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

Marlies Hoekstra, s2397595, m.hoekstra.27@student.rug.nl,
Jelmer Borst [*] and Ioanna Katidioti [*]

March 2, 2016

Abstract

Interruptions become more and more predominant in daily life and lead to more errors and slower task completion. One way to minimize the disruptiveness of interruptions is to manage at which moment in the task they occur. Interruptions are less disruptive when they take place on low mental workload moments than on high mental workload moments. In the current study, we created a system for managing the timing of interruptions in order to make them less disruptive. This interruption management system (IMS) is based on pupil dilation, a well-known indicator for mental workload. The IMS was tested with a simple task that simulates air traffic control and was compared to random interruptions. However, after testing it with 6 participants, we discovered that there is probably a flaw in the software that prevented us from drawing any conclusions about the performance of the IMS.

1 Introduction

Nowadays, interruptions are inevitable. They become more and more integrated into our daily lives and working environments. González and Mark (2004) found that the average time that information workers spent on an event (e.g. phone conversation, annotating documents, talking to colleagues) is about three minutes, and the time between major context switches is about 12 minutes. About half of the time those switches are self-initiated, but the other half of the interruptions are caused by external events. Especially external interruptions are often experienced as disruptive, as they occur in unexpected, and often bad moments. Those interruptions can lead to decreased performance (e.g. Arroyo and Selker, 2011), a delay in the resumption of the task (larger resumption lag) (e.g. Altmann and Trafton, 2007; Hodgetts and Jones, 2006; Monk et al., 2008) or more errors (e.g. Brumby et al., 2013; Bailey and Konstan, 2006).

Various features of an interruption can influence the extent of its disruptiveness. Studies have shown that longer interruptions are more disruptive than shorter ones (Hodgetts and Jones, 2006; Monk et al., 2008) and that complex interruptions or interruptions that prevent rehearsal are more disruptive than simpler ones (Hodgetts and Jones, 2006; Cades et al., 2007). It is also known that interruptions more relevant or similar to the task at hand are less disruptive than irrelevant/dissimilar interruptions (Gould et al., 2013; Arroyo and Selker, 2011).

Finally, mental workload has a great influence on the disruptiveness of interruptions. Interruptions that occur during high workload moments rather than during low workload moments can result in a higher resumption lag or more errors in the primary task (Bailey and Konstan, 2006; Monk et al., 2004; Kreifeldt and McCarthy, 1981). Bailey and Konstan (2006) reported an increase in annoyance and anxiety in high workload moment interruptions rather than in low workload moment interruptions, and Speier et al. (1999) found that worse decision making occurred after interruptions occurring during high workload moments.

[*] University of Groningen, Department of Artificial Intelligence

A popular measure for mental workload is pupil dilation. Small increases in pupil dilation occur when the mental workload gets higher (Beatty and Lucero-Wagoner, 2000). Those measurements are very robust and the effects are similar across tasks and individuals. Katidioti et al. (submitted) used pupil dilation as an indicator for workload to manage the timing of interruptions. They measured the changes in pupil dilation in real-time using an eye tracker and integrated it into an Interruption Management System (IMS) that interrupted at low workload moments. The IMS was tested in a simulation of a client service task in which the participants needed to answer emails with questions about products, after looking up the answers. During the task the participants were interrupted by a chat window with simple questions they had to answer. The IMS worked: the interruptions took place at the right (low workload) moments, and the emails were completed faster when the interruption moments were determined by the IMS than when they took place at random moments.

Züger and Fritz (2015) used machine learning techniques to create a system that could assess the interruptibility of knowledge workers. The system was based on several psycho-physiological features and proved to be highly accurate. They did not yet use the system to actually interrupt the users at opportune moments, but they did conduct interviews that showed that the desirability of a IMS's is high.

McFarlane (2002) tested four different interruption strategies: immediate (on random moments), negotiated (the user gets to choose the moment), mediated (the system chooses the moment based on estimated workload of the task) and scheduled (every 25 seconds) interruptions. A number of design goals were identified and an evaluation of the best interruption strategy for each design goal was made. For example, in order to accomplish the highest accuracy on the continuous task, the negotiated strategy can best be chosen. For most design goals, negotiated or mediated interruptions give the best result.

This study investigates whether the IMS of Katidioti et al. (submitted) generalizes to a more continuous task, namely Air Traffic Control (ATC). The task we use is a simplified version of ATC where the participant needs to control planes and land them on the runways. The interruptions consist of simple math problems.

2 Method

To test the performance of the IMS, we performed a small pilot study, in which we wanted to look at the difference in performance between IMS-managed interruptions and random interruptions, during a game. The participants started with six practice trials in order to get familiar with the task. The practice trials were followed by three blocks with different conditions (Control, Randomized, IMS) that each consisted of four super trials, which contained four trials each. To see if the interruptions were disruptive for the performance on the main task, we compared the game performance in trials with interruptions, to the performance in trials with no interruptions.

2.1 Interruption Management System

The IMS used in this study is based on the one of Katidioti et al. (submitted), in which the effect of the IMS is studied in an email-task with interruptions. The IMS interrupts the user at low workload moments identified by pupil dilation measures.

During the practice trials preceding the experiment, a *baseline* pupil dilation was recorded. During the experiment the dilation was recorded and transformed in real-time to the percentage change in pupil size (*PCPS*) by subtracting the baseline from the measured pupil size and then dividing it by the baseline. Figure 1 shows the PCPS (black line) over the timespan of 30 seconds. The PCPS was then increased by 1000 to avoid multiplication with negative numbers.

The PCPS was recorded and averaged over the past minute to obtain the *Live Average*. This Live Average was used to make up for learning curves in the game. The Live Average was then multiplied by a *Threshold Adapter* to represent

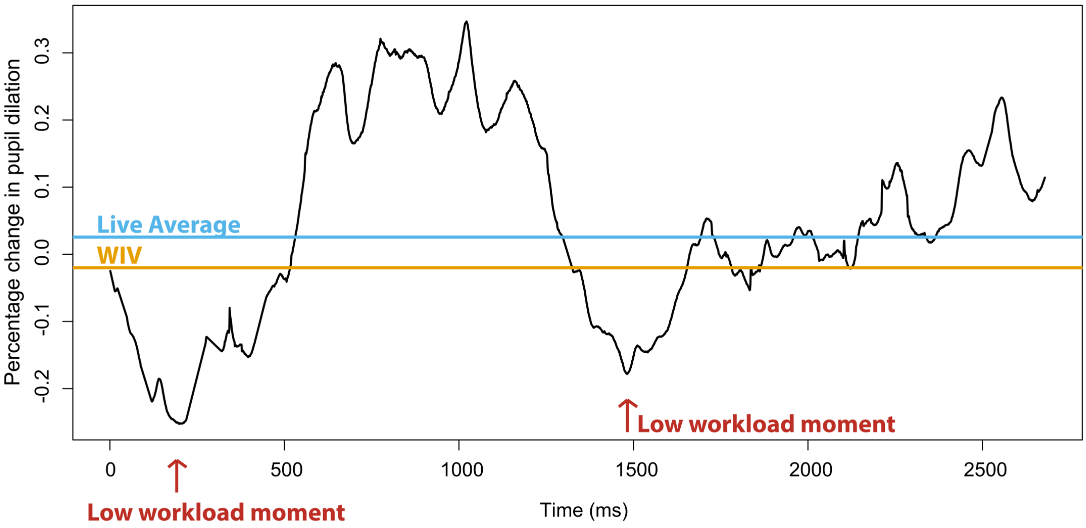


Figure 1: Graph of the PCPS over time from Katidioti et al.(2016). The two low workload moments that are indicated occur when the PCPS is below the WIV for more than 200 ms.

the Workload Identifier Value (WIV). The Live Average and WIV can both be seen in Figure 1. The Threshold Adapter started at a value of 0.997 (the same as in Katidioti et al. (submitted)) and was changed after every super trial in order to find the optimal WIV for each participant. The change was proportional (scaled by a factor of -0.001) to the deviation of the target number of interruptions (2 per trial).

The PCPS was compared to the WIV, when the PCPS was below the WIV for more than 200 ms (red arrows in Figure 1), this was regarded a low workload moment and the system could make an interruption.

The interrupting task can have workload characteristics that differ from those of the main task. Therefore, the pupil dilation was not measured during an interruption and until 5 seconds after an interruption, so that the dilation could stabilize. Also, eye blinks and off-screen gazes are ignored.

2.2 Experiment

The main task consisted of a game that simulates Air Traffic Control (ATC). ATC jobs require high mental workload capabilities and in those jobs, interruptions can have high costs. The goal of the game is to land as many planes as possible, and not to lose or crash any planes. The interruptions are simple math problems which have to be solved as accurate and fast as possible.

The game that we used is a simplified version of ATC and it is playable for people without a specific background. Also the workload and the length of the workload moments are adaptable, which were important factors for the execution of the experiment. The game is adapted from the open source game Towerx[1]. Our modified version is optimized for the purpose and setup of the experiment.

Figure 2 shows a screen shot of the game. The planes are represented by the green squares, with their names and altitudes. When the user

[1] http://pygame.org/project-Towerx+ATC+Game-1650-.html

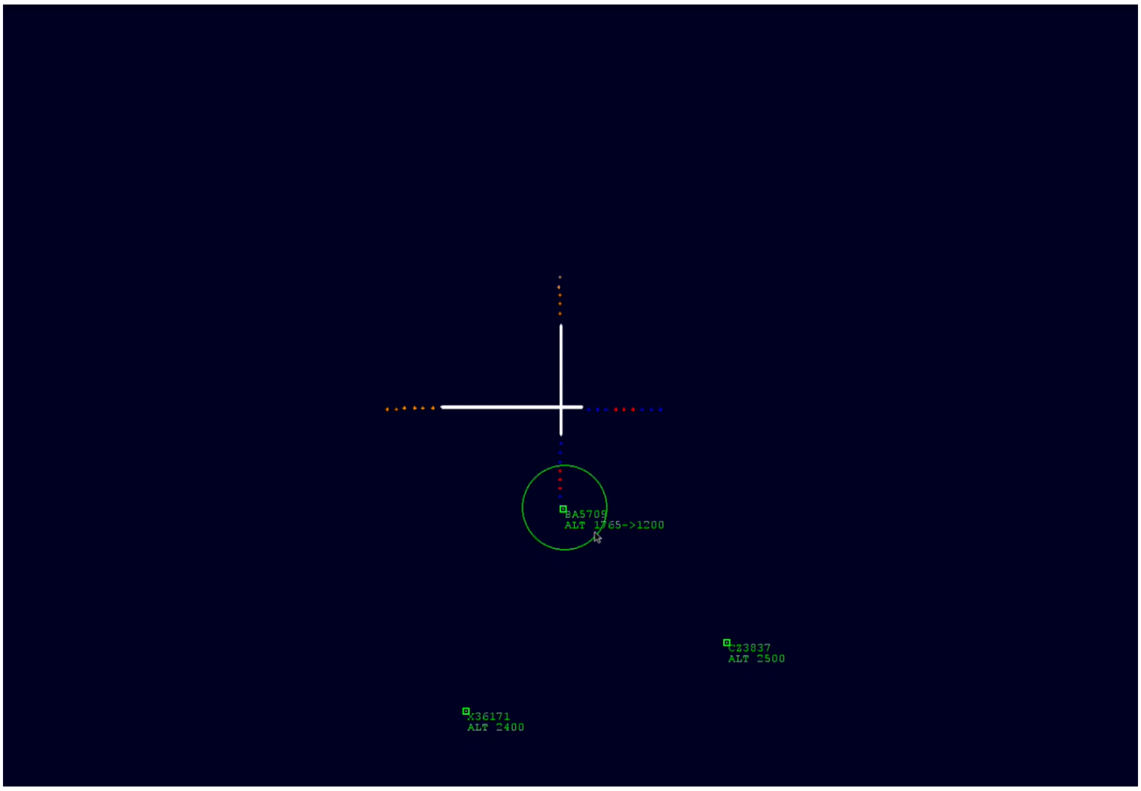
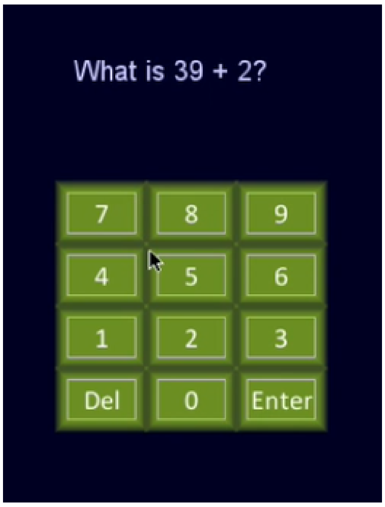


Figure 2: Screen shot of the ATC game, the green squares represent the planes. The one with the green circle is selected, and is changing altitude. The runways are in the middle of the screen and need to be approached from the right or from the bottom

places the mouse on a plane, the plane is selected and a green circle surrounds it. The runways are in the middle, and need to be approached from the right or from the bottom. In order to land, a plane must have an altitude of 1300 ft. Planes crash into each other when they have the same altitude and position, and crash into the ground when they reach an altitude below 1000 ft. Planes can also get lost, by flying out of the visual range. The score is increased by 100 points when a plane is landed, and decreased by 150 points when a plane crashes or gets lost.

The controls of the game were chosen carefully in order to minimize the occasions where the eye-tracker loses the pupil by the participant looking at the keyboard. The heading and altitude of a plane can be altered after a plane is selected. Participants were asked to change the heading with their left hand by pressing the "WASD"-keys ('W' for up, 'A' for left, 'S' for down and 'D' for right), and select the planes (by placing the mouse on top of a plane) and change the altitude (page-up to ascend and page-down to descend) with their right hand.

The interrupting task consisted of simple math problems. The problems were additions and subtractions of numbers between 10 and 90, and one and 10. An example of an interruption can be seen in Figure 3. During the interruptions, the planes continue flying. However, the participants could only see the interrupting task, so that the planes could not be seen or controlled during an interruption. The score penalty for answering a math problem incorrectly is high: 500 points. This is in order to make sure that participants do not answer the problem randomly so that they can go back to controlling the planes as quickly as possible. The reward for a correct answer is 100



What is 39 + 2?

Figure 3: Screen shot of an interruption. Participants had to solve the given problem and enter the answer with the mouse on the on-screen keypad, before submitting by clicking on the 'Enter'-button

points. After the answer is entered the score is displayed for 1.5 seconds before the participant can go back to controlling the planes, making the interruption even more disruptive. The answers to the math problems needed to be entered with the mouse on an on-screen keypad, like the one in Figure 3.

2.3 Conditions

The three conditions of the experiment were: Control, Randomized and IMS. In the Control condition the participants were never interrupted. The trials lasted about 40 seconds, and the goal was to have a mean of 2 interruptions per trial, in the Randomized and IMS condition. In the Randomized condition, a random time between the next 10 and 30 seconds was chosen, at which the interruption occurred. After the interruption, the same procedure took place. In the IMS condition, the procedure was the same but now the IMS could pick an interruption moment during the next 10 to 30 seconds according to the algorithm explained in section 2.1. When the IMS detected no suitable moment, there was no interruption. After a super trial, the Threshold Adapter was modified in order to reach the mean of 2 interruptions per trial.

The trials could have one of two different difficulty levels. The easy trials started with 3 planes, and were meant to create a low workload environment, the hard trials started with 6 planes and were meant to create a high workload environment.

2.4 Participants

The pilot study was performed with six participants (1 female). The ages of the participants range from 21 to 34 and the average age is 25.

2.5 Apparatus

The experiment was conducted in a small windowless room. The participants were seated in front of a desk and were asked to use a chin rest. An LCD monitor of 1600 x 1200 pixels with a density of 64 pixels/inch was used. The eye tracker was an Eyelink 1000 from SR Research which was placed at approximately 45 cm from the end of the desk. A calibration was performed before the experiment started and a drift correction took place in between the blocks.

2.6 Design and Procedure

The experiment took about 50 to 60 minutes per participant. Each participant was tested individually. Before the experiment started, participants were asked to read the instructions for the game[2], provided on paper.

First, participants played six practice trials. The first two practice trials lasted until all the planes were either landed or lost/crashed, so that the participant gets some time to get to know the controls and the game. The last four trials lasted 40 seconds (as in the real experiment) of which the last two included an interruption at a fixed moment, so that the procedure of the interruption could also be practiced. During those six trials the baseline pupil dilation was calculated.

The experiment itself consisted of three blocks,

[2] The instructions can be found in the appendix

each with a different condition: Control, Randomized or IMS. The order of the blocks was counterbalanced so that every participant played the blocks in a different order. In between the blocks the participants were allowed to take a short break, after which a drift correction was performed. Every block consisted of four super trials after which the Threshold Adapter was modified. The super trials consisted of two easy and two hard trials (presented in a random order) lasting 40 seconds, plus interruption time.

3 Results

To investigate the results, we looked at the number of landed planes as a measure of performance.

To know whether the IMS worked, we wanted to compare the task performance in the IMS condition to the task performance in the Randomized condition. One precondition for this comparison though, is that the number of interruptions over those two conditions is about the same. However, in the IMS condition we found an average of 3 interruptions per trial, where the Randomized condition had an average of only 1 interruption per trial. Unfortunately this means that we cannot say anything about the performance of the IMS.

Next we investigated whether the interruptions of this task were disruptive for the main task performance. To do this, we compared the performance in the Control condition (without interruptions) to the performance in the other conditions (with interruptions).

In order to analyze the data, first the sum of landed planes over every super trial per participant per condition was taken and then the number of landed planes per participant per condition was averaged.

Figure 4 shows the standard error bars of the different conditions. It shows that on average more planes were landed in the control condition than in the other two conditions. However, there is a lot of overlap in the standard error. This suggests that the difference in performance between the conditions is not significant. A one-way repeated measures ANOVA also indicated that there was no significant effect of the manner of interrupting (Control, Randomized or IMS condition) on the number of landed planes (F(2,10)=0.781, p=0.484).

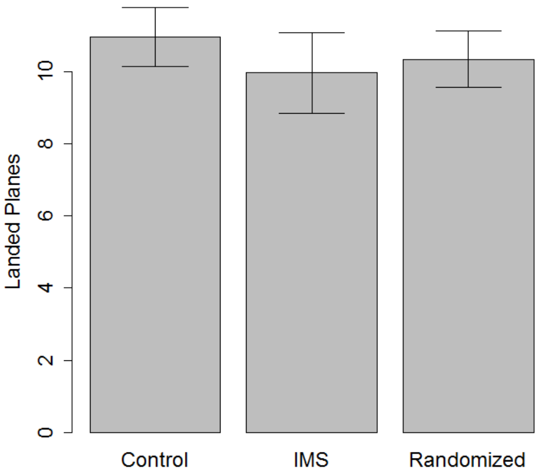


Figure 4: Error bars of the number of landed planes in a super trial per condition.

4 Discussion

The main goal of this study was to investigate whether the IMS of Katidioti et al. (submitted) can be generalized to another context, with a more continuous task. Katidioti et al. (submitted) created an IMS that used pupil dilation to measure workload. They found that their IMS worked and that it led to improved performance on an email task compared to random interruption moments. In this study we tested their IMS in a simulation of Air Traffic Control.

To determine the performance of the IMS, the number of landed planes in the Randomized condition needed to be compared to the number of landed planes in the IMS condition. A precondition for doing this, is that the number of interruptions in both conditions needs to be the same. Unfortunately there was a big difference in the number of interruptions in both conditions, which made the comparison impossible. The IMS algorithm makes up for those differences in the number of interruptions by altering the Threshold Adapter after every super trial. However, due to a yet unknown flaw in the software this did not work the way it should. When the experiment would be repeated, the first thing to do would be debugging the software to fix this flaw.

The disruptiveness of the interruptions in this study was measured by the difference in performance (number of landed planes) between the Control condition and the Randomized and IMS conditions. An ANOVA did not detect a significant difference between the conditions. Figure 5 includes the performance per participant and shows that the variance across the participants is very high. Furthermore, there were only six participants, which gives us a very small data set. Those two factors could provide an explanation for the results, as they make it harder to detect a potentially significant result.

When the study would be repeated with more participants, it is possible that the interruptions would show a significant disruptiveness. However, it is also possible that the interruptions would turn out not to be disruptive enough. The disruptiveness of the interruptions could be increased in a number of ways. One way would be to increase the duration of the score display after a problem is solved. This would prolong the interruption which would be effective for two reasons. The participant would be distracted for a longer amount of time and the planes would keep flying uncontrolled for a longer amount of time, which makes it harder to get them back on track. Another way of increasing the disruptiveness of the interruptions would be to make the math problems harder. This would have the same effects as mentioned above, as on average the participant would need to think longer before the answer could be entered. In addition to this, the working memory of the participant would be more occupied with solving the math problem, which would make the interruption even more intrusive.

A last option for future research would be to investigate whether the system could find the optimal moments for interruptions, at which the workload was low. This could be done by looking at the pupil dilation data at the moments of the interruptions.

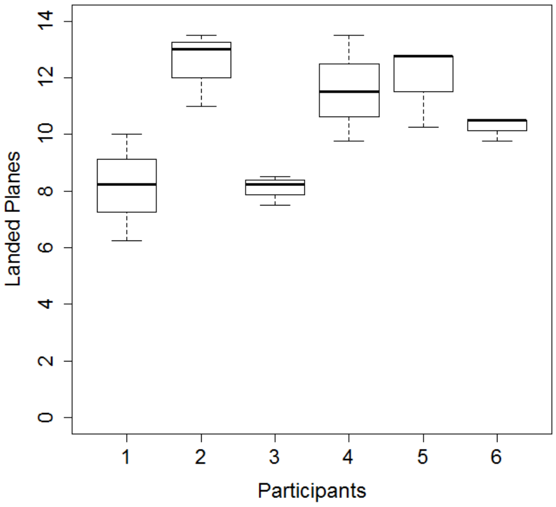


Figure 5: Boxplots for the number of landed planes in a super trial per participant.

References

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

Arroyo, E. and Selker, T. (2011). Attention and intention goals can mediate disruption in human-computer interaction. In *Human-Computer Interaction–INTERACT 2011*, pages 454–470. Springer.

Bailey, B. P. and 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. and Lucero-Wagoner, B. (2000). The pupillary system. *Handbook of psychophysiology*, 2:142–162.

Brumby, D. P., Cox, A. L., Back, J., and 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., and Monk, C. A. (2007). Does the difficulty of an interruption affect our ability to resume? In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 51, pages 234–238. SAGE Publications.

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

Gould, S. J., Brumby, D. P., and Cox, A. L. (2013). 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*, volume 57, pages 149–153. SAGE Publications.

Hodgetts, H. M. and 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.

Katidioti, I., Borst, J., Bierens de Haan, D., Pepping, T., and van Vugt, M.Kaatgen, N. Interrupted by your pupil: An interruption management system based on pupil dilation. submitted.

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

McFarlane, D. (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., Mason, G., and 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., and Boehm-Davis, D. A. (2008). The effect of interruption duration and demand on resuming suspended goals. *Journal of Experimental Psychology: Applied*, 14(4):299.

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

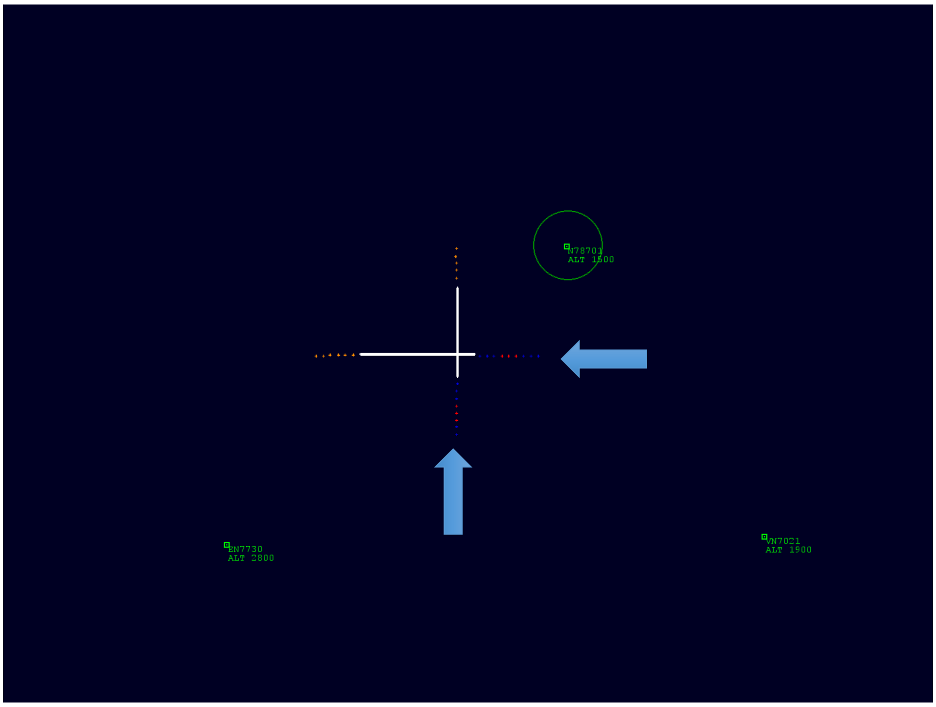
Züger, M. and Fritz, T. (2015). Interruptibility of software developers and its prediction using psycho-physiological sensors. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 2981–2990. ACM.

Appendix: Game instructions

Experiment

You are going to play a game in which you have to land as many planes as possible. At some points you will get interrupted by a simple math problem, which you have to solve as fast and accurate as you can.

Game

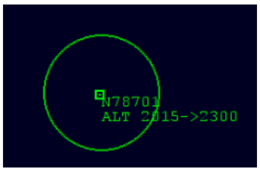


As the screenshot above shows, the game contains two runways and a number of planes, which can be landed on the runways.

In order to land, a plane must approach one of the runways in the correct direction, indicated by the arrows in the picture above, at a height of **1300ft.**

Controls:

- A plane can be selected with the mouse (clicking is not necessary). The selected plane is indicated by the green circle around it.
- The heading of the selected plane can be set with the **WASD** keys.
- The height of the selected plane can be set with the **Page Up** and **Page Down** keys. When the height of a plane is being altered, the final height is displayed on the right side of the arrow as can be seen in the picture. (This plane is ascending from 2015ft. to 2300 ft.)



Ways to lose a plane:

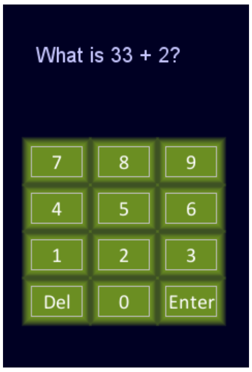
- A plane can crash into another plane if they have the same height and coordinates.
- A plane can crash into the ground when its height is lower than the ground, which is at **1000 ft.**
- A plane can fly out of the screen, in which case it is lost.

Interruption:

At some points in the game you are interrupted by a simple math problem. Try to solve this as accurate as possible, but also take the time into account, because the planes continue travelling!

The answer to the math problem needs to be entered using the mouse and the onscreen keypad. When you make a mistake you can correct it with the onscreen **Del** button, and when you have your answer you need to submit it with the onscreen **Enter** button.

After your answer is submitted, your score will be adjusted and displayed during 1.5 second.



What is 33 + 2?

Score:

You will be scored as follows:

Landed plane:	+100
Lost/crashed plane:	-150
Correctly answered problem:	+100
Incorrectly answered problem:	-500

Try to get the best score you can, good luck!