruption, pupil measurements are not taken into account by the IMS, because the interruption is a task that may have different characteristics from the main task and pupil measurements may therefore not be representative. In addition, for 5 s after the end of an interruption, there cannot be another interruption. That restriction allows pupil measurement samples to return to baseline. Finally, eye blinks are ignored.

To test the IMS, we performed a lab study that tested whether the IMS could find the optimal interruption moments, and compared its performance to random interruptions and no interruptions at all.

3. Methods

The main and the interrupting task of the experiment were based on Salvucci and Bogunovich (2010). The experiment simulates the working environment of an employee of an electronics company who has to answer clients' emails while being interrupted by chat messages. The main task was the email-answering task and the interrupting task was the chat-answering task. The windows used in the experiment are shown in Figure 2. In the actual experiment, the windows were overlapping and

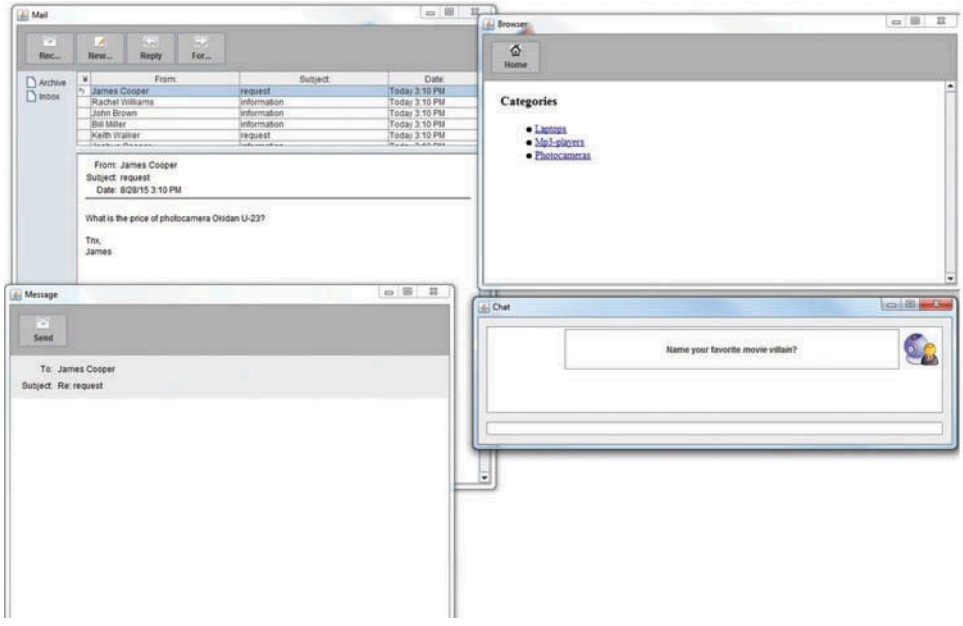

Figure 2. All the windows used in the experiment. In the top left corner there is the Mail window, top right corner is the Browser window, bottom left corner is the Composer window and bottom right corner the Chat window.

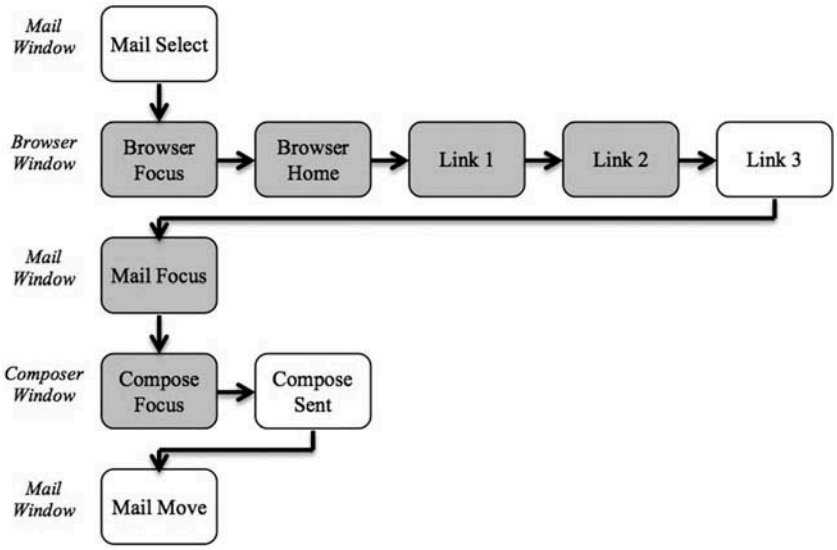

Figure 3. The sequence of the main task. The high-workload moments are indicated with gray and the low-workload moments with white.

could not be moved. The participant had to click on a window in order to see it. This forced them to remember the information in windows that were not currently visible.

The steps of the main task are shown in Figure 3. The participant first opens an email by clicking on it, reads the question (e.g., "What is the price of laptop Zanium A-63?"), goes to the simulated browser, clicks on the product category (Link 1), then the product name (Link 2) and finally the product code (Link 3). After a 2-s delay, the price of the product loaded, the participant could read it, return to the email window and press the "Reply" button. The composer window would appear, the participant had to type the price, press the "Send" button (which caused the composer window to disappear) and then drag and drop the answered email in the Archive folder.

The interrupting task simulated a casual chat conversation. The chat questions were in the form of "What is your favorite…?" (e.g., color, food, movie, book). One in four questions was a follow-up question to the previous one, asking "Which is your least favorite?," which referred to the previous question. We used these follow-up questions to engage the participants more into the simulated conversation. When an interruption occurred, the chat window appeared in front of the other windows and could not be unfocused until the participant responded. Participants were instructed to immediately answer one chat message and then continue with the email task. Although in our experiment the chat questions were of a private nature, this simulates situations in some working environments in which the employees have to give priority to the live-chat questions that clients ask them.

The email task has high- and low-workload moments (see Figure 3). High-workload moments are considered the moments where working memory was occupied by either the product name or the product price. Low-workload moments are defined as the moments that working memory was not occupied by task information. At first, after opening the email, the participant has to memorize the product type and name until finishing the search (Link 3). Link 3 is a low-workload moment, since participants no longer have to retain the product name in working memory. When the price of the product loads, there are again high-workload moments, since participants have to keep the price in their working memory until the answer is sent. A similar task analysis has been used at Salvucci and Bogunovich (2010) and Katidioti and Taatgen (2014). Both these papers have used the same e-mail task (with some small differences in Salvucci & Bogunovich, 2010). Both these papers gave participants the freedom to self-interrupt and results confirm this task analysis of high and low-workload moments.

3.1. Conditions

We used a within-subject repeated-measurement design with three conditions: Control, IMS and Random. During a Control block there were no interruptions: the participant only had to perform the email task to measure baseline performance. In a Random block, interruptions occurred at random moments. At the beginning of the block a random interruption moment between 10 and 30 s in the future was picked. The pilot study showed that the average time to complete an email sequence is about 20 s, so this interval would result in approximately one interruption per e-mail. When the designated moment arrived in the experiment, an interruption occurred. After the interruption finished, a new interruption moment between the next 10 and 30 s was picked for the next interruption, etc.

In the IMS blocks, the moments of the interruptions were determined by the IMS. As in the Random condition, a random moment between 10 s and 30 s was picked. The time

until that moment was the interruption interval. If the IMS detected no suitable interruption moments during the interruption interval, the threshold adapter was increased by 0.001, so that an interruption would be more likely in the next interval. If the IMS interrupted the participant more than once during the interruption interval, the threshold adapter was decreased by 0.001 each time an extra interruption happened. The time spent on the interruptions was added to the interruption interval. When the interval finished, another random moment between 10 and 30 s was picked, etc.

3.2. Apparatus and Setup

Participants were tested individually in a small windowless room. They were seated at a desk with a 20 inch LCD monitor with screen resolution of 1600 × 1200 pixels and screen density of 64 pixels/inch. Participants were asked to use a chin-rest during the experiment. The eyetracker was an Eyelink 1000 from SR Research, positioned approximately 45 cm from the end of the desk.

Eye fixations were measured with a sample rate of 250 Hz. Calibration and drift correction were performed before the experiment started. A calibration accuracy of 0.8° was considered acceptable. The eye tracker's default parameters were used to convert gaze positions into fixations and saccades.

3.3. Procedure

Participants started with a practice phase of 6 uninterrupted emails, during which the baseline pupil dilation was calculated. After the sixth email was archived, the first block started immediately.

The experiment consisted of three parts, each of the parts containing three blocks in random order: one Control block, one Random block and one IMS block. Each block finished after 10 emails were archived and the participant could then take a break. The experiment finished after all 9 blocks were completed. The experiment lasted approximately 50 minutes.

3.4. Participants

Twenty-six (19 female) participants were tested. Four participants were removed because they had at least two blocks where the IMS hardly interrupted them. The reason for the scarce interruptions is that it took a long time before the IMS managed to find the optimal WIV for these participants. A possible explanation for that is that the original threshold value (0.997) was too high for these participants and the IMS kept interrupting them in the beginning of the block. Although the IMS lowered the threshold, the fact that they had to deal with so many interruptions and typing of the answers probably made the participants' pupils dilate, resulting in a threshold that was too low for them. If the blocks had been longer, the IMS might have managed to find the optimal threshold. We decided to remove these participants because their uninterrupted IMS blocks are not representative of how the IMS works.

The remaining 22 participants (15 female) had a mean age of 23.2. All participants had normal or corrected-to-normal vision, gave informed consent for their participation and received monetary compensation of 8 euros.

3.5. Preprocessing

Two blocks were removed because participants did not follow instructions, one from the control condition and one from the IMS condition. Furthermore, 62 emails were removed because each of them had 4 or more interruptions. Only six of these belonged in the random condition, the rest were mostly the first emails of the IMS condition, when the IMS needed some time to find the appropriate WIV value.

In order to construct Figure 7—not for the IMS—eye blinks were removed from the pupil dilation data, starting 100 ms before the blink up to 100 ms after the blink. The removed pupil dilation data were replaced by a linear interpolation and then all data were down-sampled to 100 Hz. We calculated the percentage change in pupil dilation from a baseline, which was defined by a very slow lowess filter

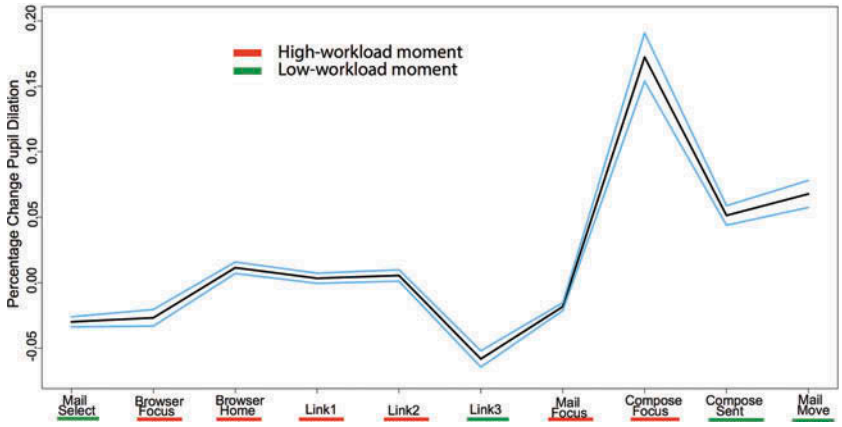

Figure 4. Average PCPS values for every move for uninterrupted emails (Control condition). The gray lines show the standard error.

[2]Parameters used: smoother span = 2/3, number of robustifying iterations that should be performed = 3, delta = 0.01 * the range of time.

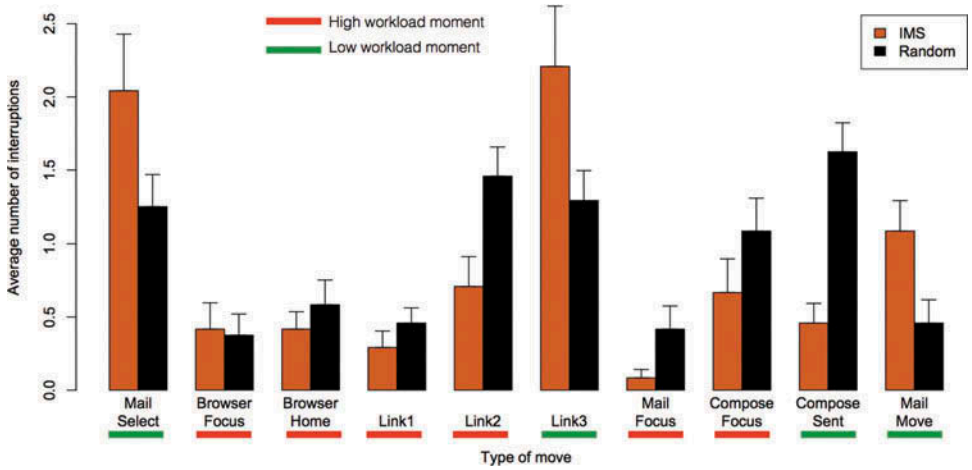

Figure 5. Average number of interruptions that occurred in each step of the main task for the IMS and the Random conditions. The high workload moments are indicated with a dark gray line and the low workload moments with a light gray line.

(a smooth curve that follows pupil dilation, with the parameter values as provided by R, 2008) given by a weighted linear least squares regression over the span (following Katidioti et al., 2014).

4. RESULTS

4.1. Pupil Dilation and Interruption Moments

Figure 4 shows the average percentage change in pupil dilation (PCPS) values at each email sequence moment for the Control condition, i.e., emails that contained no interruptions. It is clear that pupil size increases on high-workload moments and then decreases on low-workload moments. The IMS identifies the low-workload moments by the comparing the percentage change in pupil dilation of the last 200 ms with the threshold. Overall we observe the highest PCPS during the "Compose Focus" move, probably because participants had to type at that point. That is the reason why—although there is a large drop in the PCPS in the next step ("Compose Sent")—the value is still higher than that of other high-workload moments.

Link 3 is an interesting point in the task. It is a low-workload moment but only lasts about 2 s and it is in the middle of the task, between high-workload moments. Nevertheless, the size of the pupil decreases at this point (Figure 4) and the IMS is able to detect that decrease (see Figure 5).

Figure 5 shows the number of interruptions for each moment in the email sequence. The IMS succeeded partly in shifting the interruptions to the low-workload moments (green moments in Figure 5). Compared to the Random condition, using the IMS resulted in more low-workload moment interruptions and fewer high-workload moment interruptions. On average, there were 10.27 interruptions per block in the Random condition (4.7 during high-workload moments and 5.58 during low-workload moments) and 6.26 interruptions per block in the IMS condition (2.15 during high-workload moments and 4.1 during low-workload moments). Proportionally this means that in the IMS condition 66.8% of the switches occurred at low-workload moments and 33.2% at high-workload moments, whereas in the Random condition the percentages were 54.0% and 46.0%, respectively (Figure 6).

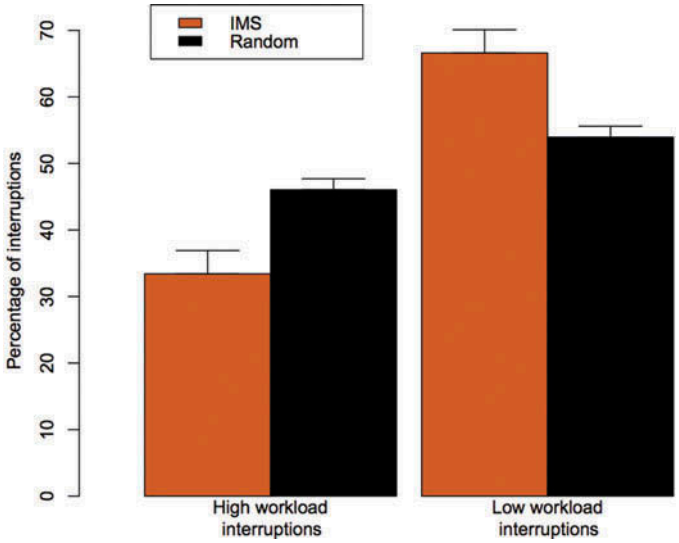

Figure 6. Average number of interruptions on high and low workload moments per block for both conditions.

According to a two-way repeated measures ANOVA with Condition (IMS vs Random) and Workload (low vs high) as factors, there were significantly more low-workload than high-workload interruptions ($F(1,21) = 21.15$, $p < .001$, $\eta_p^2 = 0.5$) and significantly more interruptions in the Random condition than in the IMS condition ($F(1,21) = 23.89$, $p < .001$, $\eta_p^2 = 0.53$). Most importantly, the interaction between Condition and Workload was significant ($F(1,21) = 5.29$, $p = .032$, $\eta_p^2 = 0.2$), indicating that the IMS changed the proportion of interruptions to more interruptions at low-workload than at high-workload moments. Furthermore, block analysis showed that the IMS increased the percentage of low-workload interruptions as the blocks progressed, from 58.04% in the first IMS block to 67.38% in the last IMS block.

Taking into account the 1-s delay in pupil dilation reaction (e.g., Hoeks & Levelt, 1993), we checked what type of moment occurred 1.1–0.9 s before the high-workload moment

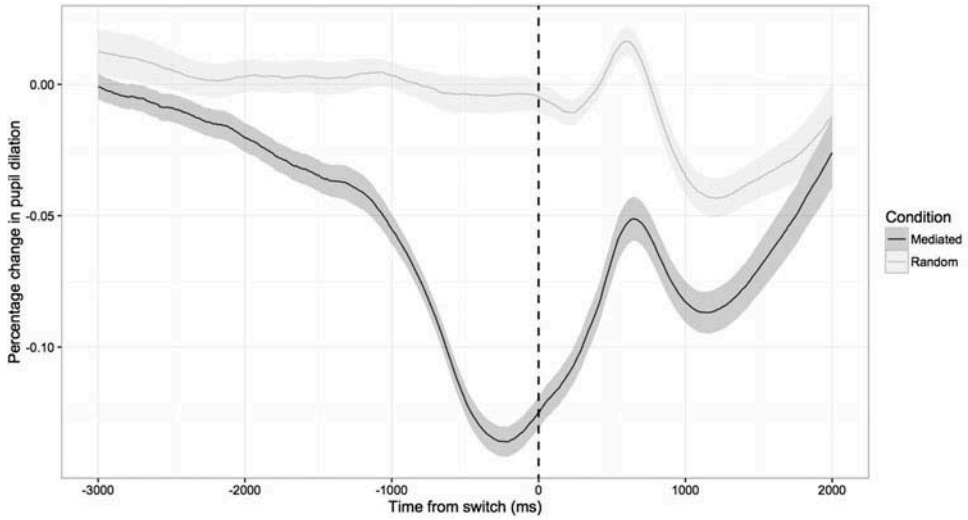

Figure 7. Average pupil dilation around the interruption point (indicated by the dashed line) for the IMS and the Random conditions. The lighter area around each line represents the standard error.

interruptions that the IMS created. Results showed that at 62.2% of the time that was indeed a high-workload moment and 37.8% of the time it was a low-workload moment.

To assess whether the IMS interrupted users when their pupil dilation was low, as we intended, we compared the average pupil dilation for the Random and the IMS condition around the interruption point, time-locked at the moment of interruption (time = 0 s, Figure 7). This figure shows that the IMS indeed interrupted users when their pupil dilation was low. In both conditions there is an increase in pupil dilation approximately 700 ms after the interruption. This increase likely reflects the pupil's reaction to the interruption itself.

4.2. Performance

In order to verify that high-workload moment interruptions are worse than low-workload ones, we compared emails in which participants were interrupted at a high-workload moment to emails in which they were interrupted at a low-workload moment (independent of condition). When interrupted at a high-workload moment, participants were significantly slower in completing the email sequence (22.37 s) than when they were interrupted at a low-workload moment (20.21 s; $t(21) = -3.76$, $p = .0012$).

The success of the IMS in detecting low-workload moments was reflected in the participants' performance on the main email task. The average time to complete an email (after removing the time spent on interruptions and emails that deviated more than 3 SDs from the mean) per condition was 18.16 (SE = 0.73) s for the Control condition, 20.30 (SE = 0.71) s for the IMS condition and 21.53 (SE = 1.19) s for the Random condition. An ANOVA revealed a significant difference between conditions ($F(2,42) = 19.18$, $p < .001$, $\eta_p^2 = 0.48$), and a pairwise t-test with Bonferroni-Holm correction revealed that the difference between the IMS and the Random condition was marginally significant ($t(21) = -2.05$, $p = .053$). Participants were significantly faster in the Control condition then in the IMS or the Random condition (both ps < .001).

Participants seldom made mistakes (such as typing the wrong price or looking up the wrong product). However, there were times that participants forgot the product name while browsing and revisited the email window in order to read it again. There were on average 0.32 (SE = 0.05) revisits per email in the Control condition, 0.44 (SE = 0.08) in the IMS condition and 0.46 (SE = 0.07) in the Random condition. An ANOVA showed that this difference is significant ($F(2,42) = 4.567$, $p = .016$, $\eta_p^2 = 0.18$), and a follow-up pairwise t-test showed that the only significant difference was between the Control and the Random condition ($t(21) = -2.82$, after a Bonferroni-Holm correction $p = .031$).

Another performance measure used often in interruption studies is the resumption lag: the time one needs to resume the main task after being interrupted. In this setup, resumption lag is the time between sending the answer in the chat and the next main task move. The resumption lag was 1.55 s for the IMS condition and 1.37 s for the Random condition, a difference that was not significant according to a t-test ($t(21) = 1.37$, $p = 0.18$, $d = 0.37$).

5. Discussion

The aim of this study was to create and test an IMS that interrupts users at optimal moments of an ongoing task. It has been shown by many studies (e.g., Gould et al., 2013; Iqbal et al., 2005; Iqbal & Bailey, 2005; Katidioti & Taatgen, 2014; Salvucci & Bogunovich, 2010) that interruptions at low-workload moments are less disruptive than interruptions at high-workload moments. In order to measure workload and find low-workload moments, we used pupil dilation, a well-known measure of cognitive workload (e.g., Beatty, 1982; Beatty & Lucero-Wagoner, 2000; Hoeks & Levelt, 1993; Iqbal et al., 2005). We developed a task-independent IMS that interrupts users when their pupil dilation drops below an adaptive value (Figure 1). We then tested

this IMS in an experimental study, using an email-and-chat setup that resembles client service in an electronics company (Salvucci & Bogunovich, 2010).

PCPS increased on the known high-workload moments of this task and decreased on the low-workload moments (Figure 4), confirming that pupil dilation reacted to the workload changes. Behavioral results showed that the IMS succeeded in interrupting participants on the low-workload moments of the main task (Figure 5). Even Link 3, which is a low-workload point lasting only about 2 s was detected by the IMS, as seen in Figure 6. In the Random condition, switches were almost equally distributed between low-workload (53.95%) and high-workload (46.05%) moments. In the IMS condition, the IMS managed to increase the difference, by interrupting participants 66.81% of the time on low-workload moments. In addition, pupil dilation results confirmed that these interruptions happened when pupil dilation was low or decreasing in the IMS condition (Figure 7). Taking all of the above into account, we can conclude that the IMS was successful in interrupting people at low-workload moments by detecting a decrease in their pupil dilation.

Performance in this setup was measured by the average time needed to complete an email sequence. High-workload interruptions (across both conditions) made participants significantly slower than low-workload interruptions. This result is in line with previous studies on the timing of interruptions (Gould et al., 2013; Iqbal et al., 2005; Iqbal & Bailey, 2005; Katidioti & Taatgen, 2014; Salvucci & Bogunovich, 2010), which all suggest that low-workload interruptions are less disruptive than high-workload interruptions.

Although there was no difference between being interrupted by the IMS or randomly in some performance measures (resumption lag and number of revisits to the email window), participants were marginally faster to complete an email in the IMS than in the Random condition. This is promising, especially given that the IMS did not result in 100% low-workload switches, but only 67%. There are two possible explanations for that. The first explanation is that since pupil dilation reacts to stimuli with a 1-s delay (Hoeks & Levelt, 1993), the behavioral task we used alternated too quickly from low- to high-workload moments for the IMS to perform optimally—and consequently, 100% low-workload switches were impossible. By the time the IMS decided that the pupil dilation was low enough to interrupt, the low-workload moment of the task might have already changed to a high-workload one—for example, opening an email is a low-workload moment, but reading it lasts only a couple of seconds and then working memory is occupied again. Analysis of the task moments that occurred 1.1–0.9 s before high-workload interruptions during the IMS blocks revealed that they were mostly also high-workload moments (62.2%). Thus, the remaining 37% of the high-workload switches might be due to a low-workload moment about a second earlier. The second explanation is that the IMS needs some time to find the optimal WIV. In some cases, that led to participants being constantly interrupted in the beginning of the IMS blocks, or not interrupted at all. The fact that the percentage of low-workload interruptions increased from 58.04% in the first IMS block to 67.38% in the last IMS block supports this explanation. In a real-life environment this should be less of a problem if pupil dilation can be measured continuously and the WIV updated throughout the day.

A substantial advantage of our IMS is that it is task-independent. Some parameters (i.e., the 10–30-s interval between interruptions that was used in order to have a fair comparison with the Random condition, the 200 ms pupil dilation judging period and the rate/amount of the threshold adaptor changing) were chosen to fit this specific task, such that we could make a valid comparison between the IMS and random interruptions. However with few minor changes (e.g., having only a set number of interruptions per hour), the IMS can be adapted to different tasks. The IMS is task-independent because it interrupts people only based on the changes in their pupil dilation, not on the properties of the main task. It can therefore be applied without first having to perform a task analysis, which is required for most other systems (e.g., Arroyo & Selker, 2011; Iqbal & Bailey, 2010). Furthermore, the pupil dilation during the interruption is not taken into account, which means that the IMS is not affected by the extent to which the interrupting task is relevant to the main task. Since our IMS is task-independent, it is possible to integrate it into an operating system and use it across tasks. For instance, simple office work could benefit from such an IMS, which could defer email and social media notifications until a low-workload moment. Naturally, it could also be employed in single-task environments such as the cockpit or air-traffic control. In those cases, the current system might be combined with a task analysis to identify crucial processes that may never be interrupted.

One issue with a pupil-dilation-based IMS is that one needs to measure pupil dilation in real time in an uncontrolled environment, where lighting conditions might affect the pupil. Although pupil dilation has traditionally been measured with expensive eye-tracking systems, in recent years webcams have rapidly become more capable, to the extent that most current webcams are high definition. Such high-quality webcams can be used to measure pupil dilation changes and calculate workload in normal office conditions (Rafiqi, Wangwiwattana, Fernandez, Nair, & Larson, 2015), promising to make eye tracking and pupil dilation widely available.

The IMS used a specific algorithm that compares 200 ms worth of pupil dilation to an adaptive threshold in order to judge whether the pupil dilation indicated a low-workload moment. We could change the algorithm to fit the task better, by comparing the pupil dilation of each step of the task to the pupil dilation of the previous step. This would be a better idea for this task, since the changes from a high to a low-workload moment can be very quick. Although this change would probably yield better results with this specific task, it would also make the IMS task-dependent. Another idea is to take the direction of the pupil change into account. Only if the pupil size keeps decreasing for a specific amount of time, then the workload decreases and it is a good moment for an interruption. Although this seems a good idea on the basis of the average data, its robustness still has to be investigated online.

IMSs are becoming a necessity in the world we live in that is full with interruptions that endanger our work. There are many studies that point out what makes interruptions disruptive and how to minimize their negative effects (e.g., Edward & Gronlund, 1998; Gould et al., 2013; Hodgetts &

Jones, 2006; Iqbal & Bailey, 2007; Jin & Dabbish, 2009) but few have implemented this knowledge in a usable system for managing interruptions. In the study presented here we focused only on one aspect of interruptions (timing of the interruption) and one psycho-physiological measure (pupil dilation), avoiding the need for extensive task analysis that is required in many previous systems. We showed that a pupil-dilation-based IMS can identify low-workload moments in real time, and interrupt users at opportune moments, leading to marginally better performance than random interruptions. Although our IMS can be further optimized—for instance by taking into account other sources of data (Züger & Fritz, 2015)—it already showed promising results.

Acknowledgments

We thank Dario Salvucci for providing the code for the experiment.

Funding

This research was funded by ERC-StG grant 283597 awarded to Niels Taatgen.

References

Altmann, E. M., & Trafton, J. G. (2002). Memory for goals: An activation-based model. *Cognitive Science*, *26*, 39–83.

Arroyo, E., & Selker, T. (2011). Attention and intention goals can mediate disruption in human–computer interaction. *Interact*, *1* (6947), 454–470.

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin*, *91*(2), 276–292.

Beatty, J., & Lucero-Wagoner, B. (2000). The pupillary system. In J. T. Cacioppo, L. G. Tassinary, & G. Berntson (Eds.), *Handbook of psychophysiology* (pp. 142–162). Cambridge, MA: Cambridge University Press.

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*. Paper presented at the SIGCHI Conference on Human Factors in Computing Systems (CHI '15), Seoul, South Korea.

Cades, D. M., Boehm Davis, D. A., Trafton, J. G., & Monk, C. A. (2007). *Does the difficulty of an interruption affect our ability to resume?* Paper presented at the Human Factors and Ergonomics Society Annual Meeting 2007, Baltimore, MD, October 1–5.

Czerwinski, M., Horvitz, E., & Wilhite, S. A. (2004). *A diary study of task switching and interruptions*. Paper presented at the SIGCHI Conference on Human Factors in Computing Systems (CHI '04), Vienna, Austria.

Edward, M. B., & Gronlund, S. D. (1998). Task interruption and its effects on memory. *Memory*, *6*(6), 665–687.

Eyrolle, H., & Cellier, J. M. (2000). The effects of interruptions in work activity: Field and laboratory results. *Applied Ergonomics*, *31*(5), 537–543.

Gillie, T., & Broadbent, D. E. (1989). What makes interruptions disruptive? A study of length, similarity, and complexity. *Psychological Research*, *50*, 243–250.

Gonzalez, V., & Mark, G. (2004). *"Constant, constant, multi-tasking craziness": Managing multiple working spheres*. Paper presented at the SIGCHI Conference on Human Factors in Computing Systems (CHI '04), Vienna, Austria.

Gould, S. J. J., Brumby, D. P., & Cox, A. L. (2013). *What does it mean for an interruption to be relevant? An investigation of relevance as a memory effect*. Paper presented at the Human Factors and Ergonomics Society Annual Meeting 2013, San Diego, CA, September 30–October 4.

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

Hodgetts, H. M., & Jones, D. M. (2006). Interruption of the tower of London task: Support for a goal activation approach. *Journal of Experimental Psychology: General*, *135*(1), 103–115.

Hoeks, B., & Levelt, W. J. M. (1993). Pupillary dilation as a measure of attention: A quantitative system analysis. *Behavior Research Methods, Instruments, & Computers*, *25*, 16–26.

Iqbal, S. T., Adamczyk, P. D., Zheng, X., & Bailey, B. P. (2005). *Towards an index of opportunity: Understanding changes in mental workload during task execution*. Paper presented at the SIGCHI Conference on Human Factors in Computing Systems (CHI '02), Minneapolis, MN.

Iqbal, S. T., & Bailey, B. P. (2005). *Investigating the effectiveness of mental workload as a predictor of opportune moments for interruption*. Paper presented at the SIGCHI Conference on Human Factors in Computing Systems (CHI '05), Portland, OR.

Iqbal, S. T., & Bailey, B. P. (2006). *Leveraging characteristics of task structure to predict the cost of interruption*. Paper presented at the SIGCHI Conference on Human Factors in Computing Systems (CHI '06), Montreal, Quebec.

Iqbal, S. T., & Bailey, B. P. (2007). *Understanding and developing models for detecting and differentiating breakpoints during interactive tasks*. Paper presented at the ACM Conference on Human Factors in Computing Systems, San Jose, CA.

Iqbal, S. T., & Bailey, B. P. (2008). *Effects of intelligent notification management on users and their tasks*. Paper presented at the SIGCHI Conference on Human Factors in Computing Systems (CHI '08), Florence, Italy.

Iqbal, S. T., & Bailey, B. P. (2010). Oasis: A framework for linking notification delivery to the perceptual structure of goal-directed tasks. *ACM Transactions on Computer–Human Interaction*, *17*(4), 1–28.

Jennings, J. R., & Van der Molen, M. W. (2005). Preparation for speeded action as a psychophysiological concept. *Psychological Bulletin*, *131*(3), 434–459.

Jin, J., & Dabbish, L. A. (2009). Self-interruption on the computer: A typology of discretionary task interleaving. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)* (pp. 1799–1808). New York: ACM. doi:10.1145/1518701.1518979

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

Kahneman, D., Tursk, B., Shapiro, D., & Crider, A. (1969). Pupillary, heart rate and skin resistance changes during a mental task. *Journal of Experimental Psychology*, *79*(1), 164–167. doi:10.1037/h0026952

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*(6), 380–396.

Katidioti, I., & Taatgen, N. A. (2014). Choice in multitasking: How delays in the primary task turn a rational into an irrational multitasker. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *56*(4), 728–736.

Kobayashi, Y., Tanaka, T., Aoki, K., & Fujita, K. (2015). Automatic delivery timing control of incoming email based on user interruptibility. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '15)* (pp. 1779–1784). New York, NY: ACM.

Laeng, B., Ørbo, M., Holmlund, T., & Miozzo, M. (2011). Pupillary Stroop effects. *Cognitive Processing*, *12*, 13–21.

Laeng, B., Sirois, S., & Gredebäck, G. (2012). Pupillometry: A window to the preconscious? *Perspectives on Psychological Science*, *7*(1), 18–27.

McFarlane, D. C. (2002). Comparison of four primary methods for coordinating the interruption of people in human–computer interaction. *Human–Computer Interaction*, *17*(1), 63–139.

Monk, C. A., Boehm Davis, D. A., & Trafton, J. G. (2004). Recovering from interruptions: Implications for driver distraction research. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*(4), 650–663.

Moresi, S., Adam, J. J., Rijcken, J., Van Gerven, P. W. M., Kuipers, H., & Jolles, J. (2008). Pupil dilation in response preparation. *International Journal of Psychophysiology*, *67*, 124–130.

Peavler, W. S. (1974). Pupil size, information overload and performance differences. *Psychophysiology*, *11*, 559–566. doi:10.1111/j.1469-8986.1974.tb01114.x

Prehn, K., Heekeren, H. R., & Van der Meer, E. (2011). Influence of affective significance on different levels of processing using pupil dilation in an analogical reasoning task. *International Journal of Psychophysiology*, *79*(2), 236–243.

R Development Core Team. (2008). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0. http://www.R-project.org

Rafiqi, S., Wangwiwattana, C., Fernandez, E., Nair, S., & Larson, E. C. (2015). *Work-in-progress, pupilware-M: Cognitive load estimation using unmodified smartphone cameras*. Paper presented at MASS 2015, SocialSens 2015, Dallas, TX, October 19–22.

Salvucci, D. D., & Bogunovich, P. (2010). *Multitasking and monotasking: The effects of mental workload on deferred task interruptions*. Paper presented at the SIGCHI Conference on Human Factors in Computing Systems (CHI '10), Atlanta, GA.

Tanaka, T., Abe, R., Aoki, K., & Fujita, K. (2015). Interruptibility estimation based on head motion and PC operation. *International Journal of Human–Computer Interaction*, *31*(3), 167–179.

van Rijn, H., Dalenberg, J. R., Borst, J. P., & Sprenger, S. A. (2012). Pupil dilation co-varies with memory strength of individual traces in a delayed response paired-associate task. *PLoS One*, *7*, e51134. doi:10.1371/journal.pone.0051134

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

About the Authors

Ioanna Katidioti earned her PhD from the Department of Artificial Intelligence of the University of Groningen in 2016.

Jelmer P. Borst is a senior postdoctoral researcher at the Department of Artificial Intelligence of the University of Groningen. He investigates how computational cognitive models can be used to inform the analysis of neural data. He earned his PhD in artificial intelligence from the University of Groningen in 2012.

Douwe J. Bierens de Haan is a master student of Human Machine Communication in the Department of Artificial Intelligence of the University of Groningen in 2016.

Tamara Pepping is a master student of Human Machine Communication in the Department of Artificial Intelligence of the University of Groningen in 2016.

Marieke K. van Vugt is an assistant professor of cognitive modeling at the Department of Artificial Intelligence of the University of Groningen. She investigates how decision making and distraction can be modeled and are implemented by the brain. She earned her PhD in neuroscience from the University of Pennsylvania in 2008.

Niels A. Taatgen is full professor of cognitive modeling at the Department of Artificial Intelligence of the University of Groningen. He earned his PhD in psychology from the University of Groningen in 1999.