Workload Assessment for Mental Arithmetic Tasks using the Task-Evoked Pupillary Response

Gerhard Marquart, Joost de Winter

Pupillometry is a promising method for assessing mental workload and could be helpful in the optimization of systems that involve human-computer interaction. The present study focuses on replicating the pupil diameter study by Ahern (1978) for mental multiplications of varying difficulty, using an automatic remote eye tracker. Our results showed that the findings of Ahern were replicated and that the mean pupil diameter and mean pupil diameter change (MPDC) discriminated just as well between the three difficulty levels as did a self-report questionnaire of mental workload (NASA-TLX). A higher mean blink rate was observed during the multiplication period for the highest level of difficulty in comparison with the other two levels. Moderate to strong correlations were found between the MPDC and the proportion of incorrect responses, indicating that the MPDC was higher for participants with a lower performance. For practical applications, validity could be improved by combining pupillometry with other physiological techniques.

Workload Assessment for Mental Arithmetic Tasks using the Task-Evoked Pupillary Response

G. Marquart and J. C. F. de Winter

Department of Biomechanical Engineering, Faculty of Mechanical, Maritime and Materials Engineering, Delft University of Technology, Mekelweg 2, 2628 CD, Delft, The Netherlands

Pupillometry is a promising method for assessing mental workload and could be helpful in the optimization of systems that involve human-computer interaction. The present study focuses on replicating the pupil diameter study by Ahern (1978) for mental multiplications of varying difficulty, using an automatic remote eye tracker. Our results showed that the findings of Ahern were replicated and that the mean pupil diameter and mean pupil diameter change (MPDC) discriminated just as well between the three difficulty levels as did a self-report questionnaire of mental workload (NASA-TLX). A higher mean blink rate was observed during the multiplication period for the highest level of difficulty in comparison with the other two levels. Moderate to strong correlations were found between the MPDC and the proportion of incorrect responses, indicating that the MPDC was higher for participants with a lower performance. For practical applications, validity could be improved by combining pupillometry with other physiological techniques.

Keywords: pupillometry, human factors, pupil diameter, cognitive load

Introduction

Mental workload is an important psychological construct that is challenging to assess on a continuous basis. A commonly used definition of mental workload is the one proposed by Hart and Staveland (1988). These authors defined workload as the “cost incurred by human operators to achieve a specific level of performance”. A valid and reliable assessment method of workload could be helpful in the optimization of systems that involve human-computer interaction, such as vehicles, computers, and simulators. One promising method for measuring workload is pupillometry, which is the measurement of the pupil diameter (e.g., Granholm & Steinhauer, 2004; Marshall, 2007; Schwalm et al., 2008; Klingner et al., 2008; Palinko et al., 2010; Goldinger & Papesh, 2012; Laeng et al., 2012).

Two antagonistic muscles regulate the pupil size, the sphincter and the dilator muscle. Activation of these muscles results in the constriction and dilation of the pupil, respectively. During a mentally demanding task, the pupils have been found to dilate up to 0.5 mm, which is small compared to the maximum dilation of about 6 mm caused by changes in lighting conditions. The involuntary reaction is also called the task-evoked pupillary response (TEPR; Beatty, 1982). In the past, TEPRs were obtained at 1 to 2 Hz by motion picture photography (Hess & Polt, 1964). This required researchers to measure the pupil diameter manually frame by frame (Janisse, 1977). Nowadays, remote non-obtrusive eye trackers are increasingly being used to automatically measure TEPRs, as these devices are getting more and more accurate.

Over the years, researchers have encountered a few challenges in pupillometry. Reflexes of the pupil to changes in luminance, for example, may undermine the validity of TEPRs. One way to achieve this is by strictly controlling luminance, but this limits the usability of pupillometry. Marshall (2000) reported to have found a valid way to filter out the pupil light reflex using wavelet transform techniques. She patented this method and dubbed it the “index of cognitive activity”. The influence of gaze direction on the measured pupil size is another issue. Whereas Pomplun and Sunkara (2003) reported a systematic dependence of pupil size on gaze direction, Klingner et al. (2008) argued that the ellipse-fitting method for the estimation of the pupil size is not affected by perspective distortion.

In the last few decades many researchers have investigated the pupillary response for different types of tasks. Typically, the dilation was found to be higher for more challenging tasks (Beatty & Kahneman, 1966; Ahern, 1978). Not only task demands have been found to influence the pupil diameter, but also factors like anxiety, stress, and fatigue. Tryon (1975) and Janisse (1977) extensively reviewed known sources regarding variation in pupil size. Back then, Janisse (1977) commented on the underexplored area of individual differences in intelligence. Ahern

(1978) continued on this topic and discovered that persons scoring higher on intelligence tests showed smaller pupillary dilations on tasks of fixed difficulty. In a more recent study, Van Der Meer et al. (2010) found greater pupil dilations for individuals with high fluid intelligence than with low fluid intelligence during the execution of geometric analogy tasks. Thus, the results are not consistent and demand further investigation.

The present study focuses on replicating the film-based pupil diameter study by Ahern (1978) for mental multiplications of varying difficulty (43 participants, 1376 trials), and is intended as a follow-up study of Klingner (2010). Klingner replicated Ahern's results with an automatic remote eye tracker and found a clear difficulty effect, with the more difficult multiplications showing a greater dilation. With more participants (30 vs. 12) and trials (1350 vs. 431) than Klingner, the present study aims to analyze the TEPRs for three levels of difficulty in high temporal detail, to provide new insights into individual differences, and to compare the effect sizes between the pupil diameter and a classic subjective measurement method of workload, the NASA-TLX. Additionally, the mean pupil diameter change rate (MPDCR) will be examined, which is a new measure introduced by Palinko et al. (2010). He expected it to be useful in assessing moment-to-moment changes in mental workload. Lastly, this study discusses the feasibility of using the pupil diameter in practical applications. One example of such an application is adaptive automation, which is "an approach to automation design where tasks are dynamically allocated between the human operator and computer systems" (Byrne & Parasuraman, 1996). As mentioned above, reliability and validity are crucial in this.

The digits in the task in this study were presented visually, in contrast to the experiment conducted by Ahern, where the digits were presented aurally. This was done to gain more temporal consistency in the presentation duration of the numbers. Like Klingner (2010), the pupil diameter was recorded with an automatic remote eye tracker.

Method

Ethics Statement

The research was approved by the Human Research Ethics Committee (HREC) of the Delft University of Technology (TU Delft). ('Workload Assessment for Mental Arithmetic Tasks using the Task-Evoked Pupillary Response', date: January 29, 2015). All participants provided written informed consent.

Participants

Thirty participants (2 women and 28 men), aged between 19 and 38 years (mean = 23, SD = 4.1 years) were recruited to volunteer in this experiment (25 MSc/BSc, 3 PhD, and 2 graduate students). Individuals wearing glasses or lenses were excluded from participation. All participants read and signed an informed consent form, explaining the purpose and procedures of the experiment and received €5 in compensation for their time.

Equipment

The SmartEye DR120 remote eye tracker, with a sampling rate of 120 Hz, was used to record the participant's pupil diameter, eyelid opening, and gaze direction while sitting behind a desktop computer (see Fig. 1, left). The pupil diameter was estimated by averaging the five longest lines found in the pupil (Wilhelm, 2010). This method is comparable to the ellipse-fitting method, since they are both unaffected by perspective distortion. In order to obtain accurate measurements, a headrest was used to avoid head displacements. The eye tracker was equipped with a 24-inch screen, which was positioned approximately 65 cm in front of the sitting participant and was used to display task-relevant information. The outcome of a task had to be entered using the numeric keypad of a keyboard. The experiment took place in a room where there was office lighting and where daylight could not enter. A screen background with variable brightness was used, which was designed to minimize the pupillary light reflex in case a participant looked away from the center of the screen (see Fig. 1, right; Marquart, 2015).

Procedure

The participants were requested to perform 50 trials of mental arithmetic tasks (multiplications of two numbers), five of which were used as a short training. The remaining 45 trials were sorted by the outcome of their multiplication and evenly divided into 3 sessions of varying difficulty (easy, medium, and hard; see Appendix A). Level 1 contained the 15 easiest multiplications (outcomes ranging between 72 and 117), Level 2 contained 15 multiplications of intermediate difficulty (outcomes between 119 and 192), and Level 3 contained the 15 hardest multiplications (outcomes between 196 and 324).

The sequence of the three sessions was counterbalanced across the participants. Each trial was initiated by the participant with a button press and started with a 4 second accommodation period, followed by a 1 second visual presentation of two numbers (multiplicand and multiplier) between 6 and 18, with a 1.5 second pause in between. The participants were asked to multiply the two numbers and type their answer on the numeric keypad 10 seconds after the multiplier disappeared (see Table 1). Thus, the total duration of one trial was 17.5 seconds (4 + 1 + 1.5 + 1 + 10). When the numbers were not presented, a double "X" was shown to avoid pupillary reflexes caused by changes in brightness or contrast.

After each session, participants were asked to fill out a NASA-TLX questionnaire to assess their subjective workload on six facets: mental demand, physical demand, temporal demand, performance, effort, and frustration (Hart & Staveland, 1988). All questions were answered on a scale from 0 % (very low) to 100 % (very high). For the performance question, 0 % meant perfect and 100 % was failure. The participants' overall subjective workload was obtained by averaging the scores across the six items. The total duration of the experiment was approximately 30 minutes.

Instructions to Participants

Before the experiment started, the participants were requested to position themselves in front of the monitor with their chin leaning on the headrest. They were instructed to stay still and keep their gaze fixed and focus (not stare) at the center of the screen throughout a trial. In addition, participants were asked to blink as little as possible, obviously without causing irritation, and to start each trial with 'a clear mind' (i.e., not thinking about the previous trial). If the participants could not complete the multiplication in time, they were instructed to enter zero as their answer.

Data Processing

The data were processed in two steps. In the first step, the missing values in the pupil diameter data (lost during recording) were removed and the signals were repaired with linear interpolation (see Fig 2, left, for an illustration). On average, 1.2% of the data were lost, so this processing step did not significantly influence the results. Step two included the removal of the blinks and the poor-quality data. During a blink, the eyelid opening rapidly diminishes to zero and then increases in a few tenths of a second until it is fully open again. It is impossible to track the pupil diameter while blinking. These instances in time were removed from the data (for a detailed description of how the blinks were identified and removed, see Appendix B). The pupil diameter quality signal (provided by SmartEye software) was used to filter out the poor quality data. This signal ranges from 0 to 1, with values close to 1 indicating a good quality (SmartEye AB, 2013). All data points with a pupil diameter quality below 0.75 were removed. Trials containing less than 70% of the original data were excluded from the analysis. Of the initial 1350 trials from 30 participants, 1110 trials spread of 29 participants passed these criteria. The results of one participant (45 trials) were discarded completely. The gaps in the remaining trials were again filled using linear interpolation (see Fig 2, right), a process that does not substantially alter the data according to Beatty and Lucero-Wagoner (2000).

The last 0.4 seconds of the accommodation period (3.6–4 s) were defined as the pupillary baseline, as was done by Klingner (2010). The mean pupil diameter of the baseline period of each trial was subtracted from each trial to accommodate for any possible shifts or drifts. The mean pupil diameter change (MPDC) for each participant was then obtained by averaging all trials per level of difficulty. Similarly, the mean pupil diameter (MPD) for each participant was obtained but then without subtracting the mean pupil diameter of the baseline period. The MPDCR was calculated for each participant as the average velocity (mm/sample) or change in MPD between two points in time. In order to compare the three difficulty levels, the MPD and MPDC were analyzed at eight fixed points in time from the multiplier and calculation periods. Both measures were reported such that a complete picture of the pupillary behavior could be given (Beatty & Lucero-Wagoner, 2000). The MPDCR was assessed across the seven interim periods.

In addition to these analyses, the mean blink rate (MBR) for two different periods in time was calculated and Pearson's *r* correlation coefficients were obtained between the MPDC and the NASA-TLX and responses. Cohen's d_z effect size (see Eq. 1) was calculated to determine at which points in time the differences in MPDC between the three levels of difficulty were largest.

$$Cohen's\, d_z = \frac{|M_i - M_j|}{\sqrt{SD_i^2 + SD_j^2 - 2 * r * SD_i * SD_j}} \qquad (1)$$

Statistical Analyses

The pupil diameter measures (MPD, MPDC, and MPDCR), the blink rates (MBR), and the results of the NASA-TLX questionnaire were analyzed with a one-way repeated measures ANOVA. Tukey's honest significant difference test was used with a significance level of 0.05 to determine whether pairs of conditions were significantly different from each other. To determine whether the Pearson correlation coefficients were significantly different from zero, a Bonferroni correction was applied. Thus, because 24 correlation coefficients were calculated (8 points in time * 3 levels of difficulty), the significance level was reduced to 0.002 (0.05/24).

Results

Mean Pupil Diameter (MPD)

The mean pupillary response during the mental multiplication task of 29 participants is shown in Figure 3a. It can be seen that the MPD was higher for the higher of levels of difficulty at all points in time. The pattern of the MPD was similar for all levels during the first ten seconds. Hereafter, the response seems different for each level and was split for further analysis in seven periods with eight points (see Fig. 3b). The points are indicated by a 'P' and the numbers of the periods are shown in parentheses.

The means and standard deviations of the MPD for the eight points in time and three levels of difficulty are shown in Table 2, together with the effect sizes and the p-values of the one-way repeated measures ANOVA and the pairwise comparisons. The results confirm that the MPD was significantly higher for the more difficult levels at all points in time and between most of the conditions.

Mean Pupil Diameter Change (MPDC)

Figures 4a shows the MPDC of 29 participants as a function of the level of difficulty. As mentioned above, this measure takes into account the shift of the baseline by subtracting the mean of the baseline period of each trial. The difference between the three pupillary responses during the calculation period can now be seen more clearly. Again, the multiplier and calculation were split into seven periods by eight points (see Fig. 4b).

The results of the analysis of the MPDC at the eight points in time and three levels of difficulty are shown in Table 2. It shows that a significant difference occurred at points 4 to 8 and that the effect size was largest at point 7.

A scatterplot of the MPDC at points 1, 5 and 8 of Level 1 versus Level 3 gives insight into the differences between individuals (see Fig. 5). The MPDC of Level 3 lies above the unity line for 16, 28, and 29 of the 29 participants for the three points respectively, and has a range of about 1 mm.

Mean Pupil Diameter Change Rate (MPDCR)

Figure 6 shows the MPDCR of the 29 participants as a function of the difficulty level, for the seven periods. A positive value indicates overall pupil dilation during that period and a negative value means overall contraction of the pupil diameter. In the first two periods, the diameter increased with approximately equal velocity for the three levels. During the other periods, the velocities decreased and became negative. Significant differences were found between the three conditions (see also Table 2).

Self-reported workload (NASA-TLX)

The results of the NASA-TLX questionnaire are shown in Figure 7. For almost all items, the TLX score was significantly higher for the more difficult multiplications (see also Table 2). Only the subjective physical workload did not differ significantly across the levels of difficulty.

Responses

The percentage correct responses for Levels 1, 2, and 3 were respectively 94.2%, 93.8%, and 69.2%. Figure 8 shows the MPD for Level 3 of all trials, and separated for correct and incorrect responses. Too few incorrect answers were given for the other two levels and the results for these levels are therefore not reported. The MPD of the incorrect responses shows the same pattern as the one of the correct responses for the first twelve seconds. From this moment onward, the MPD belonging to the trials with incorrect responses was higher. A significant difference was observed at point 2 and 8 between the two lines when the same eight-point analysis was used (see Appendix C).

Effect Size

The effect size estimate Cohen's d_z was calculated for the MPDC between pairs of difficulty levels for every point in time. Figure 9 shows the results. Large effect sizes arose after approximately 11 seconds since the start of the trial, especially between Levels 1 and 3.

Correlations

The results of the correlation analyses between the MPDC, NASA-TLX, and proportion of incorrect responses are shown in Table 3. For the MPDC, the table shows overall positive correlations, for the eight points in time and for the three different levels of difficulty. Between the MPDC and the percentage of incorrect responses, two statistically significant positive correlation coefficients were observed at points 1 and 2. Furthermore, Table 3 shows that people who experienced higher subjective workload (i.e., a higher NASA-TLX score) generally gave more incorrect responses.

Blinks

Figure 10 shows the MBR of all participants and sorted per level of difficulty during a period with low (2–6.5 s) and high (6.5–13 s) mental demands. The MBR of Level 3 during the second period was significantly higher than those of Level 1 and 2. More details can be found in Table 2.

Discussion

Pupil Diameter Results

The results showed that the overall MPD was higher for the higher levels of difficulty. Points 7 and 8 showed the largest differences. These findings demonstrate that the mean or baseline of the pupil diameter can shift during mental activity. If the pupil was given more time to recover from the previous trial, by increasing the length of the accommodation period, the difference of the MPD between the three levels of difficulty in the first period would probably have been smaller.

A remarkable finding is the behavior of the MPD during the first three seconds of the accommodation period (0–3 s). Where a clear decline from the start or a low horizontal line might be expected, the MPD starts to decline only after three seconds. This unexpected effect may have been caused by the fact that participants looked away from the center of the screen when their outcome to the multiplication had to be entered. Although the responses were not given during the accommodation period, the fluctuation could be an aftereffect because the trials came in relatively quick succession. During the presentation of the multiplicand and the pause (4–6.5 s) the MPD decreased further, at a slower pace however, which seems to indicate memory load (cf. Kahneman & Beatty, 1966). This small increase of the pupil diameter after the presentation of the first number was also observed by Ahern (1978) and Klingner (2010).

What is notable in the MPDC figure (Fig. 4) is that the pupillary behavior among the three difficulty levels was highly similar during the first few seconds after the presentation of the multiplier (6.5–9 s). This might be due to the strategy that the participants used. One can imagine that the first step in each multiplication, regardless of its difficulty, is similar. For example, the first step for many people of the Level 1 multiplication 7x14 would probably be 7x10. This is comparable to the first step of the Level 3 multiplication 14x18, which would then be 14x10. These observations are in line with the TEPRs obtained by Ahern (1978). She also observed a similar response among the three levels of difficulty at the beginning of the calculation. The MPDC during the other periods was found to differ significantly between the three levels, particularly when Levels 1 and 2 were compared to Level 3. This finding is in accordance with the results in the scatterplot (Fig. 5), where 28 and 29 of the 29 participants had a higher MPDC for Level 3 than for Level 1, for points 5 and 8, respectively.

The results of the MPDCR illustrate that the effect sizes are smaller when compared to the results of the MPDC measure. It does provide, however, a clear understanding of when the muscles of the pupil relax and hence when the mental workload decreases.

Self-reported Workload

According to the results of the NASA-TLX questionnaire, the classification of the arithmetic tasks was done properly, since a statistically significant difference was found in the subjective mental workload across all three levels. The big contrast between the subjective mental and physical workload underlines that the task was predominantly mentally demanding. Not to be overlooked are the roles of the subjective temporal demand and frustration. Looking at the increase of the MPD of the incorrect responses after 12 seconds for Level 3 (Fig. 8), it is plausible that, although only one significant difference was found, this increase was caused by the time pressure of the task or the frustration of not having solved the multiplication yet, instead of increased task demands.

Correlation Analyses

At the first two points in particular, moderate to strong correlations were found between the MPDC and the proportion of incorrect responses. A similar but weaker effect was obtained between the MPDC and the NASA-TLX. It may not be surprising that the strongest correlations were found at points 1 and 2, considering the fact that at these points in time probably all participants were still calculating. Once the task has been completed, the pupil diameter decreases again (cf. Kahneman, 1966 for similar findings in a memory paradigm). Since this decline does not occur at the same time for each trial, this causes higher variability and lower correlation coefficients. Apart from that, the results seem to indicate that the MPDC was higher for participants who gave more incorrect responses and experienced a higher workload. This could help in determining the feasibility of using the pupil diameter in adaptive automation. Combining the pupil diameter with other assessment methods could help increase validity and robustness. Correlations of similar size between the pupil diameter and the proportion of incorrect responses and NASA-TLX were respectively found by Payne et al. (1986) and Recarte et al. (2008).

Another interesting question related to Figure 8 showing the trials with the correct versus incorrect responses is: were the participants really trying to complete the task or did they give up on the task because it was too difficult? If the latter were the case, one would expect an early decline of the MPD. But the opposite is true, instead. A small increase of the MPD was measured, suggesting that the participants were trying hard to complete the task.

Blink Rate

The relation between mental workload and blink rate has been unclear (Kramer, 1990; Recarte et al., 2008; Marquart et al., in press). The results in the present study show that the MBR is significantly higher for Level 3 than for Level 1 and 2 during mentally demanding periods. However, the differences between Level 1 and 2 and the two periods in time are small. The MBR therefore appears to be less sensitive than the MPDC and more suited for the detection of a task's overall mental workload, because of its low temporal resolution.

Conclusions and Recommendations

It is concluded that the results of Ahern (1978) and Klingner (2010) have been accurately replicated with the SmartEye DR120 remote eye tracker. The partial eta squared effect sizes (η_p^2) for point 7 and 8 of the MPD, MPDC, and NASA-TLX are approximately the same (~0.6), which demonstrates that pupil diameter measurements can be just as valid as the NASA-TLX. An attempt was made to provide more insight into the individual differences of TEPRs by means of a correlation analysis. Results showed a few moderate to strong correlations at the beginning of the calculation period between the MPDC and the NASA-TLX, on the one hand, and the ratio of incorrect responses, on the other.

Thus, it seems possible to assess workload by tracking the pupil diameter. However, the validity of pupil diameter measurements may need improvement before it could be implemented in practice. One possible way to do this is by combining pupillometry with other physiological measures, such as blink and heart rate (Kahneman et al., 1969; Molen et al., 1989; Just et al., 2003; Satterthwaite et al., 2007; Haapalainen et al., 2010). Additionally, future research could focus on improving signal analysis techniques that filter out effects other than mental workload, such as the light reflex.

The supplementary materials provide the measurement data, software, and scripts that would allow others to reproduce these results:
https://www.dropbox.com/s/fbaz0cvcoxnu98q/Supplementary_Material_Gerhard_Marquart.zip?dl=0

References

Ahern, S.K. (1978). Activation and intelligence: Pupillometric correlates of individual differences in cognitive abilities. *Unpublished doctoral dissertation,* University of California, Los Angeles.

Beatty, J. & Kahneman, D. (1966). Pupillary changes in two memory tasks. *Psychonomic Science, 5*, 371-372.

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

Beatty, J. & B. Lucero-Wagoner, B. (2000). The pupillary system. In J. Cacioppo, L. Tassinary & G. Berntson (Eds.), *Handbook of Psychophysiology*, Cambridge: Cambridge University Press.

Byrne, E.A. & Parasuraman, R. (1996). Psychophysiology and adaptive automation. *Biological Psychology, 42*, 249-268.

Goldinger, S.D. & Papesh, M.H. (2012). Pupil Dilation Reflects the Creation and Retrieval of Memories. *Psychological Science, 21*, 90-95.

Granholm, E. & Steinhauer, S.R. (2004). Pupillometric measures of cognitive and emotional processes. *International Journal of Psychophysiology, 52*, 1-6.

Hart, S. & Staveland, L. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. Hancock & N. Meshkati (Eds.), *Human Mental Workload*. Amsterdam: North Holland Press, pp. 139-183.

Hess, E.H. & Polt, J.M. (1964). Pupil sizes in relation to mental activity during simple problem solving. *Science, 143*, 1190-1192.

Janisse, M.P. (1977). Pupillometry: The psychology of the pupillary response. *Hemisphere Publishing Co*.

Just, M.A., Carpenter, P.A., & Miyake, A. (2003). Neuroindices of cognitive workload: Neuroimaging pupillometric and event-related potential studies of brain work. *Theoretical Issues in Ergonomics Science, 4*, 56-88.

Kahneman, D., Tursky, B., Shapiro, D. & Crider, A. (1969). Pupillary, heart rate, and skin resistance changes during a mental task. *Journal of Experiment Psychology, 79*, 164-167.

Kahneman, D. & Beatty, J. (1966). Pupil Diameter and Load on Memory, *Science, 154*, 1583-1585.

Klingner, J., Kumar, R. & Hanrahan, P. (2008). Measuring the task-evoked pupillary response with a remote eye tracker. *Proceedings of the 2008 Symposium on Eye Tracking Research & Applications*, 69-72.

Klingner, J. (2010). Measuring cognitive load during visual tasks by combining pupillometry and eye tracking. *Ph.D. dissertation, Stanford University Computer Science Department*.

Kramer, A.F. (1990). Physiological metrics of mental workload: a review of recent progress. In D.L. Damos (Ed.), *Multiple-task performance*, (pp. 279-328). Taylor & Francis, London.

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

Marquart, G., Cabrall, C., & De Winter, J.C.F. (in press). Review of eye-related measures of drivers' mental workload. *6th International Conference on Applied Human Factors and Ergonomics.*

Marquart, G. (2015). Pupil Light Reflex Suppression by Variable Screen Brightness. *Unpublished manuscript.* Delft University of Technology.

Marshall, S. (2000). Method and apparatus for eye tracking and monitoring pupil dilation to evaluate cognitive activity. *U.S. patent no.* 6,090,051.

Marshall, S. (2007). Identifying cognitive state from eye metrics. *Aviation, space, and environmental medicine*, 78 (Supplement 1), B165-B175.

Molen, M.W., Boomsma, D.I., Jennings, J.R., & Nieuwboer, R.T. (1989). Does the heart know what the eye sees? A cardiac/pupillometric analysis of motor preparation and response execution. *Psychophysiology, 26*, 70-80.

Palinko, O., Kun, A., Shyrokov, A. & Heeman, P. (2010). Estimating cognitive load using remote eye tracking in a driving simulator. *Proceedings of the 2010 Symposium on eye-tracking research & applications*, 141-144.

Payne, D.T., Parry, M.E. & Harasymiw, S.J. (1968). Percentage of pupillary dilation as a measure of item difficulty. *Perception & Psychophysics, 4*, 139-143.

Pomplun, M. & Sunkara, S. (2003). Pupil dilation as an indicator of cognitive workload in human-computer interaction. *Proceedings of the International Conference on Human-Computer Interaction*, 3, 542-546.

Recarte, M.A., Pérez, E., Conchillo, A. & Nunes, L.M. (2008). Mental Workload and Visual Impairment: Differences between Pupil, Blink, and Subjective Rating. *The Spanish Journal of Psychology*, *11*, 374-385.

Satterthwaite, T.D., Green, L., Myerson, J., Parker, J., Ramaratnam, M., & Buckner, R.L. (2007). Dissociable but inter-related systems of cognitive control and reward during decision making: evidence from pupillometry and event-related fMRI. *Neuroimage, 37*, 1017-1031.

Schwalm, M., Keinath, A. & Zimmer, H.D. (2008). Pupillometry as a method for measuring mental workload within a simulated driving task. In F. Flemisch, B. Lorenz, H. Oberheid, K. Brookhuis, & D. De Waard, (Eds.), *Human Factors for assistance and automation*, 75-88. Maastricht, Netherlands: Shaker Publishing.

SmartEye AB. (2013). Programmer's Guide, Revision 1.3.
Tryon, W.W. (1975). Pupillometry: A survey of sources of variation. *Psychophysiology, 12*, 90-93.
Wilhelm, T. (2010). *Accuracy and precision of the pupil size measured with SmartEye Pro 5.6.0*. Smart Eye AB.

Table 1

Timeline of an individual trial

Period	Start time (s)	End time (s)	Symbol
Accommodation	0.0	4.0	X X
Baseline	3.6	4.0	X X
Multiplicand	4.0	5.0	0 8
Pause	5.0	6.5	X X
Multiplier	6.5	7.5	1 6
Calculation	7.5	17.5	X X
Response	17.5	when pressing enter key	N/A

Table 2

Mean Pupil Diameter Change (MPDC), Mean Pupil Diameter Change Rate (MPDCR), NASA-TLX, and Mean Blink Rate (MBR). The means (M) and standard deviations (SD) of 29 participants are shown per level of difficulty of the multiplications. P1-P8 refers to the eight points in time, while (1)-(7) refers to the seven periods.

	Level 1	**Level 2**	**Level 3**	**p-value**	**Effect size**	**Pairwise comparison of levels**		
	M (SD)	M (SD)	M (SD)		η_p^2 (η_G^2)	1 vs. 2	1 vs. 3	2 vs. 3
MPD (mm) (N = 29)								
P1	3.770 (0.456)	3.804 (0.467)	3.881 (0.490)	**0.004**	0.18 (0.01)	0.555	**0.003**	0.051
P2	3.814 (0.480)	3.865 (0.486)	3.954 (0.516)	**$1.94*10^{-4}$**	0.26 (0.01)	0.242	**$1.33*10^{-4}$**	**0.019**
P3	3.919 (0.471)	3.979 (0.481)	4.061 (0.531)	**0.001**	0.22 (0.01)	0.224	**$6.53*10^{-4}$**	0.069
P4	3.902 (0.456)	4.003 (0.478)	4.116 (0.522)	**$2.02*10^{-5}$**	0.32 (0.03)	**0.048**	**$1.10*10^{-5}$**	**0.024**
P5	3.836 (0.429)	3.949 (0.488)	4.140 (0.521)	**$7.14*10^{-9}$**	0.49 (0.06)	**0.025**	**$5.26*10^{-9}$**	**$8.79*10^{-5}$**
P6	3.767 (0.451)	3.894 (0.490)	4.127 (0.518)	**$1.98*10^{-9}$**	0.51 (0.09)	**0.026**	**$2.25*10^{-9}$**	**$2.77*10^{-5}$**
P7	3.720 (0.428)	3.815 (0.474)	4.130 (0.500)	**$3.50*10^{-12}$**	0.61 (0.12)	**0.104**	**$9.63*10^{-10}$**	**$1.81*10^{-8}$**
P8	3.693 (0.437)	3.781 (0.460)	4.114 (0.493)	**$1.03*10^{-12}$**	0.63 (0.14)	**0.148**	**$9.59*10^{-10}$**	**$4.45*10^{-9}$**
MPDC (mm) (N = 29)								
P1	-0.001 (0.087)	0.004 (0.115)	0.024 (0.085)	0.474	0.03 (0.01)	0.977	0.486	0.613
P2	0.043 (0.094)	0.065 (0.118)	0.097 (0.120)	0.064	0.09 (0.04)	0.583	0.052	0.351
P3	0.148 (0.148)	0.179 (0.148)	0.203 (0.152)	0.178	0.06 (0.02)	0.548	0.153	0.685
P4	0.131 (0.179)	0.203 (0.149)	0.259 (0.171)	**0.001**	0.21 (0.10)	0.085	**$8.69*10^{-4}$**	0.220
P5	0.064 (0.204)	0.148 (0.164)	0.282 (0.205)	**$7.26*10^{-8}$**	0.44 (0.19)	**0.036**	**$4.31*10^{-8}$**	**$4.20*10^{-4}$**
P6	-0.005 (0.196)	0.094 (0.193)	0.270 (0.228)	**$1.54*10^{-9}$**	0.52 (0.24)	**0.022**	**$1.93*10^{-9}$**	**$2.67*10^{-5}$**
P7	-0.051 (0.186)	0.015 (0.207)	0.273 (0.226)	**$6.52*10^{-14}$**	0.66 (0.33)	**0.116**	**$9.56*10^{-10}$**	**$1.31*10^{-9}$**
P8	-0.078 (0.179)	-0.018 (0.208)	0.259 (0.248)	**$1.72*10^{-12}$**	0.62 (0.32)	**0.251**	**$9.62*10^{-10}$**	**$3.61*10^{-9}$**
MPDCR (mm/sample) (N = 29) (x $1*10^{-3}$)								
(1)	0.361 (0.698)	0.513 (0.657)	0.611 (0.942)	0.143	0.07 (0.02)	0.450	0.124	0.719
(2)	0.586 (0.676)	0.632 (0.578)	0.592 (0.662)	0.902	0.00 (0.00)	0.909	0.998	0.930
(3)	-0.094 (0.641)	0.134 (0.587)	0.309 (0.415)	**0.006**	0.17 (0.08)	0.150	**0.004**	0.324
(4)	-0.371 (0.438)	-0.305 (0.477)	0.130 (0.443)	**$3.67*10^{-5}$**	0.31 (0.20)	0.820	**$8.06*10^{-5}$**	**$6.03*10^{-4}$**
(5)	-0.383 (0.475)	-0.302 (0.491)	-0.070 (0.567)	**0.044**	0.11 (0.06)	0.797	**0.042**	0.168
(6)	-0.257 (0.438)	-0.438 (0.477)	0.017 (0.443)	**$4.96*10^{-4}$**	0.24 (0.15)	0.235	**0.040**	**$3.30*10^{-4}$**

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1105v1 | CC-BY 4.0 Open Access | rec: 21 May 2015, publ: 21 May 2015

(7)	-0.152 (0.475)	-0.184 (0.491)	-0.080 (0.567)	**0.694**	0.01 (0.01)	0.964	0.832	0.681
NASA-TLX (%) (N = 30)								
Total	20.744 (12.783)	30.883 (13.060)	48.658 (14.441)	**$1.65*10^{-16}$**	0.71 (0.43)	**$1.80*10^{-4}$**	**$9.56*10^{-10}$**	**$1.90*10^{-9}$**
Mental	33.833 (21.037)	46.833 (16.942)	70.180 (16.746)	**$1.67*10^{-12}$**	0.61 (0.41)	**0.004**	**$9.57*10^{-10}$**	**$4.00*10^{-7}$**
Physical	16.000 (17.291)	19.000 (18.773)	19.833 (20.235)	0.152	0.06 (0.01)	0.314	0.155	0.913
Temporal	18.667 (15.025)	29.167 (18.293)	53.167 (23.359)	**$3.67 *10^{-12}$**	0.60 (0.37)	**0.021**	**$9.60*10^{-10}$**	**$1.38*10^{-7}$**
Performance	10.033 (11.643)	20.667 (16.904)	40.433 (22.509)	**$8.09*10^{-11}$**	0.55 (0.35)	**0.014**	**$1.01*10^{-9}$**	**$3.74*10^{-6}$**
Effort	28.000 (18.782)	43.133 (17.362)	63.500 (21.502)	**$9.89*10^{-12}$**	0.58 (0.37)	**$9.37*10^{-4}$**	**$9.61*10^{-10}$**	**$9.87*10^{-6}$**
Frustration	17.933 (17.187)	26.500 (23.713)	44.833 (28.542)	**$2.51*10^{-9}$**	0.49 (0.19)	0.057	**$2.99*10^{-9}$**	**$1.50*10^{-5}$**
MBR (N = 30)								
(2–6.5 s)	0.256 (0.301)	0.215 (0.199)	0.321 (0.492)	0.084	0.03 (0.02)	0.869	0.714	0.406
(6.5–13 s)	0.238 (0.217)	0.265 (0.232)	0.369 (0.336)	**0.008**	0.16 (0.04)	0.799	**0.008**	**0.041**

Note. Statistically significant differences are indicated in boldface.

Table 3

Pearson's r correlations between the mean pupil diameter change (MPDC), percentage of incorrect responses, and the overall NASA-TLX scores, for the three levels of difficulty.

	Level 1	**Level 2**	**Level 3**	**Mean**
	r (p-value)	r (p-value)	r (p-value)	r (p-value)
MPDC vs. Overall NASA-TLX (N = 29)				
P1	-0.009 (0.961)	0.195 (0.310)	0.201 (0.296)	0.355 (0.059)
P2	-0.131 (0.498)	0.288 (0.130)	0.079 (0.685)	0.247 (0.195)
P3	-0.035 (0.857)	0.045 (0.818)	0.009 (0.964)	0.040 (0.836)
P4	0.303 (0.109)	0.066 (0.733)	0.030 (0.878)	0.272 (0.153)
P5	0.243 (0.204)	0.115 (0.554)	0.010 (0.956)	0.168 (0.384)
P6	0.211 (0.272)	0.196 (0.307)	-0.016 (0.934)	0.139 (0.472)
P7	0.175 (0.363)	0.203 (0.290)	0.163 (0.397)	0.226 (0.238)
P8	0.056 (0.766)	0.258 (0.176)	0.163 (0.399)	0.215 (0.262)
MPDC vs. % Incorrect responses (N = 29)				
P1	0.353 (0.060)	0.438 (0.017)	0.349 (0.063)	0.643 (**$1.70*10^{-4}$**)
P2	0.228 (0.233)	0.505 (0.005)	0.264 (0.166)	0.561 (**0.002**)
P3	0.069 (0.722)	0.256 (0.180)	0.130 (0.500)	0.196 (0.309)
P4	0.306 (0.106)	0.254 (0.183)	0.122 (0.528)	0.312 (0.099)
P5	0.232 (0.224)	0.159 (0.409)	0.027 (0.887)	0.199 (0.302)
P6	0.064 (0.740)	0.205 (0.285)	0.016 (0.932)	0.123 (0.525)
P7	0.048 (0.803)	0.321 (0.090)	0.087 (0.653)	0.226 (0.238)
P8	0.063 (0.744)	0.249 (0.193)	0.137 (0.477)	0.218 (0.255)
Overall NASA-TLX vs. % Incorrect responses (N = 30)				
	0.566 (**0.001**)	0.352 (0.056)	0.532 (0.002)	0.580 (**$7.91*10^{-4}$**)

Note. Statistically significant correlations are indicated in boldface.

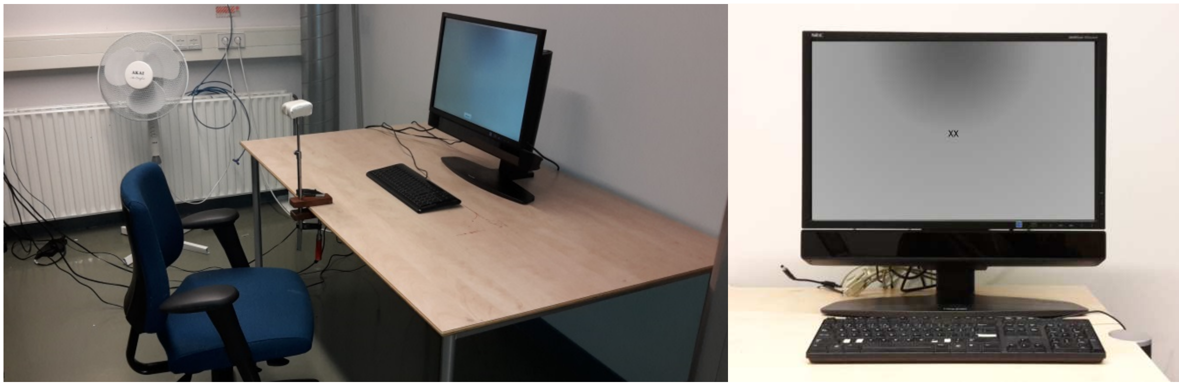

Figure 1. Experimental equipment. Left: eye tracker, monitor, table, headrest, chair, keyboard. Right: task display.

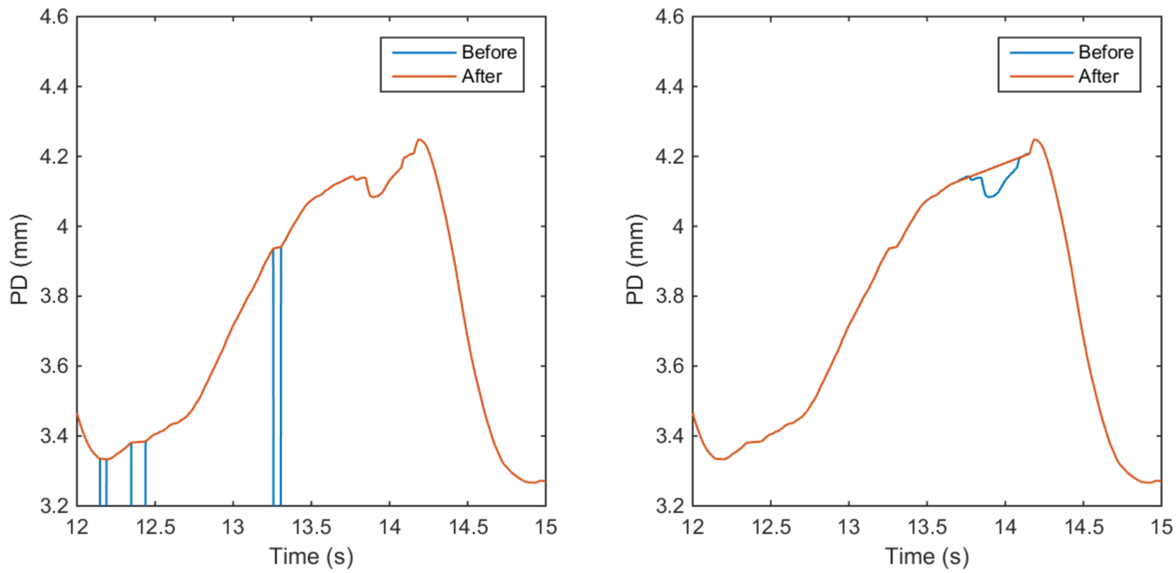

Figure 2. Example of the data processing steps. Left: Pupil diameter (PD) before and after linear interpolation for missing values. Right: PD before and after blink and poor quality data removal and linear interpolation.

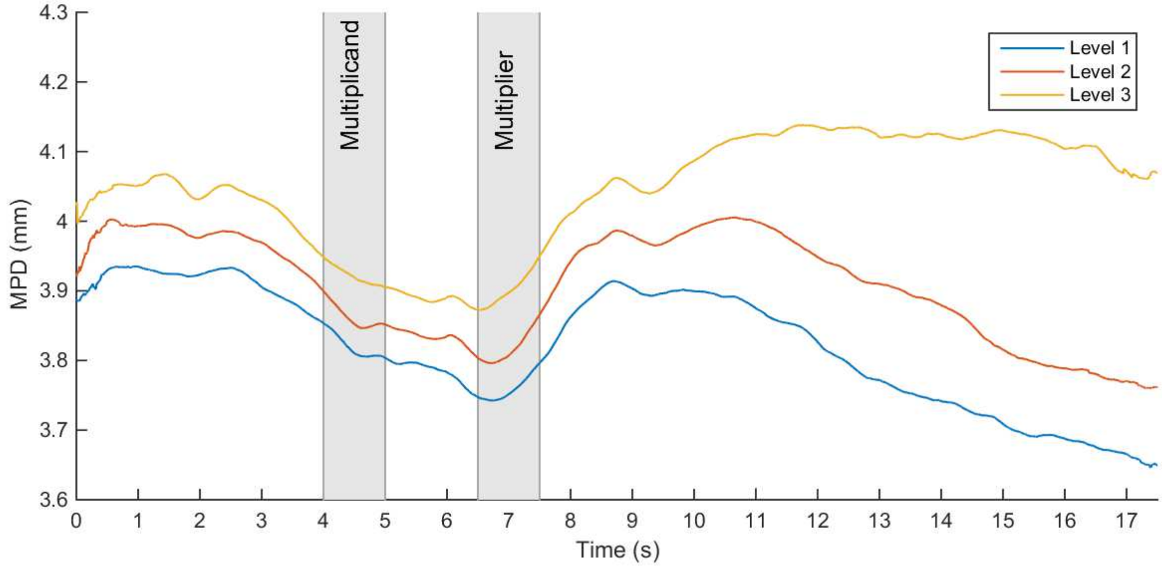

Figure 3a. Mean pupil diameter (MPD) during the mental multiplication task of 29 participants, for the three levels of difficulty. The grey bars represent the periods where the multiplicand and multiplier were shown on the screen. The numbers were masked by an “XX” during the rest of the trial.

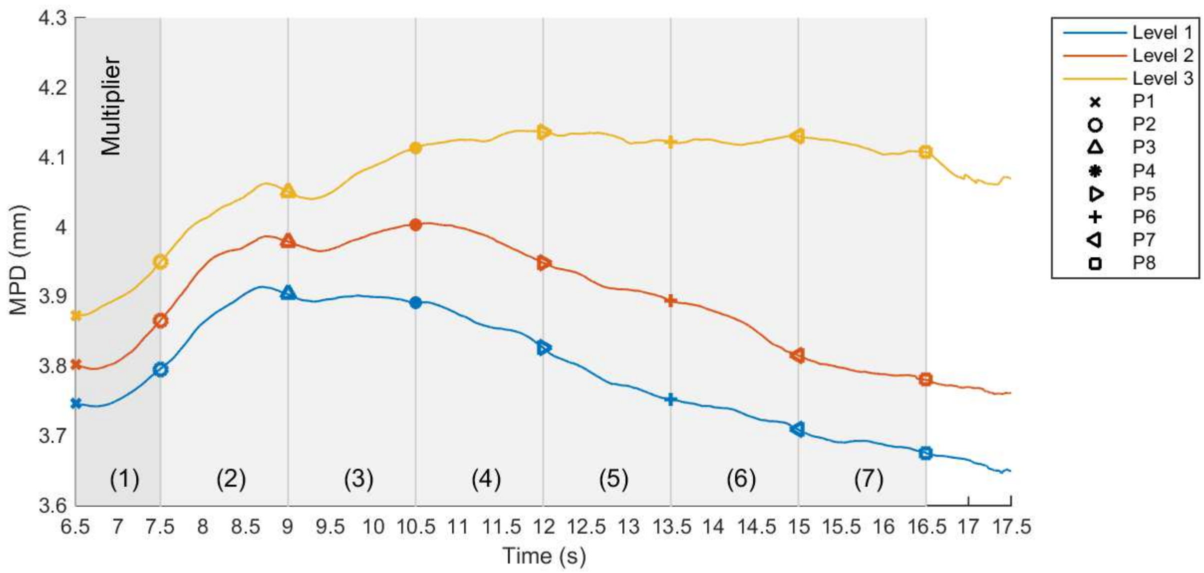

Figure 3b. Mean pupil diameter (MPD) during the presentation of the multiplier and the calculation period of 29 participants, for the three levels of difficulty. The seven periods are indicated in parenthesis.

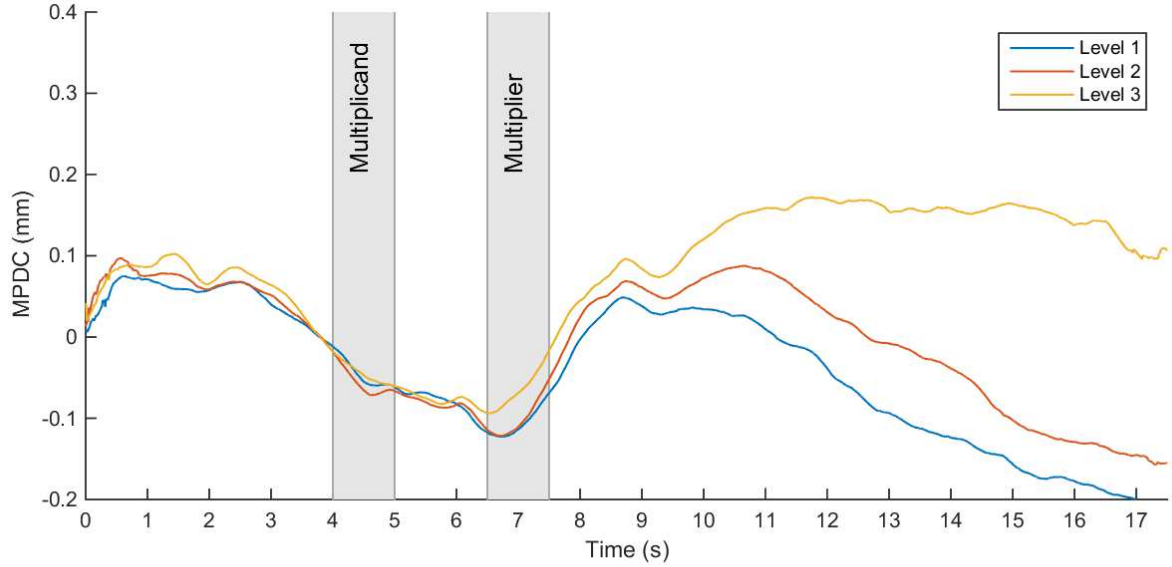

Figure 4a. Mean pupil diameter change (MPDC) during the mental multiplication task of 29 participants, for the three levels of difficulty. The grey bars represent the periods where the multiplicand and multiplier were shown on the screen. The numbers were masked by an "XX" during the rest of the trial.

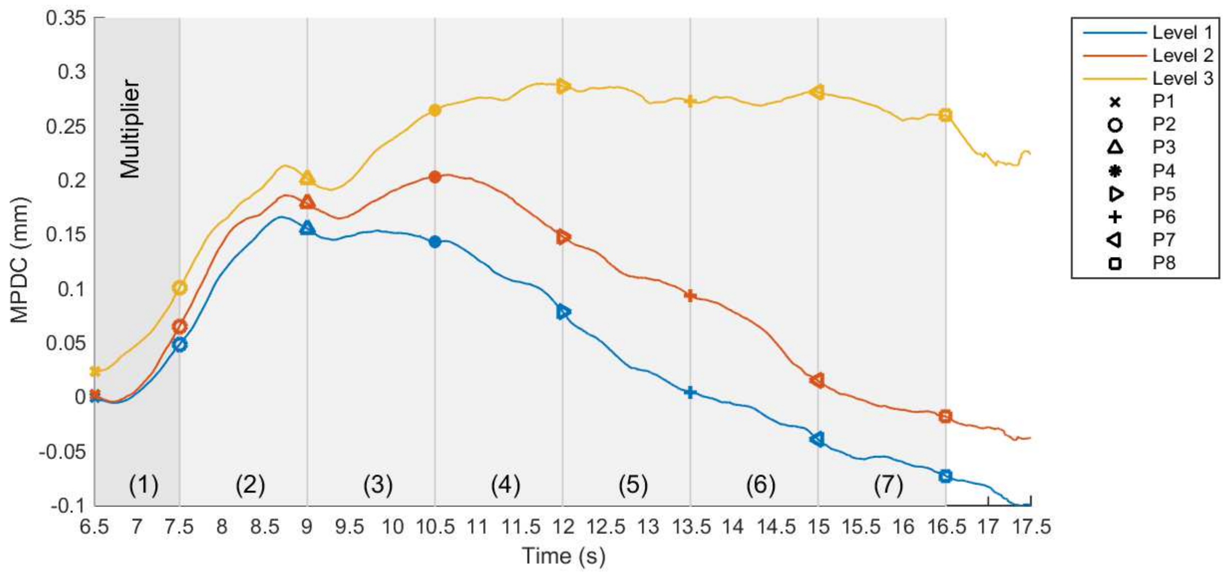

Figure 4b. Mean pupil diameter change (MPDC) during the presentation of the multiplier and the calculation period of 29 participants, for the three levels of difficulty.

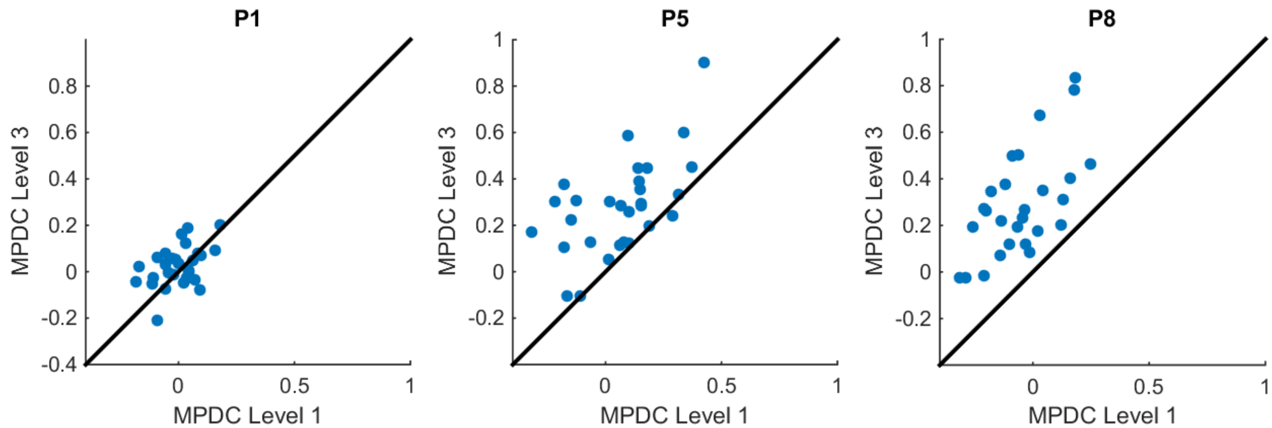

Figure 5. Scatterplot of the mean pupil diameter change (MPDC; blue dots) of 29 participants at point 5 of Levels 1 and 3. Also depicted is the unity line (solid black).

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1105v1 | CC-BY 4.0 Open Access | rec: 21 May 2015, publ: 21 May 2015

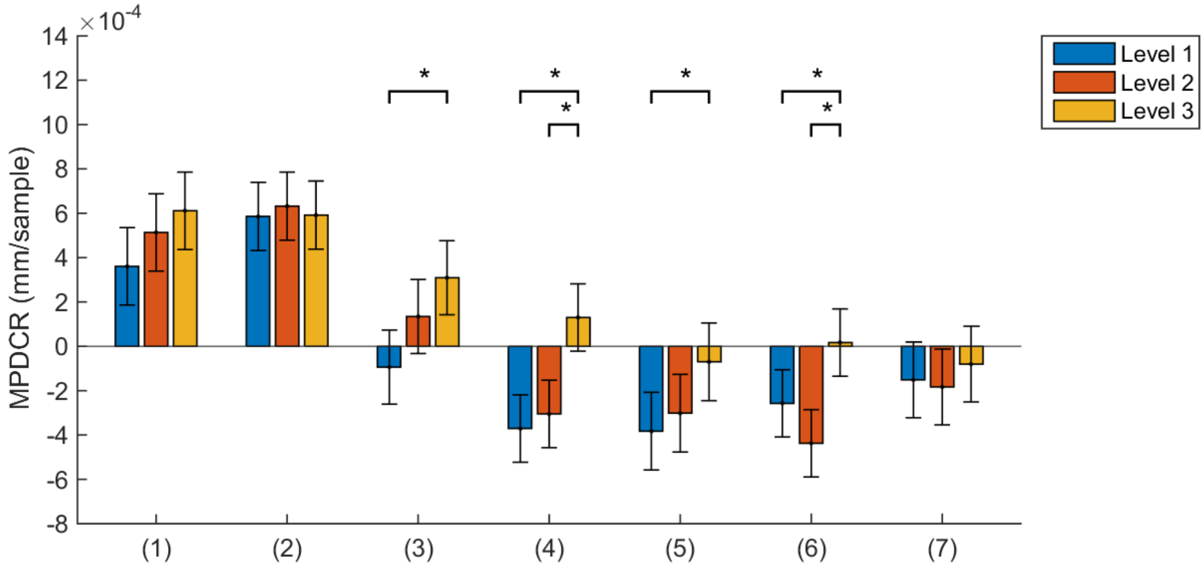

Figure 6. Mean pupil diameter change rate (MPDCR) of 29 participants as a function of difficulty level, for seven periods in time during the presentation of the multiplier and the calculation period. The asterisks indicate significant differences between the levels of difficulty.

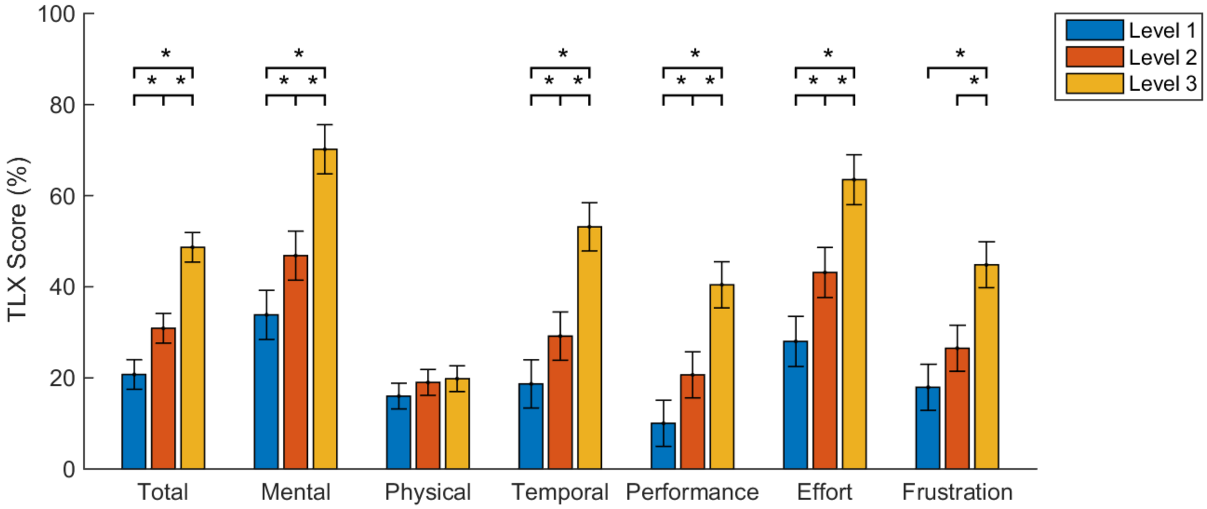

Figure 7. Results of the NASA-TLX questionnaire.

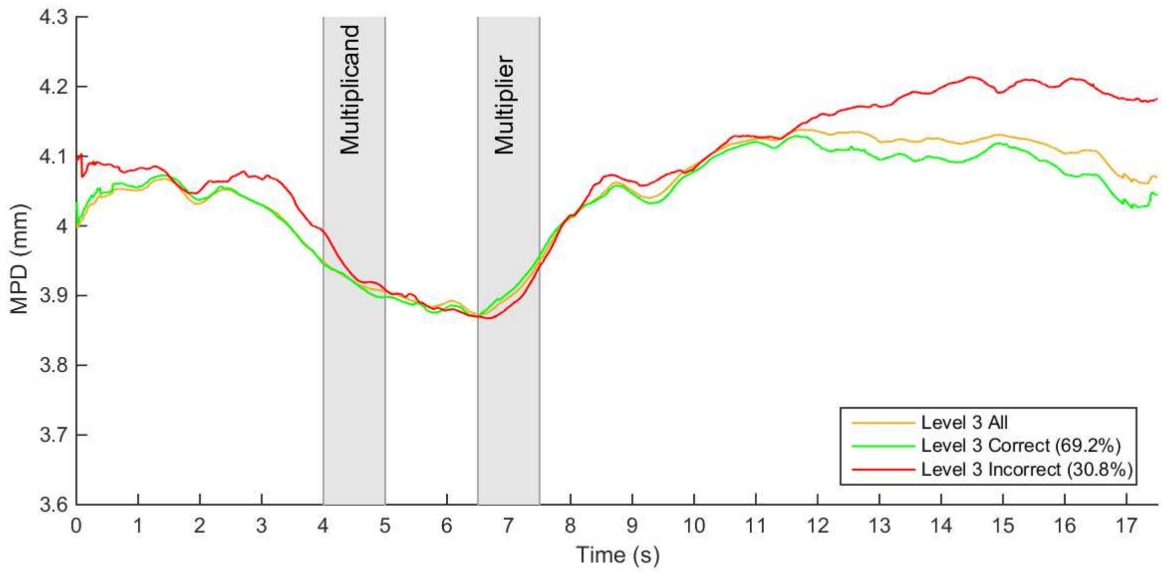

Figure 8. Mean pupil diameter (MPD) during the mental multiplication task of 29 participants for Level 3. The grey bars represent the periods where the multiplicand and multiplier were shown on the screen. The numbers were masked by an "XX" during the rest of the trial.

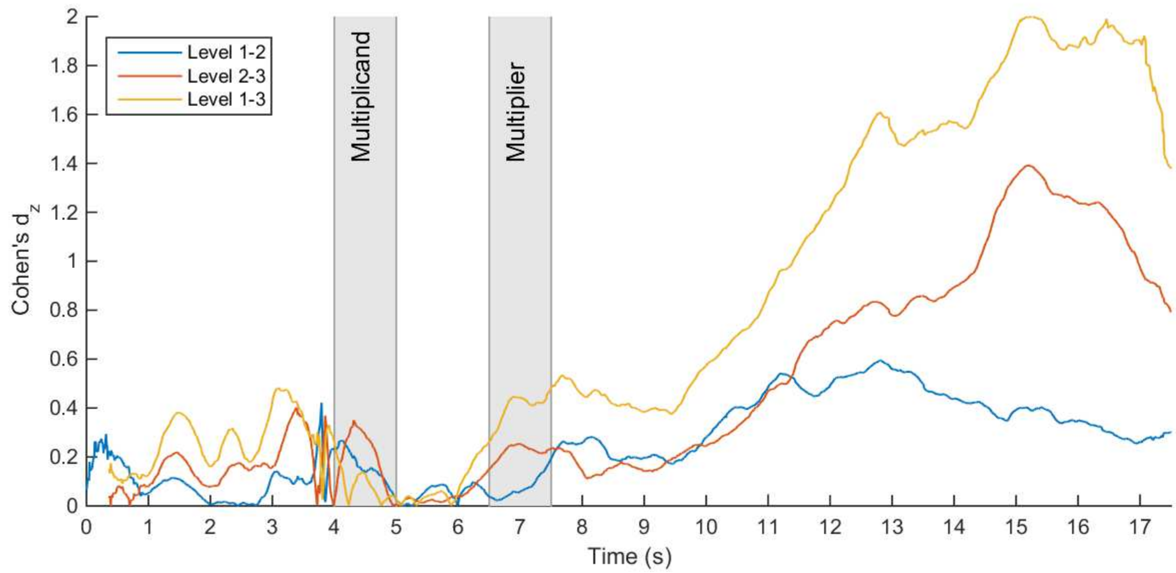

Figure 9. Cohen's d_z for the mean pupil diameter change (MPDC) between the three levels difficulty. The grey bars represent the periods where the multiplicand and multiplier were shown on the screen. The numbers were masked by an "XX" during the rest of the trial.

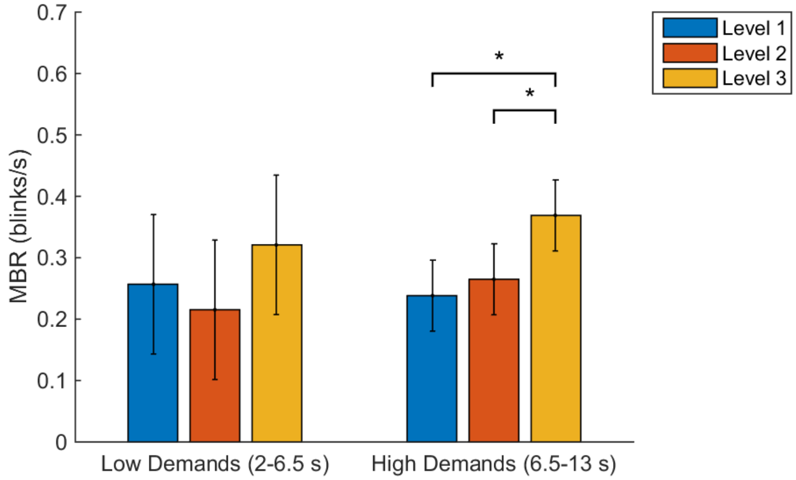

Figure 10. Mean blink rate (MBR) of 30 participants during a period with low and high mental demands, for three levels of difficulty.

Appendix A. Classification of arithmetic tasks.

Three levels of arithmetic task difficulty were used for the full-scale experiment. Each task consisted of calculating the multiplication between two digits ranging from 5 to 18. The tasks were sorted from easy to hard by the outcome of their multiplication. It was assumed that multiplications with a lower outcome were easier than those with a higher outcome. So in this case the easiest task was 5x12 and the hardest was 18x18. The digits 10, 11 were excluded in this method, since they were considered to be too easy. This left 63 possible multiplications, with the assumption that AxB and BxA were equally difficult.

The multiplications were then distributed over three different levels of difficulty (easy, medium and hard), all containing 21 possible multiplications. In order to make a clear distinction between the three levels of difficulty, the first six multiplications were removed from each level. Table A.1 shows the removed and selected multiplications of the three levels. Note that the smallest digit of a pair is put down first, but during the experiment they were presented to the participant in randomized order.

Table A.1

All possible multiplications between 6 and 18 (10, 11 and 15 are excluded), sorted by difficulty and classified into three different levels (Level 1 being the easiest and Level 3 being the hardest).

	Level 1			**Level 2**			**Level 3**	
Removed	5	12		7	16		13	15
	5	13		8	14		14	14
	5	14		9	13		12	17
	6	12		7	17		13	16
	5	15		8	15		14	15
	6	13		7	18		12	18
Selected	5	16		9	14		13	17
	6	14		8	16		14	16
	7	12		9	15		15	15
	5	17		8	17		13	18
	5	18		8	18		14	17
	6	15		9	16		15	16
	7	13		12	12		14	18
	6	16		9	17		15	17
	8	12		12	13		16	16
	7	14		9	18		15	18
	6	17		12	14		16	17
	8	13		13	13		16	18
	7	15		12	15		17	17
	6	18		13	14		17	18
	9	12		12	16		18	18

Appendix B. Blink identification and removal

During a blink, the eyelid opening rapidly diminishes to zero and then increases in a few tenths of a second until it is fully open again (see Fig. B.1, solid blue line). It is impossible to track the pupil's diameter while blinking. These instances in time should therefore be removed from the data. The recordings of the eyelid opening were used to identify the blinks in the pupil diameter data. A threshold of 75% of the mean eyelid opening was used to make a clear distinction between blinks and no blinks as depicted in the figure by the dashed red line.

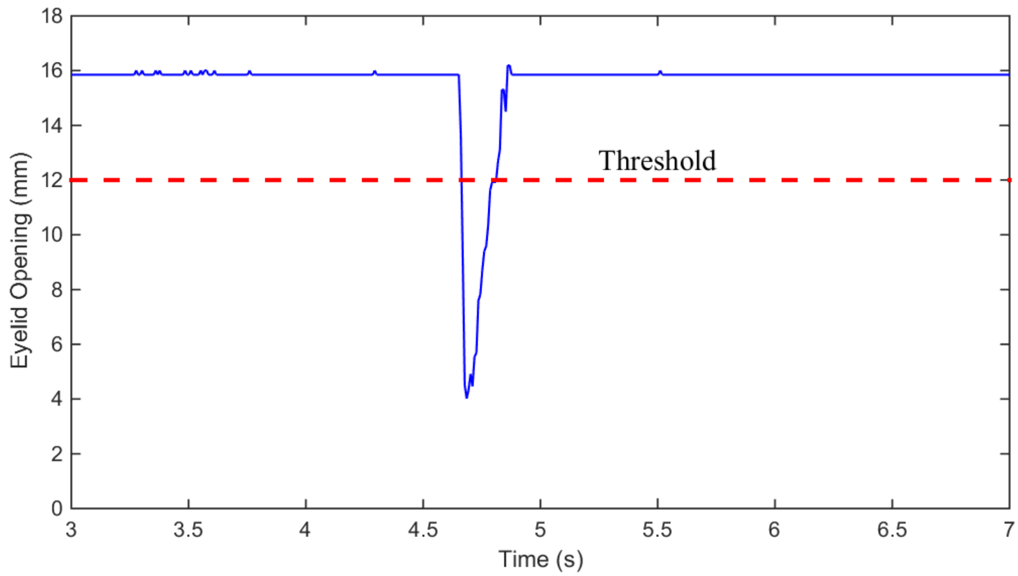

Figure B.1. Sample of the recordings of the eyelid opening showing a typical blink (blue) and the threshold (red) used to identify it.

As can be seen in the figure, it takes some time to cross the threshold and the blink has not been completed after the eyelid opening signal crossed the threshold line for the second time. That is why 12 additional data points (~0.1 s) were removed from the data before the blink and 36 additional data points (~0.3 s) after the blink.

Appendix C. Eight-point analysis of correct and incorrect responses

The results of the eight-point analysis for the correct and incorrect responses of difficulty Level 3 are shown in Table C.1 and Figure C.1.

Table C.1

Mean Pupil Diameter (MPD). The means (M) and standard deviations (SD) of 25 participants are shown for Level 3 of the multiplications, and separated for correct and incorrect responses. P1-P8 refers to the eight points in time.

	Level 3 All	**Level 3 Correct**	**Level 3 Incorrect**	**p-value**	**Effect size**	**Pairwise comparison of conditions**		
	M (SD)	M (SD)	M (SD)		η_p^2 (η_G^2)	1 vs. 2	1 vs. 3	2 vs. 3
MPD (mm) (N = 25)								
P1	3.915 (0.490)	3.919 (0.508)	3.862 (0.494)	0.140	0.08 (0.00)	0.991	0.222	0.178
P2	3.985 (0.516)	4.002 (0.524)	3.933 (0.550)	0.027	0.14 (0.00)	0.803	0.112	**0.027**
P3	4.080 (0.531)	4.079 (0.534)	4.060 (0.566)	0.642	0.02 (0.00)	1.000	0.685	0.703
P4	4.138 (0.522)	4.132 (0.526)	4.116 (0.589)	0.638	0.02 (0.00)	0.975	0.636	0.767
P5	4.157 (0.521)	4.135 (0.534)	4.157 (0.577)	0.662	0.02 (0.00)	0.711	1.000	0.709
P6	4.135 (0.518)	4.109 (0.529)	4.190 (0.599)	0.063	0.11 (0.00)	0.732	0.250	0.056
P7	4.144 (0.500)	4.126 (0.517)	4.197 (0.556)	0.224	0.06 (0.00)	0.906	0.421	0.220
P8	4.125 (0.493)	4.084 (0.516)	4.203 (0.575)	**0.049**	0.12 (0.01)	0.672	0.240	**0.042**

Note. Statistically significant differences are indicated in boldface.

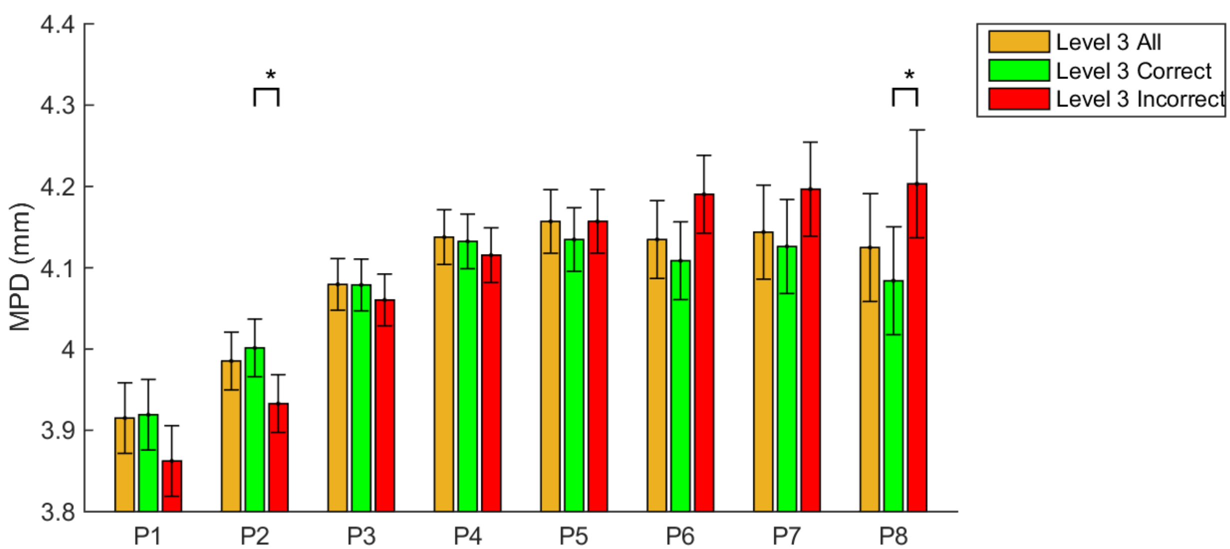

Figure C.1. Mean pupil diameter (MPD) of 25 participants for Level 3, and separated for correct and incorrect responses.