An Interruption Management System based on Pupil Dilation tested in Air Traffic Control

Bachelor's Project Thesis

Cedric Fekken, s2400529, c.e.fekken@student.rug.nl

Supervisor: Jelmer Borst

Abstract: In daily life we often make errors due to being interrupted while performing our tasks. Often these interruptions happen when we are in the middle of a high workload moment. These errors could possibly be reduced if we could manage at which moment we would be interrupted in such a way that we will only be interrupted during low mental workload moments. In our study we created an Interruption Management System based on pupil dilations that tries to interrupt participants during low workload moments. We tested our system by having participants perform an air traffic control task during which participants would be interrupted. The performance of the IMS will be compared to random interruptions. Results show that the IMS could not distinguish the low mental workload moments from the high mental workload moments in this air traffic control task.

1. Introduction

Interruptions are becoming more and more prevalent in our current society. Easier contact through the internet and our phones makes it a lot easier to get interrupted. González and Mark (2004) found that information workers on average spent three minutes on an event, which could be anything from a phone conversation to writing down information, and the time between switching main tasks is around 12 minutes. These interruptions would often not be initiated by the workers themselves as about half of these interruptions were actually caused by external events. These external interruptions were often in unexpected moments and were experienced as being a lot more disruptive. When interruptions are disrupting the main task generally the performance of the workers is decreased (e.g. Arroyo and Selker, 2011), there is a delay in the resumption of their main task (e.g. Altmann and Trafton, 2007; Hodgetts and Jones, 2006; Monk et al., 2008) or they make more errors (e.g. Brumby et al., 2013; Bailey and Konstan, 2006).

Some interruptions are more disruptive than others, several studies have been done looking for what features of these interruptions affected this disruptiveness. The length of the interruptions seems to matter as longer interruptions tend to be more disruptive than shorter ones (Hodgetts and Jones, 2006; Monk et al., 2008). The complexity of the interruption also seems to matter as interruptions that are more complex or prevent rehearsal also tend to be more disruptive (Hodgetts and Jones, 2006; Cades et al., 2007). The topic of the interruption also has an effect as interruptions that are more similar to the main task seem to be less disruptive than interruptions that are less similar (Gould et al., 2013; Arroyo and Selker, 2011).

Another very important influence on the disruptiveness of interruptions is when these interruptions occur. When interruptions occur during high workload moments rather than low workload moments can result in a higher resumption lag or more errors in the main task (Bailey and Konstan, 2006; Monk et al., 2008; Kreifeldt and McCarthy). Participants reported an increase in their annoyance and anxiety when they were interrupted during high workload moments compared to during low workload moments (Bailey and Konstan, 2006) and tend to make worse decisions in their main tasks just after an interruption during a high workload moment (Speier et al., 1999).

A way to measure the mental workload in people is to use pupil dilation. The pupil dilation gets increases slightly when the mental workload increases (Beatty and Lucero-Wagoner, 2000). These measurements have been found to be very robust and the effects are similar across tasks and individuals. Iqbal and colleagues (2005) found that measuring the percentage change in pupil dilation was an effective way to measure the mental workload of participants and suggested that these measurements can be used in a system that can make more effective decisions about when to interrupt users. Iqbal and Bailey (2005) continued their studies on this showing that a system could be made to interrupt participants based on the measured mental workload. They found that interruptions done at the predicted low workload moments caused less resumption lag and annoyance and fostered more social attribution.

Züger and Fritz (2015) created a system with the help of machine learning techniques that could assess the interruptibility of knowledge workers. The system used several psycho-physiological features to determine this interruptibility and proved to be highly accurate. No interruptions were actually performed at these opportune moments but interviews were conducted that showed that the desirability of a system that can manage interruptions is high.

Katidioti and colleagues (2016) have used pupil dilation to indicate the mental workload of participants to manage the timing of interruptions. An eye tracker was used to measure the changes in pupil dilation in real time and these measurements were used in an Interruption Management System (IMS) to try to interrupt participants during low workload moments. To test this IMS an email task was created. In this task participants had to answer emails which contained questions about products for which the answer had to be looked up. High workload moments during this task were moments during which participants had to remember product information. Low workload moments during this task were moments during which the working memory of the participants was not occupied. The interruptions were in the form of a chat window in which the participants were asked simple questions that they had to answer. Figure 1 shows that the IMS had success in this task and it managed to

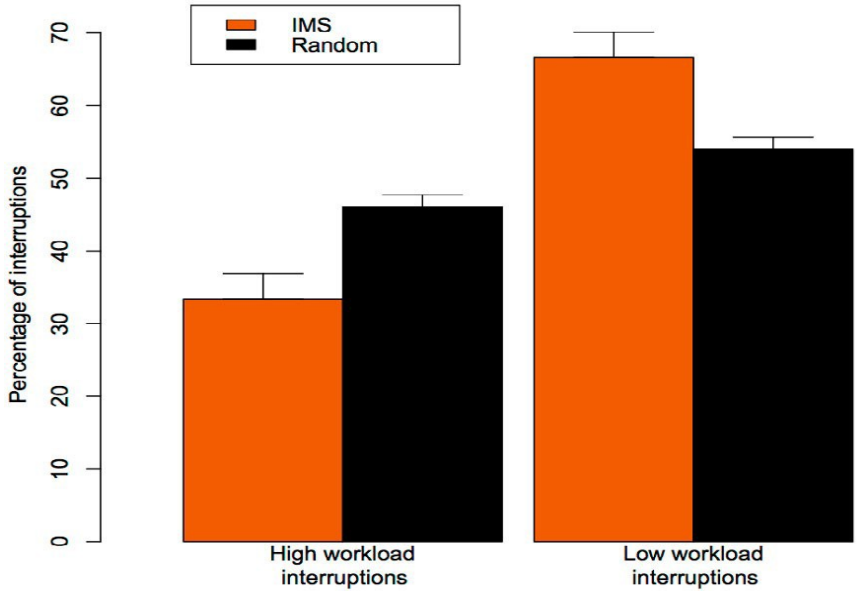

Figure 1: A graph from Katidioti et al. (2016) showing the distribution of interruptions amongst the conditions.

interrupt participants at the low workload moments. The participants completed the tasks faster when the IMS was managing the interruptions than when the interruptions were at random moments.

In this study the IMS of Katidioti and colleagues (2016) will be tested to see if it generalizes to a more continuous and complex task, namely Air Traffic Control (ATC). In this task participants will have to control and land planes on a runway. During this task participants will be interrupted with simple mathematical problems.

2. Method

To see if the interruption management system works, a study was conducted to find out if the IMS can interrupt participants during low workload moments and if this improves their performance on an air traffic control (ATC) task. During this study, participants first had six practice trials to familiarize themselves with the controls and the task. After these practice trials there were 15 blocks of trials, five blocks for every condition. The different conditions were the IMS blocks, the randomized blocks and the control block. Each of these blocks consisted of two super trials, which then contained one low workload trial and one high workload trial. All these blocks are randomized in order for every participant. To see how effective the IMS is, we looked at if the IMS properly interrupted participants at low workload moments instead of high- workload moments. The performance of the participants during the IMS trials will also be compared to their performance during the randomized and control trials.

2.1. Interruption Management System

The algorithm of the IMS in this study is based on the algorithm described in the paper of Katidioti et al. (2016). In their task the IMS was first used to interrupt participants during low workload moments in an email task. The IMS identifies these low workload moments

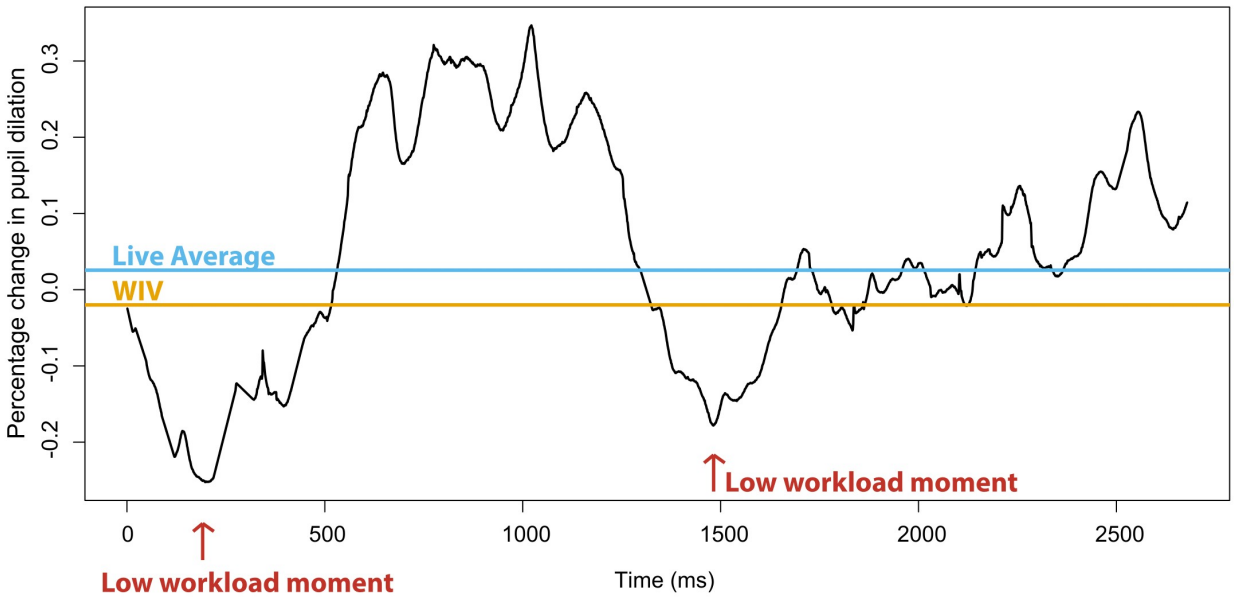

Figure 2: A graph from Katidioti et al. (2016) showing the PCPS and two low workload moments.

by measuring the pupil dilation of the participant.

The IMS works by taking the baseline pupil dilation during the practice trials of the experiment. The baseline pupil dilation is transformed in real-time to a percentage change in pupil size (PCPS). This is done by subtracting the baseline from the measured pupil size and then dividing it by the baseline. After this transformation 1000 is added to this value to avoid having to multiply with negative numbers. Values slightly below 1000 would indicate a pupil size smaller than the previously recorded pupil dilation and thus a lower workload and vice versa.

The PCPS values created in this transformation are then stored and averaged to create the Live Average, the Live Average holds the PCPS values of the last minute and averages them. Figure 2 shows how these values worked in Katidioti et al. (2016). The IMS uses this Live Average to compare if the workload of the participant is higher or lower than it was over the previous minute. To estimate the exact workload of the participant the Live Average is multiplied by the Threshold Adapter to create a Workload Index Value (WIV). The Threshold Adapter is initiated at a value of 0.997 which was found to have the best results in the study done by Katidioti et al.(2016). This value represents a way for the IMS to learn the optimal threshold per participant as it can change depending on the amount of interruptions that took place in every IMS super trial. For every IMS trial a target number of interruptions was set to two, the Threshold Adapter would then be increased or decreased depending on how much the the actual amount of interruptions in the trials deviated from the target number. This change was done with a scale of 0.001 to make sure the changes would happen gradually and not overshoot the ideal WIV for the participant.

The IMS determines if a participant's workload is low enough by comparing the current PCPS to the WIV. When the PCPS is below the WIV for more than 200 ms, the IMS identifies the current situation as a low workload moment, telling the game that an interruption could take place. Whenever a participant was not looking at the screen or blinked the camera would not record their pupil size and therefore not generating a PCPS, in this way they are ignored by the system. To avoid a situation in which an interruption could possibly have a lower workload than the main task and thus possibly trigger a loop of interruptions, the pupil dilation was not measured during interruptions and the 5 seconds after an interruption, effectively putting a minimum timer between interruptions.

2.2. Experiment

The experiment consisted of a game simulating an ATC task. The ATC tasks require the participant to keep track of multiple planes and generally required a relatively high mental workload. In the game participants have to land as many planes as possible while making sure none of them get lost or crash. During the game the participants will be interrupted by simple mathematical problems which should be easy and quick to solve for every participant.

The game is a modified version of the open source game Towerx[1]. The game is modified to allow the IMS to handle interruptions. Playing the game requires no background knowledge and the game is heavily modifiable in how the trials are set up, thus allowing the experiment to have expected low and high workload moments.

The planes are represented by the small green squares on the screen. The airfield where the planes have to land have two landing strips, a horizontal one and a vertical one. The currently selected plane is the green square with the circle around it. The selected plane can be changed by hovering the mouse over a different plane, the green circle will then swap to the plane the participant is currently hovering over.

Figure 3 shows a screenshot of what participants will see during the main task. Every plane will have two lines of information next to it, the top line is its name and the bottom line is its altitude. Planes will maintain the same altitude unless ordered to descend or ascend. When planes collide and are within 200 feet of altitude of each other they will crash and become uncontrollable grey crosses on the screen. Similarly when a plane flies off the screen it will become lost and count as crashed and if the plane goes below 1000 feet altitude the plane will crash on the ground. To be able to land on the landing strips the altitude of the plane has to be between 1101 and 1499 and have to approach the landing strips from the sides of the flashing lights, this is the bottom for the vertical landing strip and the right side for the horizontal landing strip. When a plane reaches the landing strip at the

1 http://pygame.org/project-Towex+ATC+Game-1650.html

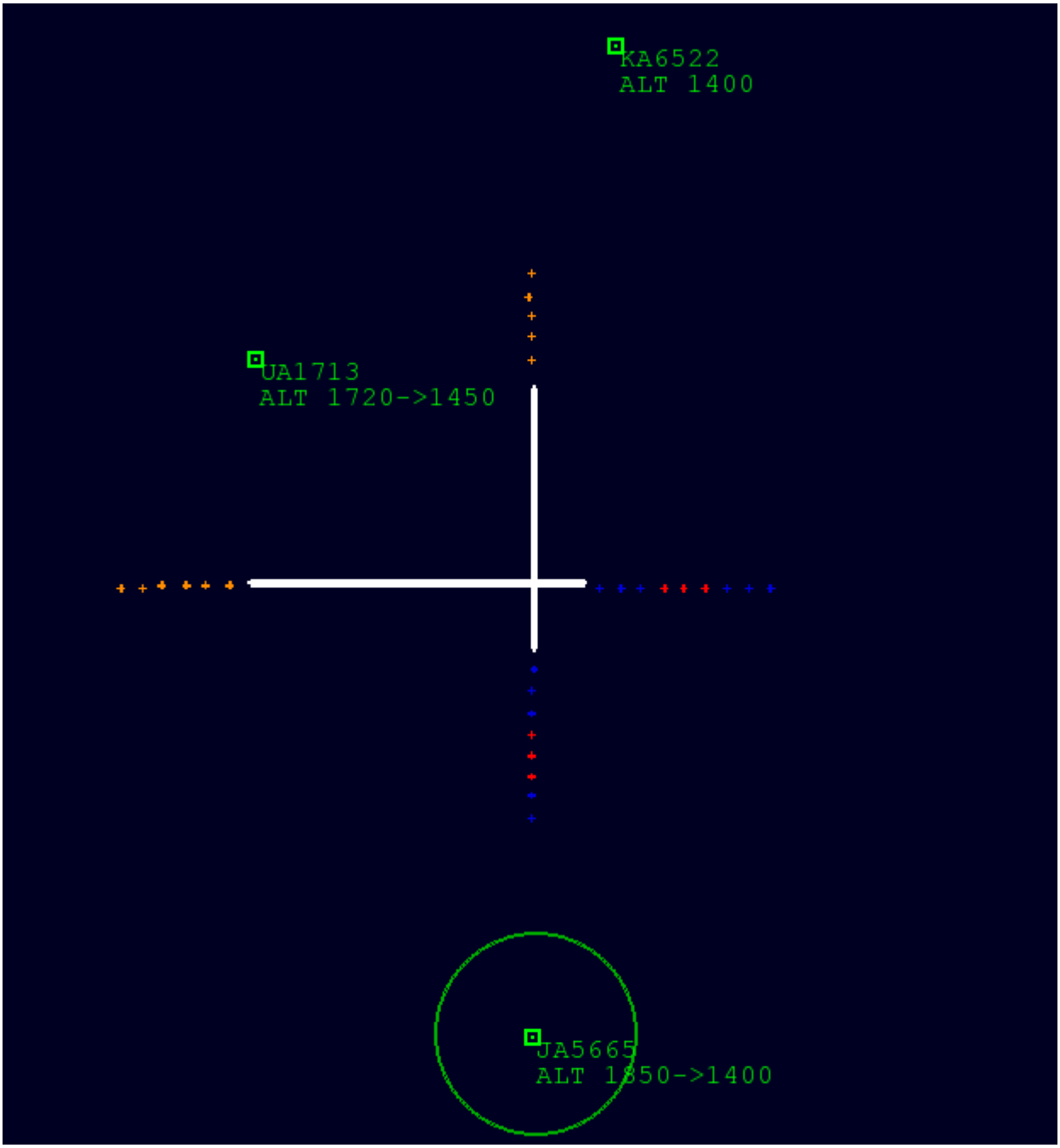

Figure 3: A screen shot of the Air Traffic Control game. This is what participants will see during the main task. The green squares are planes and the one with the circle around it is the currently selected plane. The lines in the middle are the runways and need to be approached from the red and blue sides.

correct location at the correct altitude the plane will land and the participant will no longer have to control that plane. To control the direction of the plane, press 'W' to send it north, 'A' for west, 'S' to send it south and 'D' to send it to the east. Planes will keep flying in the same direction at the same speed and the same key does not have to be pressed again to keep it moving in the same direction. To change the altitude the 'Page Up' button can be pressed to increase the altitude by 50 and the 'Page Down' button to decrease it by 50.

The interruptions consisted of an easy mathematical equation having either an addition or a subtraction with one

Figure 4: An example of an interruption. Participants had to solve the given problem by clicking on the right numbers and then submit their answer by clicking on 'Enter'.

number between 10 and 90 and the other between 1 and 10. Figure 4 shows one of these interruptions. Participants have to click the numbers of the answer and then click on the enter button. It is advised for participants to solve these tasks quickly as the planes will keep flying as normal during the interruption.

The performance of the participants on the task is measured by a score that they can see for a short while after solving an interruption task. The scores are determined by the following factors: landing a plane gives 100 points, crashing or losing a plane gives -150 points, successfully solving an interruption task gives 100 and incorrectly solving an interruption gives -400 points.

2.3. Conditions

The experiment has three conditions: IMS, randomized and control. The control condition featured no interruptions. In the randomized condition participants would be randomly interrupted between 10 and 40 seconds into the trial. This would be the same every randomized trial. For the IMS condition, the trial would start the same except that the IMS would pick when a good moment would be for an interruption. Since there is no maximum time between interruptions in the IMS condition it is possible that there is not a single interruption during a trial if the workload is high and the IMS cannot find a suitable moment to interrupt.

Every super trial consisted of two trials, one easy trial and one hard trial. The easy trial started with 3 planes and the hard trial with 6 planes. Every trial would take 40 seconds by default and participants are not meant to have enough time to always land all the planes. This is done to increase the workload as the participants have to land planes quickly for them land in time. The time of these trials is increased by the time spent by the participant on interruptions.

2.4. Participants

The study was performed with 17 participants, of which 4 were male. The ages of the participants range from 20 to 33 and the average age is 25. Of the experiments, the data of 4 participants was unfit for analysis due to the camera being unable to pick up the correct pupil dilation.

2.5. Apparatus

The experiment was performed in a small room that has no windows. The researcher and the participant were not within sight of each other during the actual experiment. The participant was seated on a chair in front of a desk and had to use a chin rest. The camera was an Eyelink 1000 from SR Research which was placed between the monitor and the keyboard. Eye fixations were measured with a sample rate of 250 Hz. Before the practice trials a calibration was done and after every break a drift correction was done.

2.6. Design and Procedure

The experiment took approximately 90 minutes per participant. The participants were doing the experiment one at a time. Before the practice trials and the calibration, participants were asked to read the instructions for the ATC, which was provided for them on paper. After reading the instructions the researcher would explain anything that might have been unclear about how to play the game.

After the instructions a calibration was performed. Then the participants played six practice trials. These practice trials were split into three groups, first two trials which last until all planes have left the field through landing, crashing or getting lost. Then there were two trials that lasted 40 seconds and were similar to the control trials. The last two practice trials had set interruptions happening during the trial to make sure the participant knows how to solve them. During these six trials the baseline pupil dilation was calculated. The participants get to ask any final questions about how to play the game and then the first block of the actual experiment begins. There were breaks after the fifth and tenth block.

3. Results

To analyse the results, we looked at the performance of the Interruption Management System and the disruptiveness of the interruptions.

The performance of the IMS was tested by comparing the performance of participants in the IMS and Randomized trials. A precondition for this comparison is that the number of interruptions in the IMS and Randomized trials are close to equal. This precondition is made to ensure that the comparison between the IMS and Randomized trials is fair. A higher amount of interruptions leads to worse performance so if the amount of interruptions are not equal the comparison will be in favour of the condition with the least interruptions. Figure 5 shows the average interruptions per trial per block. It shows that on average the IMS trials had more interruptions per trial than the Randomized trials. This means that the precondition for this performance comparison was not met and therefore nothing can be said about the performance of the IMS.

Another way the performance of the IMS was tested was by comparing the interruptions in low- and high workload trials. This will be done by taking the

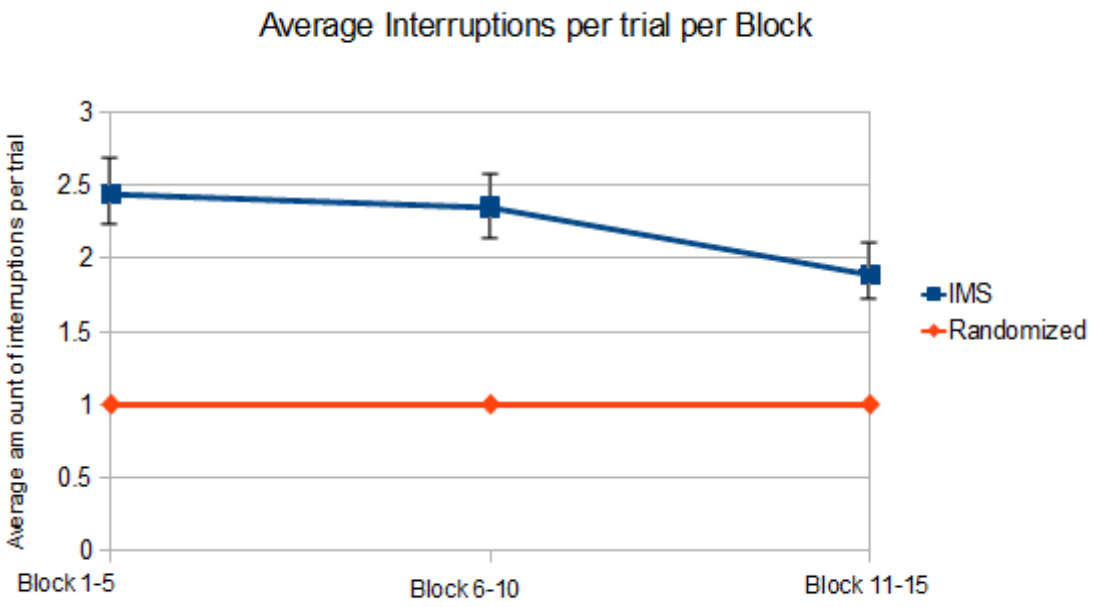

Figure 5: Average number of interruptions per trial per block for the IMS and Randomized conditions.

total number of interruptions during the IMS trials. In total during there were 312 interruptions during low workload trials and 309 during high workload trials. This is no relevant difference and therefore we can say that the IMS did not manage to interrupt participants more often during the predicted low workload moments.

Next the disruptiveness of the interruptions during the task was measured. This was done by comparing the performance of the participants in the three conditions. The performance was measured using a score and the score was averaged per condition per participant and then all the participants' averages were averaged. Figure 6 shows the performance per condition with standard error bars. It shows that on average, participants had a better performance during the Control condition than in the IMS or Randomized conditions and participants performed better in the Randomized condition than in the IMS condition. There is no overlap in the standard error, this suggests that there is a significant performance difference between the conditions. To confirm this, a one-way repeated measures ANOVA was done which showed that the way of interrupting (Control, Randomized or IMS condition) had a significant effect on the performance of participants during the main task ($F(2,39)=8.597$, $p=0.0008067$).

4. Discussion

In this study the main goal was to investigate whether the IMS of Katidioti et al. (2016) can be generalized to other contexts, in this case an Air Traffic Control task. The IMS of Katiodioti and colleagues (2016) uses pupil dilation to measure the mental workload. In their study the IMS worked and participants had an improved performance during an email task with the IMS compared to randomized interruptions.

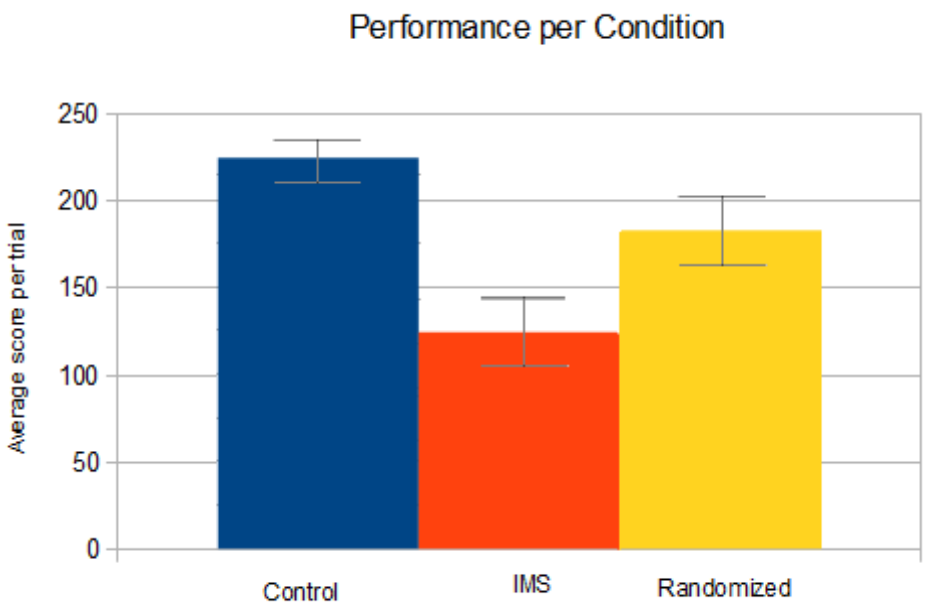

Figure 6: The average performance of participants per trial per condition with standard error bars.

To measure the performance of the IMS in this study the performance of participants in the IMS and Randomized trials was compared. The precondition for this comparison is that the number of interruptions during the IMS trials and the Randomized trials is close to equal. This was unfortunately not the case as the amount of interruptions during the IMS trials was a lot higher.

The IMS uses a Threshold Adapter to change the threshold depending on how many interruptions there were during IMS trials. As seen in Figure 5 the IMS recognizes that there are too many interruptions and the Threshold Adapter changes to reduce the amount of interruptions. To reduce the chance of overshooting the ideal Threshold the changes made to the Threshold Adapter are relatively small, therefore it was not able to reduce the amount of interruptions by the IMS by enough to match the amount of interruptions in the Randomized condition. A way to meet the precondition would be to increase the amount of interruptions in the Randomized trial to 2, this would mean the amount of interruptions in both the IMS and Randomized condition would be a lot closer and thus a fair comparison could be made.

Another way to test the performance of the IMS was by comparing the amount of interruptions in low- and high workload trials. The amount of interruptions during the IMS condition in low- and high workload trials was unfortunately almost equal whereas it would be expected that the IMS would interrupt participants more often during the low workload trials. To further analyse why the IMS might have interrupted participants a lot during high workload trials we looked at how many planes were remaining during the interruptions. When a participant lands or loses multiple planes early on in a high workload trial the workload for the trial would be reduced, therefore triggering more interruptions which would still be counted as interruptions

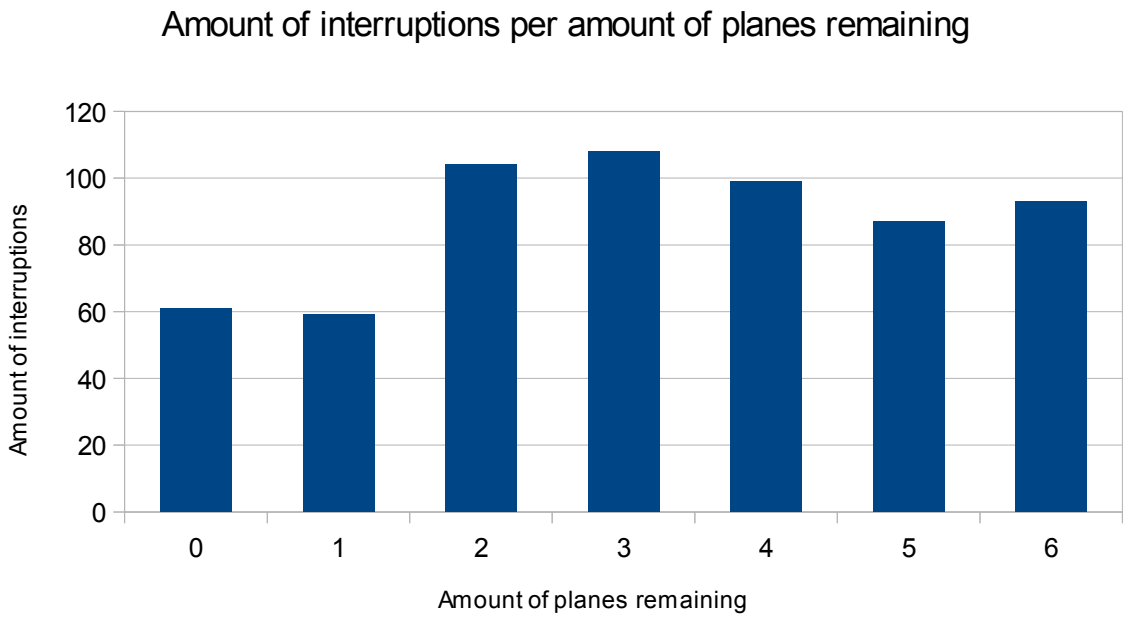

Figure 7: The amount of interruptions with the amount of planes still remaining when the interruption started.

during high workload trials. Figure 7 shows the amount of interruptions with the amount of planes still remaining when the interruption started. There are more interruptions when there are 0-3 planes on the field than when there are 4-6 planes on the screen, where interruptions when there are 0-3 planes would be low workload moments and 4-6 planes remaining on screen can only happen in high workload trials. The total amount of interruptions with 0-3 planes remaining is 332 and the interruptions with 4-6 planes remaining is 279. This does show that in total there are more interruptions during low workload moments than during high workload moments, however the amounts are still a lot closer to each other than expected, the IMS interrupts way too often during high workload trials. The amount of interruptions is the highest when there are 2-4 planes remaining. An explanation for this could be that participants will most likely spend the most time with 2-4 planes on the screen.

A possible reason that the IMS interrupted more often than expected in high workload trials has to do with the risk taking behaviour of participants. During the high workload trials participants have to put more effort into making sure that none of the planes crash or get lost. This leads to participants having less time and incentive to actually try to land the planes. The mental workload is expected to be a lot lower when participants aren't trying to land the planes. During low workload trials participants are more likely to take the risk of putting in extra effort into trying to land the planes as there are only 3 planes on the screen. This might lead to the IMS recognizing more low workload moments during the high workload trials than expected.

Another possible reason for the high amount of interruptions could be that during the start of a trial the participant has to analyse the new locations of all the planes and making sure they aren't close to crashing or getting lost. This could lead to a spike in mental workload at the start of the trial which will then be followed by a period of lower mental workload as none of the planes are in immediate danger of crashing or getting lost. This sequence of events would lead to a low workload moment even during a high workload trial, which might lead to more interruptions during the high workload trials. This effect would be less prevalent during low workload trials as there aren't as many planes at the start of a trial.

The disruptiveness of interruptions was measured by comparing the performance of the participants in every condition. A one-way repeated measures ANOVA was done to see if there was a significant difference between the conditions. The ANOVA did find a significant difference in the scores which was expected as there was no overlap in the standard error bars. The participants did perform better during the Control condition than either the IMS or Randomized conditions which shows that when there were no interruptions participants perform better. We hoped to find that participants performed better during the IMS condition than during the Randomized condition however as Figure 5 shows there were a lot more interruptions during the IMS condition. This indicates that the amount of interruptions did have an impact on the performance of the participants and thus were disruptive enough.

For future research the amount of Randomized interruptions could be increased to be able to have a fair comparison between the IMS and the Randomized conditions. More research can be done looking into exactly why the amount of interruptions in the low- and high workload trials seem to be close to equal as there seems to be no clear reason in the data.

References

Altmann, E. M., & Trafton, J. G. (2007). Timecourse of recovery from task interruption: Data and a model. *Psychonomic Bulletin & Review*, *14*(6), 1079-1084.

Arroyo, E., & Selker, T. (2011, September). Attention and intention goals can mediate disruption in human-computer interaction. In *IFIP Conference on Human-Computer Interaction* (pp. 454-470). Springer Berlin Heidelberg.

Bailey, B. P., & Konstan, J. A. (2006). On the need for attention-aware systems: Measuring effects of interruption on task performance, error rate, and affective state. *Computers in human behavior*, *22*(4), 685-708.

Beatty, J., & Lucero-Wagoner, B. (2000). The pupillary system. *Handbook of psychophysiology*, *2*, 142-162.

Brumby, D. P., Cox, A. L., Back, J., & Gould, S. J. (2013). Recovering from an interruption: Investigating speed–accuracy trade-offs in task resumption behavior. *Journal of Experimental Psychology: Applied*, *19*(2), 95.

Cades, D. M., Davis, D. A. B., Trafton, J. G., & Monk, C. A. (2007, October). Does the Difficulty of an Interruption Affect our Ability to Resume?. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 51, No. 4, pp. 234-238). SAGE Publications.

González, V. M., & Mark, G. (2004, April). Constant, constant, multi-tasking craziness: managing multiple working spheres. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 113-120). ACM.

Gould, S. J., Brumby, D. P., & Cox, A. L. (2013, September). What does it mean for an interruption to be relevant? An investigation of relevance as a memory effect. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 57, No. 1, pp. 149-153). SAGE Publications.

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.

Iqbal, S. T., Adamczyk, P. D., Zheng, X. S., & Bailey, B. P. (2005, April). Towards an index of opportunity: understanding changes in mental workload during task execution. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 311-320). ACM.

Iqbal, S. T., & Bailey, B. P. (2005, April). Investigating the effectiveness of mental workload as a predictor of opportune moments for interruption. In *CHI'05 extended abstracts on Human factors in computing systems* (pp. 1489-1492). ACM.

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*, (just-accepted).

Kreifeldt, J. G., & McCarthy, M. E. (1981). Interruption as a test of the user-computer interface.

Monk, C. A., Boehm-Davis, D. A., Mason, G., & 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.

Monk, C. A., Trafton, J. G., & Boehm-Davis, D. A. (2008). The effect of interruption duration and demand on resuming uspended goals. *Journal of Experimental Psychology: Applied*, *14*(4), 299.

Speier, C., Valacich, J. S., & Vessey, I. (1999). The influence of task interruption on individual decision making: An information overload perspective. *Decision Sciences*, *30*(2), 337-360.

Züger, M., & Fritz, T. (2015, April). Interruptibility of software developers and its prediction using psycho-physiological sensors. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 2981-2990). ACM.