Caffeine effects of a cup of coffee on vigilance and attention in a realistic scenario

Geertje van Bergen[1], Maykel van Miltenburg[2], Ruud van Stiphout[3], Alwin van Drongelen[2], Lea Riesenbeck[2], Jane Sieters[2], Garmt Dijksterhuis[1], Monique Vingerhoeds[1], Esther Aarts[4]

PUBLIC

WAGENINGEN
UNIVERSITY & RESEARCH

Caffeine effects of a cup of coffee on vigilance and attention in a realistic scenario

Authors: Geertje van Bergen[1], Maykel van Miltenburg[2], Ruud van Stiphout[3], Alwin van Drongelen[2], Lea Riesenbeck[2], Jane Sieters[2], Garmt Dijksterhuis[1], Monique Vingerhoeds[1], Esther Aarts[4]

[1] Wageningen Food and Biobased Research, Wageningen University & Research, Wageningen, NL
[2] Department of Safety and Human Performance, Royal Netherlands Aerospace Centre, Amsterdam, NL
[3] imec at OnePlanet Research Center, Wageningen, NL
[4] Centre for Cognitive Neuroimaging, Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, NL

Wageningen Food & Biobased Research
Wageningen, December 2024

Public

Report 2610
DOI: 10.18174/683400

WAGENINGEN
UNIVERSITY & RESEARCH

Version: Final
Reviewer: Jurriaan Mes
Approved by: Mark Bouwens
Carried out by: Wageningen Food and Biobased Research, Royal Netherlands Aerospace Centre, Amsterdam, imec at OnePlanet Research Center, Centre for Cognitive Neuroimaging, and Donders Institute for Brain, Cognition and Behaviour
Subsidised and commissioned by: the Dutch Ministry of Agriculture, Fisheries, Food Security and Nature
This report is: Public

The research that is documented in this report was conducted in an objective way by researchers who act impartial with respect to the client(s) and sponsor(s). This report can be downloaded for free at https://doi.org/10.18174/683400 or at www.wur.eu/wfbr (under publications).

The research leading to this report received funding from the Ministry of Agriculture, Nature and Food Quality and a consortium of partners (DSM Nutritional Products Ltd., Thales Nederland B.V., De Staat der Nederlanden, represented by the Minister of Defense, Royal Netherlands Aerospace Centre-NLR, TimeTools, Koninklijke Douwe Egberts B.V., Stichting IMEC Nederland and Radboud University Nijmegen) under Grant Agreement DFI-AF-19010 Food for vigilance.

PO box 17, 6700 AA Wageningen, The Netherlands, T + 31 (0)317 48 00 84, E info.wfbr@wur.nl, www.wur.eu/wfbr.

Contents

Preface

A growing number of professionals work outside regular office hours, many of them in a type of job that brings psychological or physical stress while requiring optimal alertness. These professionals are found in air traffic control, transportation, airplane piloting, emergency work, etc., both in civil and military contexts. Even a tiny lapse in their alertness can carry large risks for themselves and others. There is a well-known link between the amount and type of food ingested and one's level of alertness, but most evidence is anecdotal or based on laboratory studies that may not allow translation to real working circumstances. In the project Food for Vigilance, a consortium of societal parties teamed up with knowledge partners to study the link between food and alertness in real working circumstances (https://www.wur.nl/nl/onderzoek-resultaten/kennisonline-onderzoeksprojecten-lvvn/kennisonline/food-for-vigilance.htm).

Vigilance is typically assessed through performance on a simple behavioural task, such as the Psychomotor Vigilance Test (PVT), while selective attention is measured through the oddball paradigm. These tests are widely used for their simplicity and predictive ability, but they are criticized for lacking ecological validity. While flying an aircraft can be monotonous for pilots, it requires more than detecting visual or auditory targets in isolation, as the cockpit is a dynamic environment. This complex, multi-task setting is simulated in the NASA Multi-Attribute Task Battery (MATB), which replicates various tasks pilots perform during flight, including planning, monitoring, and vigilance.

This report describes a study aimed to assess the suitability of MATB-II for measuring the effects of interventions that enhance vigilance and attention in professionals, specifically civil pilots. To achieve this, the study compared the effects of caffeine, administered through two common beverages (regular vs. decaffeinated coffee), on MATB-II performance with their effects on simpler, lab-based equivalents.

Acknowledgements

The authors like to thank Meeke Ummels and Floris van den Oever for their help in recruiting participants, and Alexander van Eekelen for helping with selecting the most appropriate questionnaires.

Summary

Professionally realistic multi-task environments, such as the NASA Multi-Attribute Task Battery (MATB-II), are employed for measuring vigilance and attention in (military) aviation professionals. However, it is unclear whether well-known intervention effects on performance during simple lab-based tasks, such as those of caffeine, translate to these more realistic working situations.

In a preregistered, double-blind, randomized, controlled repeated-measures experiment (https://osf.io/2zubx), we compared the performance of thirty-five civil pilots during vigilance- and attention-related tasks in simple (psychomotor vigilance task; auditory oddball detection) versus multitask environments (MATB-II system monitoring; MATB-II communications) after consuming regular vs. decaffeinated coffee (respectively, containing ~98 vs. ~5 mg of caffeine per 125 ml cup).

For vigilance tasks, no coffee intervention effects were found. Instead, a reversed task repetition effect was found, with participants being slower in session 2 in the simple task environment, but faster in session 2 in the complex environment. For attention-related tasks, regular coffee improved performance accuracy in the simple, but not the multitask environment. Coffee versus decaf effects in the simple task environment did not correlate with those in the complex task environment, neither for vigilance nor for selective attention. However, an experiment-wide increase in sleepiness was attenuated if participants drank regular coffee in the second session. This finding was supported by heart rate and eye blink measures.

Results suggest that intervention-related findings do not easily translate to different vigilance- and attention-related tasks if task environments differ in complexity. The MATB-II multi-task environment, in its current form, is perhaps more suitable for assessing intervention effects on physiological measures of fatigue and vigilance than on cognitive performance.

Keywords

Vigilance (sustained attention); Attentional processes; Cognitive task analysis, Physiological measurement; Fatigue

1 Introduction

A growing number of professionals work outside regular office hours, in jobs that require optimal vigilance (i.e., the ability to remain attentive for sustained periods of time) and selective attention (i.e., the ability to focus on specific input while ignoring irrelevant information). Airline pilots are highly susceptible to vigilance decrements during flight operations, with potentially devastating consequences; maintaining and improving vigilance and attention in this group of professionals is hence crucial to reduce safety risks (e.g., Casagrande, 2011; Federal Aviation Administration, 2010).

A well-known vigilance-enhancing strategy is caffeine intake. Caffeine is among the most consumed psychoactive substances in the world, predominantly in the form of coffee and tea (Barone & Roberts, 1996; van Dam et al., 2020). Prior research using has extensively shown that caffeine has beneficial effects on vigilance (e.g., Baker & Theologus, 1972; Cooper et al., 2021) and selective attention (e.g., Lorist et al., 1994, 1996; Pan et al., 2000; for reviews, see, e.g., Rogers & Smith, 2011; Einöther & Giesbrecht, 2012; McLellan et al., 2016).

Vigilance is typically measured as performance on a simple behavioural task, i.e., the Psychomotor Vigilance Test (PVT; Dinges & Powell, 1985), where participants respond as fast as possible to a series of stimuli appearing on a screen at random time intervals. Vigilance decrements are reflected in slower responses; this effect enhances as the duration of the PVT increases (a.k.a. time-on-task effects; Lim & Dinges, 2008; Basner & Dinges, 2011). The PVT is specifically suitable for measuring effects of sleep deprivation, but it has also been used to measure vigilance-enhancing intervention effects in well-rested subjects (e.g., Fine et al., 1994; Lanini et al. 2016; Cooper et al., 2021).

Selective attention is typically measured as performance on another simple task, i.e., the oddball paradigm (Squires et al., 1975), where participants respond to infrequently occurring target stimuli while ignoring or differentially responding to frequently occurring distractors. Selective attention capacity is reflected in performance accuracy (false alarms and/or misses) and/or reaction times. This classic paradigm has been widely used in psychology research, especially in combination with electrophysiological measurements (Hermann & Knight, 2001).

Their simplicity and predictive power explain and validate the widespread use of these classic cognitive tests, yet their ecological validity can be, and has been, criticised (e.g., Kibler, 1965; Adams, 1987; Koelega, 1993; Al-Shargie, 2019). Although flying an aircraft may, at times, be unexciting for pilots, it requires more than readily detecting visual targets on an empty screen or auditory targets in a quiet room, as a cockpit is a much more stimulating environment.

Such a complex multi-task environment is mimicked in the NASA Multi-Attribute Task Battery (MATB, Comstock & Arnegard, 1992; revised version MATB-II: Santiago-Espada et al., 2011), a computer-based-performance task specifically designed to evaluate operator performance. MATB incorporates various tasks and activities that aviators perform in flight (such as planning, monitoring and vigilance). MATB allows for inducing controlled variation in the simulated cockpit environment (e.g., by systematically varying levels of (sub)task complexity and/or automatization), making it a highly valuable experimental platform. It has been widely used to investigate effects of fatigue, sleep deprivation and workload on operator performance (E.g., Molloy & Parasuraman, 1996; Caldwell et al., 2004; Smith & Gevins, 2005; Valk & Simons, 2008; Carlozzi et al., 2010).

Two of the subtasks incorporated in MATB are equivalents of the PVT and oddball task. The system monitoring subtask (SYSMON) requires participants to detect and respond to warning lights appearing in the display at random time intervals (similar to the PVT). The communications subtask (COMM) requires participants to selectively listen for and respond to radio messages directed at their aircraft (i.e., oddball paradigm). The crucial difference with their simple equivalents is that SYSMON and COMM are embedded in a combination of simultaneously executed tasks in the same display, thus providing an ecologically more valid way of assessing vigilance and attention performance under professional working conditions (see also Navarro et al., 2018).

Although MATB SYSMON and COMM provide outcome measures that equal their simple lab-based counterparts (e.g., Kong et al., 2022), we are unaware of studies using MATB to measure vigilance and

attention-enhancing intervention effects (but see Kourtido-Papadeli et al., 2002; Papadelis et al., 2003 who measured caffeine effects on performance on a different MATB subtask).

Aim

The aim of the current study is to fill this gap and investigate the suitability of MATB-II for measuring vigilance- and attention-enhancing intervention effects in professionals (i.e., civil pilots). To this end, effects of caffeine, which was manipulated as a component of two common beverages (regular vs. decaffeinated coffee), on MATB-II SYSMON and COMM performance were directly compared with their simple lab-based equivalents (PVT and auditory oddball detection, AOD, respectively).

If caffeine effects on MATB-II SYSMON and COMM performance translate well to the better known simple lab-based performance measures, we hypothesize (a) the (known) positive effect of caffeine on PVT performance (i.e., faster responses) to be positively related to its effect on MATB-II SYSMON performance, and (b) the (known) positive effect of caffeine on oddball task performance to be positively related to its effect on MATB-II COMM performance (i.e., higher accuracy).

In addition, we explored ways of successfully assessing effects of vigilance-enhancing interventions in professionals beyond behavioural performance measures. To this end, intervention effects were also assessed with physiological measures (blink behaviour and heart rate) throughout the experiment.

The research leading to this report received funding from the Ministry of Agriculture, Nature and Food Quality and a consortium of partners (DSM Nutritional Products Ltd., Thales Nederland B.V., De Staat der Nederlanden, represented by the Minister of Defense, Royal Netherlands Aerospace Centre-NLR, TimeTools, Koninklijke Douwe Egberts B.V., Stichting IMEC Nederland and Radboud University Nijmegen) under Grant Agreement DFI-AF-19010 Food for vigilance (https://www.wur.nl/nl/onderzoek-resultaten/kennisonline-onderzoeksprojecten-lvvn/kennisonline/food-for-vigilance.htm).

2 Method

2.1 Participants

36 civil pilots (in training) participated in the study. Participants were recruited through an email campaign amongst several Netherlands-based flight academies; inclusion criteria were (a) being a civil pilot or following education to become a civil pilot in the practical phase, (b) aged 18 - 40 years, (c) non-smoking, and (d) being a habitual coffee consumer (1-6 cups per day). One participant attended only one session, leaving 35 participants (2 female, mean age 28.6 years, SD 5.4, range 18-39) for analysis. One participant responded to the wrong target sound in the auditory oddball detection task; this person was excluded from analyses of this task only. A summary of participant characteristics (full sample and per treatment order group) is provided in Table 1.
Informed consent was obtained from each participant, and all received a monetary compensation for their participation (40 euros). The study complied with the tenets of the Declaration of Helsinki and was approved by the Institutional Review Board at Wageningen University & Research.

2.2 Investigational product

To be close to realistic behaviour having one cup of coffee before starting an activity, the investigational products were regular coffee (Senseo Strong coffee pads, Douwe Egberts, JDE, the Netherlands), containing ~98 mg of caffeine per 125 ml cup, and decaffeinated coffee (Senseo Decaf coffee pads, same company), containing ~5 mg of caffeine per 125 ml cup. Coffee pads were individually packed and coded by the coffee company. Products were prepared in a Senseo coffee machine (Philips) and served immediately without milk, sugar or sweeteners. The coffee machine was rinsed after each use to minimize potential caffeine spill-over between sessions.

Table 1 ***Sample characteristics: demographics, self-reported trait and state variables of the full sample and per treatment order group (regular coffee first or decaffeinated coffee first).***

	Full sample N=35	**Coffee-first N=17**	**Decaf-first N=18**
N(%) female	2 (6%)	1 (6%)	1 (6%)
Age (M±SD)	28.6 ± 5.4	27.4 ± 4.9	29.8 ± 5.6
Cockpit experience (in years, M±SD)	6.28 ± 5.2	5.1 ± 3.9	7.4 ± 6.1
TRAIT VARIABLES			
Chronotype			
Moderate morning person	6 (17%)	3 (18%)	3 (17%)
None of both	24 (69%)	11 (65%)	11 (72%)
Moderate evening person	5 (14%)	3 (18%)	2 (11%)
Habitual caffeine consumption (mg/day)	174 ± 75	170 ± 67	179 ± 83
Caffeine sensitivity			
High	1 (3%)	0 (0%)	1 (6%)
Moderate	21 (60%)	9 (53%)	12 (67%)
Low	12 (34%)	7 (41%)	5 (28%)
(unknown)	1 (3%)	1 (6%)	0 (0%)
General impression sleep quality			
Good	34 (97%)	16 (94%)	18 (100%)
Bad	1 (3%)	1 (6%)	0 (0%)
General impression sleep duration			
Good	25 (71%)	12 (71%)	13 (72%)
Bad	10 (29%)	5 (29%)	5 (28%)

STATE VARIABLES			
Sleep quality past 24 hrs (0-100)			
Session 1	75.7 ± 10.9	73.2 ± 7.3	78.1 ± 13.3
Session 2	74.5 ± 12.2	74.5 ± 11.0	74.4 ± 13.6
Sleep duration past 24 hrs (hrs)			
Session 1	7.6 ± 1.1	7.5 ± 1.1	7.6 ± 1.1
Session 2	7.3 ± 1.3	7.5 ± 1.5	7.0 ± 1.1

2.3 Tasks

Fig1 shows an example display of the multitask environment (MATB-II) with the vigilance-related (SYSMON) and attention-related (COMM) subtasks outlined, together with example displays of their simple lab-based equivalents.

Vigilance: MATB-II system monitoring (SYSMON). Participants had to respond as fast as possible to the appearance of a red warning light in the upper left corner of the MATB-II display at random time intervals by pressing F6 on the keyboard. No responses to other system monitoring events (i.e., green warning lights, scales) were required. The 30-minutes MATB-II test included 120 SYSMON events in total.

Vigilance: Psychomotor Vigilance test (PVT). Participants had to respond to a red square appearing in the centre of a computer screen at random time intervals between 1 and 4 seconds by pressing the space bar on the keyboard. There were 150 trials in total. It took on average 8 minutes to complete the PVT.

Selective attention: MATB-II communications (COMM). Participants had to listen for and respond to call signs directed at their own aircraft ("NASA504") by adjusting the radio channel and radio frequencies according to the instructions in the message using the arrow keys on the keyboard; irrelevant call signs (messages directed at four other aircrafts: "CITRIS", "ACEY", "DELTA", "FEDEX") had to be ignored. The 30-minutes MATB-II test included 66 COMM events, of which 22 relevant and 44 irrelevant call signs (ratio 1:2), occurring at random time intervals.

Selective attention: Auditory oddball detection (AOD). Participants responded as fast as possible to an infrequently occurring target sound (e.g., "three"; 33% of trials) by pressing the left arrow key on the keyboard, and to frequently occurring distractor sounds (four other numbers, e.g., "one", "two", "four", "five"; 67% of trials) by pressing the right arrow key. There were 300 trials in total (divided into 3 blocks with self-timed breaks in between); inter-trial intervals randomly varied between 1 and 2 seconds. Completing the AOD task took 10 minutes on average.

To mimic a realistic working environment, the MATB-II test additionally included a tracking subtask (in which participants had to keep the aircraft inside a central square with a joystick; top centre) and a resource management subtask (where participants had to manage fuel levels in tanks by (de)activating pumps from other tanks; bottom centre); performance on these subtasks was not analysed. Settings for all MATB-II subtasks (i.e., number of events, response options) were kept constant across participants and sessions.

The MATB-II event script was programmed using XML language. PVT and AOD were administered using Presentation® software (version 18.0, Neurobehavioral Systems, Inc., Berkeley, CA). All tests were administered on a standard Windows 10 laptop with a 24-inch LED-backlit LCD monitor. Participants wore over-ear headphones and used a standard QWERTY USB keyboard, an optical mouse and a joystick to provide responses during the (sub)tasks.

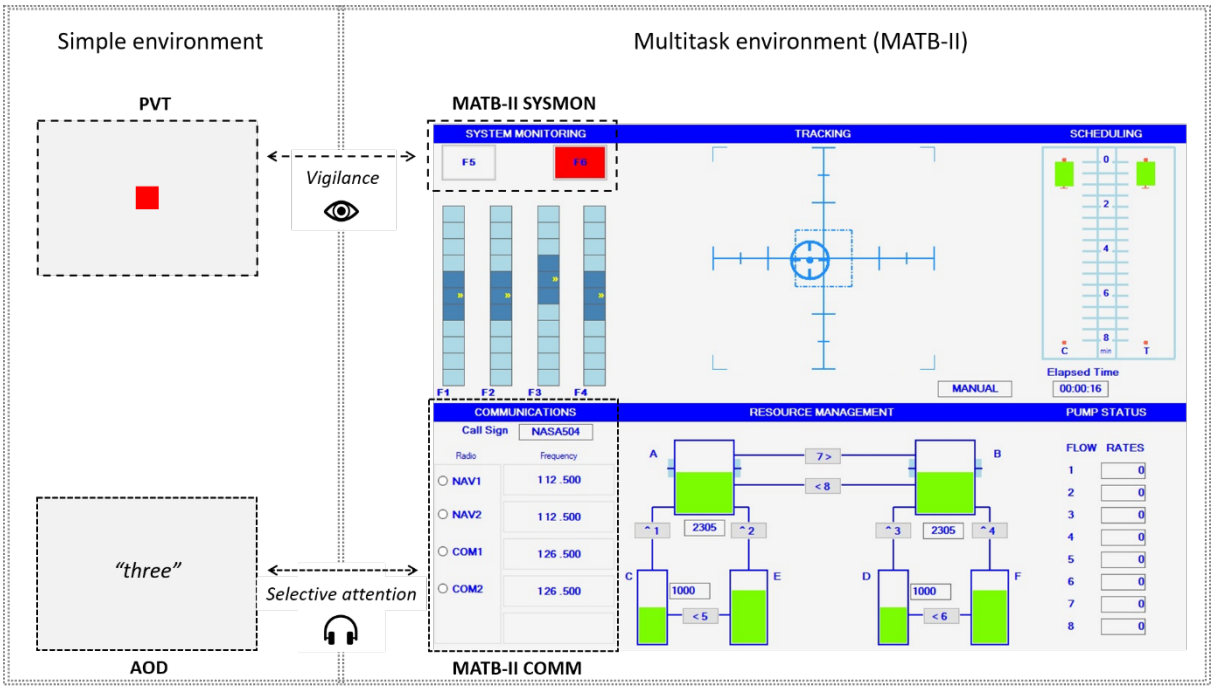

Figure 1 ***Example displays of the multitask environment (MATB-II; right) and the simple task environments (left). For vigilance, performance on the MATB-II system monitoring subtask (SYSMON) was compared with performance on the psychomotor vigilance test (PVT); for selective attention, performance on the MATB-II communications subtask (COMM) was compared with performance on the auditory oddball detection task (AOD)***

2.4 Design and procedure

The study followed a double-blind, randomized, controlled, repeated-measures design. Participants attended two experimental sessions (which were minimally one week and maximally two weeks apart) at the Royal Netherlands Aerospace Centre (NLR) in Amsterdam, the Netherlands, in February and March, 2021. All sessions were scheduled in the morning between 8.30 and 11.30am; scheduled times were the same for both sessions for all but two participants (due to conflicting calendars - for these participants, sessions were 90 minutes apart). Participants were instructed to abstain from coffee, tea and other caffeinated products (e.g., energy drinks) ≥ 12 hours, from alcohol ≥ 24 hours, and from (non-prescribed) drugs ≥ 48 hours prior to the session, and to consume their usual breakfast prior to both sessions.

A schematic representation of a full experimental session is provided in Figure 2. At the start of each session, participants consumed the investigational product (either regular coffee or decaf; treatment order counterbalanced across participants), and a smartwatch (Empatica E4) was attached around their wrist to monitor their heart rate continuously throughout the session. An ingestion period of 30 minutes was included to ensure that caffeine would be (almost) completely absorbed during performance of the first task (which is around 45 minutes after oral ingestion in adults, with peak plasma levels ranging from 15-120 minutes; Fredholm et al., 1999; Nehlig, 2018), and that all computer tasks would be executed well before the assumed half-life of caffeine in the human body (approximately 4 hours; Nehlig, 2018). During this ingestion period, participants performed a 10-minute practice version of the MATB-II test to familiarize themselves with (session 1) or refresh their memory of (session 2) the multitask environment, and answered a number of questions assessing trait and state variables. After the ingestion period, they performed the three computer tasks consecutively (MATB-II – PVT – AOD; task order was kept constant across sessions and participants to minimize between-subjects differences) during which their blinking behaviour was registered via a Tobii Eye Tracker 4C at a 90 Hz sampling rate. After the MATB-II test, participants indicated their experienced workload by filling out the NASA Task Load Index (NASA-TLX; Hart et al., 1988); after the last

task, they answered an additional number of questions assessing subjective task performance and product liking. At five time points during the session, participants indicated their level of sleepiness on a 9-point scale (Karolinska Sleepiness Scale [KSS]; Akerstedt & Gillberg, 1990). At the end of the second session, participants were debriefed about the goal of the experiment, after which they were asked to guess at which session they had consumed decaf and answered a number of additional questions assessing caffeine-related traits. A full experimental session took 90 minutes on average.

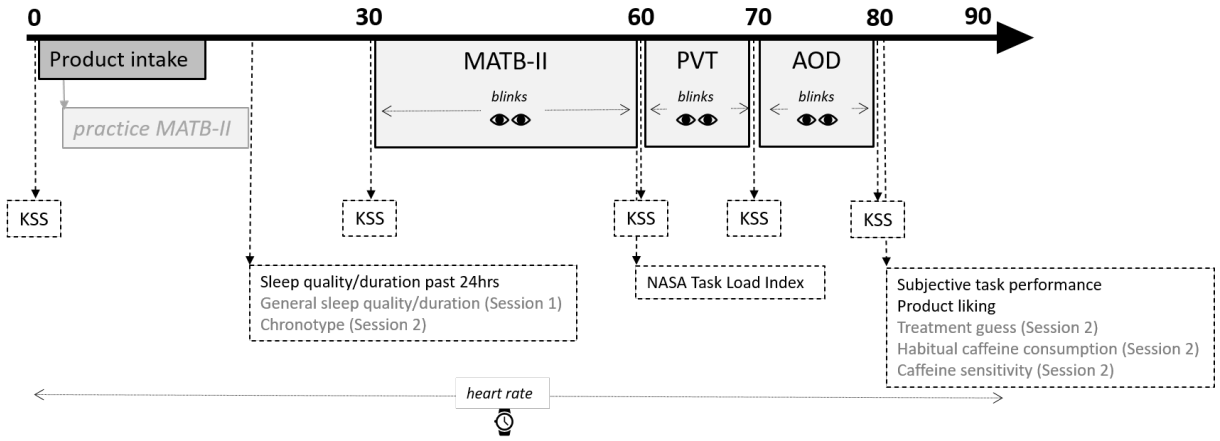

Figure 2 ***Timeline of each experimental session (in minutes). Light grey boxes represent the computer tests. Dashed boxes represent self-report measures (measures assessed in both sessions in black, measures assessed in one session in grey). KSS: Karolinska sleepiness scale. Horizontal arrows represent time intervals of blink (obtained via eye-tracker) and heart rate (obtained via smartwatch) measurements.***

2.5 Data analysis

Data pre-processing, analysis and visualization were performed in R version 4.0.5 (R core team, 2021) and RStudio version 1.4 (RStudio Team, 2020).

2.5.1 Behavioural performance measures

Preregistered analyses. Initial analyses followed a preregistered data analysis plan (doi: 10.17605/OSF.IO/2ZUBX: https://osf.io/2zubx). For PVT and MATB-II SYSMON, RTs were selected as primary outcome measure. For each participant, RTs were aggregated per coffee condition. Miss trials, defined as trials in which no response was given before the next stimulus appeared (PVT) or within 15 seconds (SYSMON), and commission errors (RTs <100 ms) were excluded (PVT: 0.23% data loss; SYSMON: 0.06% data loss). For MATB-II COMM and AOD, response accuracy was selected as primary outcome measure, from which a discriminability measure (d') was derived. Hit rates were calculated by dividing the number of correct responses to target trials (or relevant call signs) by the total number of target trials; false alarm rates were calculated by dividing the number of incorrect responses to distractor trials (or irrelevant call signs) by the total number of distractor trials. d' was calculated as the difference between z-transformed hit rates and false alarm rates. For each participant, a d' was calculated per coffee condition.
First, as a manipulation check, mean RTs and d' were compared between coffee conditions using paired samples t-tests. Next, difference scores were calculated by subtracting mean RTs/d' in the regular coffee condition from mean RTs/d' in the decaf condition (averaged over treatment orders). Condition difference scores were compared between vigilance-related tasks on the one hand, and attention-related tasks on the other, using Spearman's correlation analyses (as the data was not normally distributed, Shapiro-Wilk tests, p's<.001). The standard p<.05 criterion was used to determine significance.
Exploratory non-preregistered analyses. Besides the preregistered analyses, mean RTs and d' were compared between tests sessions using paired samples t-tests. Differences across treatment order groups were assessed by means of 2 (session 1 vs. session 2; within-subjects) x 2 (regular-first vs. decaf-first, between-subjects) ANOVAs.

2.5.2 Physiological measures

Blinks were detected by identifying gaps without gaze data from the Tobii 4C that were in the expected range of 75-300 ms long. Number of blinks, mean blink duration and standard deviation of blink duration were calculated in 1-minute non-overlapping time windows. The averages of each blink variable for the duration of each task were compared. The Empatica E4 recorded mean heart rate in non-overlapping 1-minute windows and was used to assess physiological arousal. Besides averages for the duration of each task, mean heart rate was compared also at time window 0-10 minutes after coffee consumption and in the last 10 minutes before starting MATB-II. Differences between sessions and treatment order groups were assessed using repeated-measures ANOVAs.

2.5.3 Additional measures

Self-reported task-related and product-related measures were compared across sessions and treatment order groups using repeated-measures ANOVAs. For KSS ratings, ANOVAs additionally included Time (5 levels) as a within-subjects factor.
Participant demographics, trait and state variables were compared between treatment order groups (regular coffee first or decaf coffee first) using independent t-tests. Because of the interactions with Treatment Order and the small number of participants per treatment order group, potential mediating effects of individual trait and state variables on task performance (as proposed in the preregistration as exploratory analyses) were not performed.

2.5.4 Sample size calculation

A power calculation administered in G*Power revealed that a sample size of 31 participants suffices to detect a correlation of r=0.5 at an alpha of 0.05 with 90% power in a one-tailed bivariate normal model. Moreover, with a sample size of 31, our study is sufficiently powered (90%) to detect a moderate within-subject effect of coffee vs. decaf (Cohen's d z = 0.6) using a two-tailed, paired-samples t-test; prior studies investigating effects of caffeine (vs. placebo) on vigilance in healthy, rested volunteers have reported similar effect sizes (e.g., Fine et al. 1994: d=0.6; Lanini et al. 2015: Hedge's g = 0.59). To compensate for potential drop-outs and/or technical issues, we recruited 36 participants.

3 Results

3.1 Vigilance: PVT vs. MATB-II SYSMON

The results from the preregistered analyses provided no evidence for an effect of regular (vs. decaf) coffee on mean RTs in either task environment (PVT: difference 1.4ms, t(34)=0.36, p =.721; SYSMON: difference -30ms, t(34)=-1.32, p=.196). Also, no evidence was found for a correlation of treatment difference scores between task environments (ρ=-0.11, p=.516). However, (exploratory non-preregistered) analyses showed that mean RTs significantly differed across tests sessions, both for PVT (t(34)=3.03, p=.004, d_z=0.39) and SYSMON (t(34)=2.78, p=.009, d_z=0.36). For the PVT, mean RTs increased in the second relative to the first session (difference 10.5 ms), indicating that participants were slower when performing the PVT for the second time. For SYSMON, the session effect went in the opposite direction: mean RTs decreased in the second vs. first session (difference -58.6 ms), i.e., participants were faster when performing SYSMON for the second time. No evidence was found that session effects differed across treatment orders (i.e., regular-first vs. decaf-first, PVT: p=.755; SYSMON: p=.182), suggesting that test-retest effects occurred irrespective of when participants drank regular or decaf coffee (see Fig3).

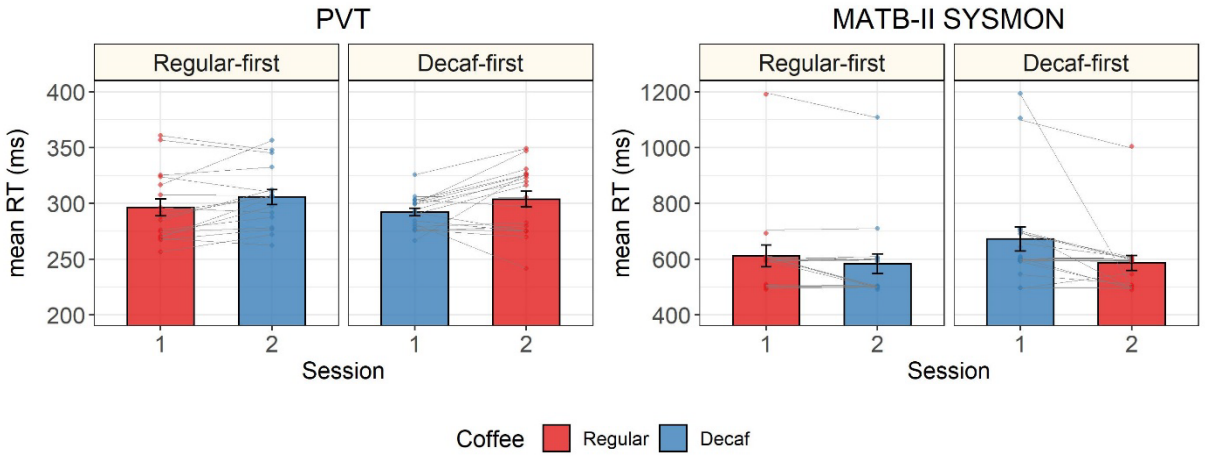

Figure 3 ***Mean RTs across test sessions per treatment order group for PVT (left) and MATB-II SYSMON (right). Red bars indicate regular coffee sessions; blue bars indicate decaf sessions. Grey lines represent individuals; error bars show ±SEM***

3.2 Selective attention: AOD vs. MATB-II COMM

Performance accuracy was high for both selective attention tasks. For AOD, participants correctly responded to target sounds (hits) in 94.3%, and incorrectly responded to distractor sounds (false alarms) in 0.8% of the cases, yielding an average d' of 4.09. A (two-tailed) paired-samples t-test showed a trend in difference between regular coffee and decaf in the hypothesized direction: performance improved after having consumed regular (d'=4.18) vs. decaf coffee (d'=4.01), difference 0.17, t(33)=1.80, p=.082, d_z=0.30 (Fig4, left). No evidence was found that this effect differed across treatment order groups (p=.75). For MATB-II COMM, performance was virtually at ceiling: participants almost always correctly responded to relevant call signs (99.4%) and hardly responded to messages directed at other aircrafts (0.03%), yielding an average d' of 4.23. Paired-samples t-tests provided no evidence for a difference in discriminability between the regular coffee (d'=4.23) and decaf condition (d'=4.22), difference 0.01, t(34)=0.14, p=.89 (Fig4, right). No evidence was found that coffee effects on MATB-II COMM performance differed between treatment orders (p=.42), nor for a correlation of d' difference scores between task environments (ρ=-0.02, p=.92).

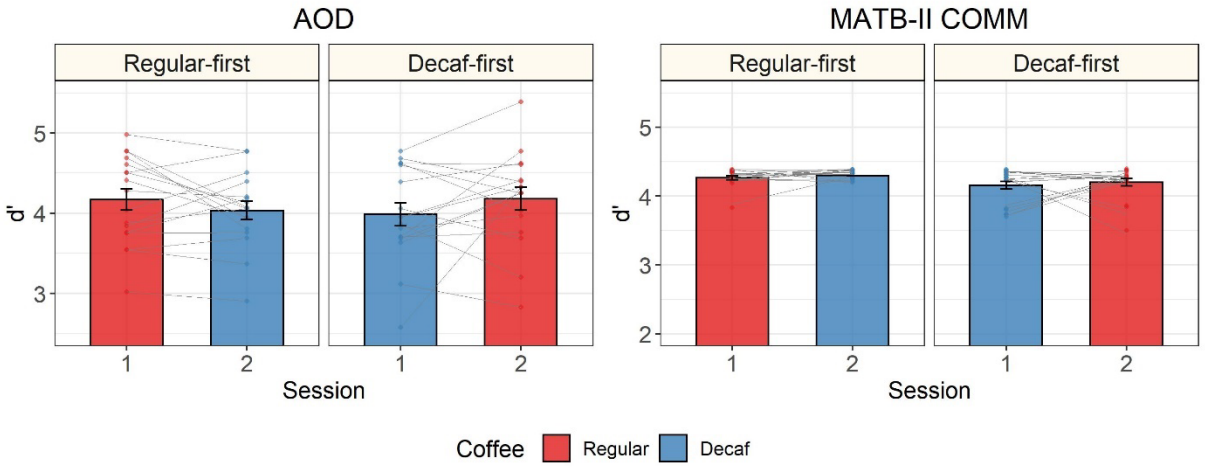

Figure 4 ***Discriminability (d') scores in regular (red) vs. decaf coffee (blue) sessions per treatment order group for MATB-II COMM (left) and AOD (right). Lines represent individuals; error bars show ±SEM***

Together, findings suggest that regular (vs. decaf) coffee improved attention-related performance only in the simple task environment (AOD), possibly due to the ceiling effect observed for MATB-II COMM. Results moreover suggest that selective attention-related task performance is less susceptible to test-retest effects than vigilance-related task performance.

3.3 Sleepiness/arousal

Results from the analyses of KSS ratings, blink behaviour and heart rate are summarized in Table 2. As shown in Fig5 (top panel), KSS ratings gradually increased as the experimental sessions progressed, but a significant Time x Session x Treatment Order effect indicated that sleepiness patterns over time differed across sessions and treatment order groups. In the regular-first group, participants felt sleepier in the second (decaf) session relative to the first (regular coffee) session; KSS ratings differed at all time points (p's<.021), except for T0 (i.e., prior to product intake, $p=1$). For the decaf-first group, sleepiness patterns over time did not differ between sessions. This self-reported sleepiness pattern was confirmed by an objective sleepiness measurement: mean blink duration increased over the course of the experiment, and a difference between sessions (with overall longer blinks in Session 2) was again restricted to the regular-first group (Fig5, middle panel). Analyses of blink rate and blink duration variability showed similar patterns (results are provided in the Supplement). An experiment-wide repetition effect was also found in a physiological measurement of arousal: heart rate was reduced in the second relative to the first session. Fig5 (bottom panel) suggests that the session difference was again restricted to the regular-first group, but this could not be statistically confirmed. Taken together, results from these additional measures show that participants were more sleepy and less aroused during the second test session, but that this test-retest effect was attenuated in the decaf-first group, suggesting that regular coffee counteracted an experiment-wide boredom effect.

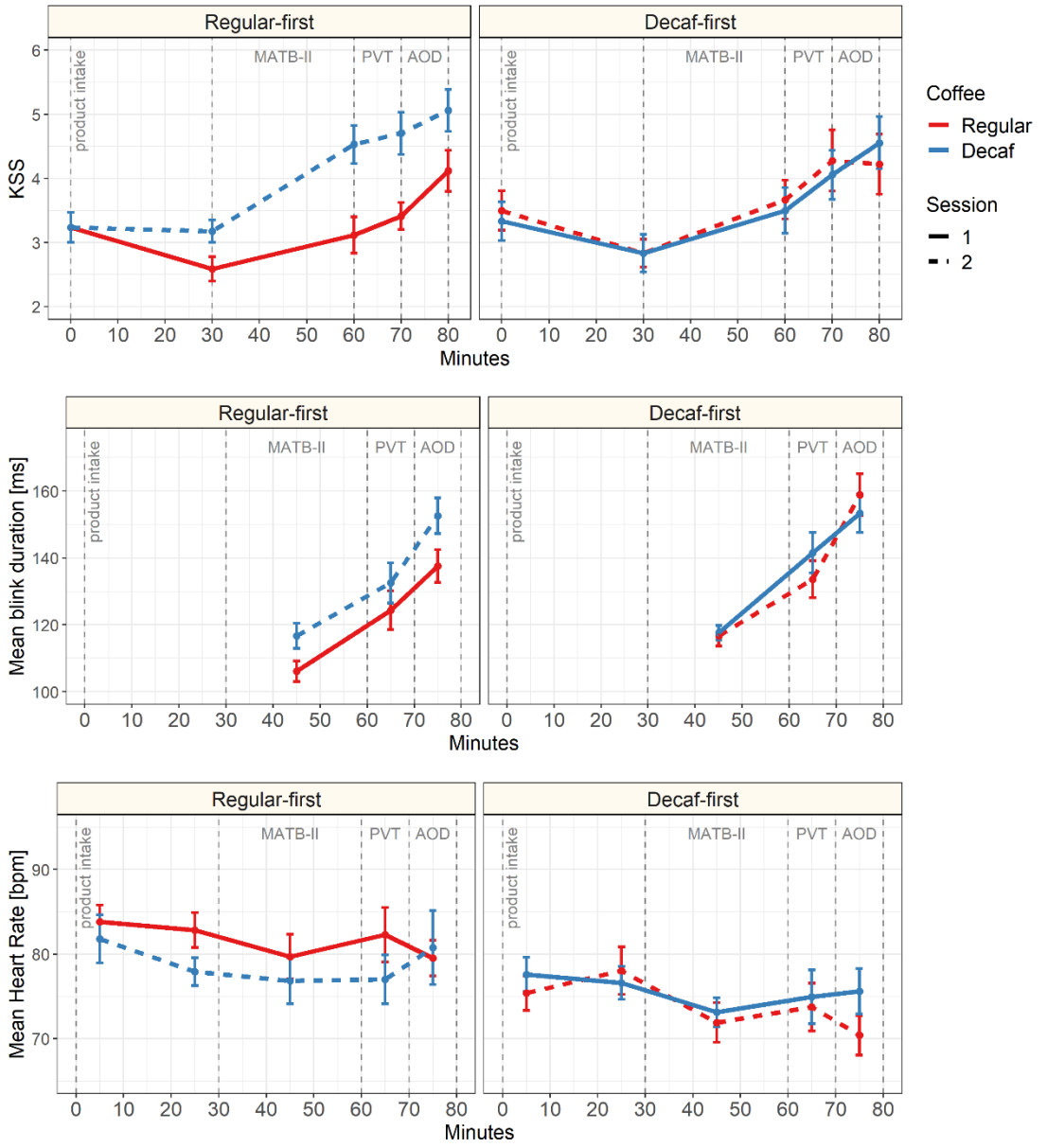

Figure 5 ***Mean KSS ratings (top panel: 1 = extremely alert; 5 = neither alert nor sleepy; 9 = extremely sleepy/fighting sleep); mean blink duration (middle panel) and mean heart rate (bottom panel) over time in Session 1 (solid lines) and Session 2 (dashed lines) for the regular-first (left) and the decaf-first group (right). Red lines reflect regular coffee sessions, blue lines reflect decaf sessions. Error bars show ±SEM***

Table 2 ***Result summary of repeated-measures ANOVAs of KSS ratings, mean blink duration (in milliseconds) and heart rate (in beats per minute): F-statistics, significance levels (*p<.05; **p<.01; ***p<.001) and effect sizes (η^2_G). Significant effects (p<.05) in bold.***

	ALL		Regular-first		Decaf-first	
	F	η^2_G	F	η^2_G	F	η^2_G
KSS ratings						
Time	**23.2*****	.16	**14.5*****	.24	**9.83*****	.12
Session	**9.65****	.03	**22.9*****	.14	0.04	.0002
Treatment Order	.02					
Time x Session	**3.22***	.01	**6.60*****	.06	0.743	.005
Time x Treatment Order	.48					
Session x Treatment Order	**7.82****	.02				
Time x Session x Treatment Order	**3.04***	.01				
Blink duration (ms)						
Time	**66.9*****	.35	**32.0*****	.34	**35.2*****	.36
Session	**7.35***	.02	**24.8*****	.11	0.05	
Treatment Order	2.65					
Time x Session	2.68		1.63		1.69	
Time x Treatment Order	.18					
Session x Treatment Order	**9.36****	.03				
Time x Session x Treatment Order	.54					
Heart rate (bpm)						
Time	2.66		1.17		2.16	
Session	**4.36***	.01	**6.33***	.01	0.97	
Treatment Order	3.83					
Time x Session	.08		1.28		1.54	
Time x Treatment Order	.62					
Session x Treatment Order	.168					
Time x Session x Treatment Order	2.72					

3.4 Self-report measures

3.4.1 Task-related

The analysis of global NASA-TLX scores (i.e., experienced MATB-II workload: 0=very low, 100=very high), showed a main effect of Session, $F(1,32)=6.08$, $p=.019$, $\eta^2_G=0.03$, which was significantly mediated by Treatment Order, $F(1,32)=6.25$, $p=.018$, $\eta^2_G=0.03$. Pairwise comparisons per treatment order group showed that the decaf-first group experienced a lower workload during the second (regular coffee) session (M=25.7) relative to the first (decaf) session (M=36.0), difference 10.3, $t(16)=4.11$, $p<.001$, $d_z=0.62$. For the regular-first group, experienced workload did not differ between sessions (M=33.7 in both sessions; p=.98). As for subjective task performance ratings (1=very bad, 10=very good), a main effect of Task ($F(1.5,47.9)=53.1$,$p<.001$, $\eta^2_G=0.29$) indicated that performance ratings for AOD (M=6.7) were significantly

lower than for PVT (M=8.1) and MATB-II (M=8.3) (p's <.001). No evidence was found for a difference between sessions and/or treatment order groups for any of the tasks (all F<1.5, p's≥.24).

3.4.2 Product-related

Participants did not distinguish between regular coffee (M=6.4) and decaf coffee (M=6.2) in terms of liking (p=.216). Moreover, treatment guessing behaviour was at chance (Table 3): 18 out of 35 participants correctly guessed when they had decaf coffee, confirming the effectivity of the placebo. In fact, the majority of participants guessed they had decaf coffee in the second session, which suggests that they attributed the experiment-wide repetition effect to a lack of caffeine.

Table 3 ***Decaf coffee session guesses in the full sample and per treatment order. Asterisks indicate correct guesses.***

	Full sample N=35	**Regular-first N=17**	**Decaf-first N=18**
Decaf coffee session guess			
Session 1	6 (17%)	1 (6%)	5* (28%)
Session 2	23 (66%)	13* (76%)	10 (56%)
I don't know	6 (17%)	3 (18%)	3 (17%)

4 Discussion

This study was set out to investigate the suitability of a realistic professional multi-task working environment (MATB-II) for measuring intervention effects on vigilance and attention in pilots. To this end, the effects of regular vs. decaffeinated coffee on performance on two MATB-II subtasks (SYSMON, COMM) were directly compared with coffee effects on performance on their simple lab-based equivalents (PVT, AOD).

Findings provided no evidence that coffee effects on vigilance- and attention-related performance correlated across task environments. In fact, for vigilance performance, no evidence was found for the hypothesized positive effect of regular (vs. decaf) coffee at all, neither for the MATB-II subtask (SYSMON) nor for its simple equivalent (PVT). Instead, a test-retest effect was found for both vigilance tasks, which went in opposite directions: whereas task repetition led to faster responses in the multitask environment (MATB-II SYSMON), it slowed down responses in the less stimulating environment (PVT). As for attention-related performance, findings did show the hypothesized beneficial effect of regular (vs. decaf) coffee (regardless of test session), but only in the simple environment (AOD); in the multitask environment (MATB-II COMM) attention-related performance did not differ between coffee conditions, likely due to an observed ceiling effect. Furthermore, subjective and objective measures of sleepiness and arousal revealed an experiment-wide test-retest effect, such that participants were more sleepy and less aroused during the second test session. This effect was attenuated in the decaf-first group, suggesting that regular (vs. decaf) coffee counteracted an experiment-wide boredom effect. The mediating effect of coffee on experiment-wide boredom was reflected in the experienced workload during the MATB-II task, but not in vigilance-related nor in attention-related task performance.

Our vigilance-related findings contrast with a wealth of studies showing beneficial effects of caffeine on PVT performance (e.g., Fine et al., 1994; Lanini et al. 2016; Cooper et al., 2021; McLellan et al., 2016), but are in line with Doan and colleagues (2006), who investigated caffeine effects on repeated cognitive performance in pilots during a 9-hour simulated flight. The authors found beneficial effects of caffeine (vs. placebo) on repeated performance of multiple cognitive tasks, but failed to find effects on a Scanning Visual Vigilance Test (which is highly similar to a PVT – although their Scanning Visual Vigilance Test lasted longer than the PVT used in our study). The authors speculatively ascribed their findings to a lack of motivation caused by the (reportedly) monotonous nature of the scanning vigilance task. Our findings fit with this interpretation: although regular coffee was an effective countermeasure against experiment-wide boredom (as indicated by sleepiness and arousal measures), it could not sufficiently counteract the combined boredom induced by repeatedly performing a simple task in a simple environment (I.e., PVT), nor could it further enhance performance in an inherently stimulating multitask environment (i.e., MATB-II). Put differently: our findings raise the suggestion that the inherent (de) stimulating nature of the task environment had a stronger impact on vigilance than our intervention.

Findings from the attention-related tasks showed that regular coffee (marginally) improved performance in the simple environment (AOD), but not in the multitask environment (MATB-II COMM) where performance accuracy was at ceiling. This finding might relate to the way in which the oddball paradigm is implemented in MATB-II. That is, participants can only respond to radio messages after having listened to the full length of the stimulus (nine seconds on average). Whereas in the auditory oddball detection (AOD) much faster responses (within two seconds after having heard a one-word stimulus) are required, making it more prone to false alarms and thus yielding a more sensitive discriminability measure. As such, the MATB-II communication task (COMM) seems less suitable for measuring intervention effects on selective attention when compared with simpler equivalents such as the AOD.

As mentioned above, both the results of the subjective (KSS) and objective measures of sleepiness (blink rate and duration) and arousal (HR) suggest that regular coffee could counteract an experiment-wide boredom effect. Participants were more sleepy and less aroused during the second test session, but this test-retest effect was mitigated in the group that had regular coffee during this session. Although we are not aware of studies with the exact same set-up as ours, similar effects of caffeine on (continuous) attention (Frewer & Lader, 1991), subjective fatigue (Horne & Reyner, 1996; Michael et al., 2008), blink characteristics (Michael et al., 2008), and heart rate (Green et al.,1996) have been reported before; yet the evidence, especially during normal, non-sleep deprived, conditions is limited. Earlier, Loke (1988) found that

caffeine intake of 0, 200, and 400 mg showed nonsignificant effects on cognitive, learning, and memory performance in a double-blind study. However, caffeine did decrease the level of boredom and relaxation, while increasing other subjective mood ratings (anxiousness, tenseness, and nervousness). The author argued that the reduction in boredom/fatigue was associated with the repetitive nature of the task and the period during which the tasks were repeated, which is in accordance with our findings. Our findings furthermore show that non-verbal, objective physiological measures of arousal show similar effects of caffeine interventions as self-reported, subjective measures.

4.1 Limitations

The absence of indications for the hypothesized effect of caffeinated coffee on PVT performance may in part be related to the fact that our study included well-rested subjects (instead of more widely studied sleep-deprived individuals; e.g., Lim & Dinges, 2008) and because we used a relatively moderate dose of caffeine (~98 mg) when compared with more commonly used doses (~200 mg or more; see McLellan et al., 2016). However, the positive (one-tailed) effect of regular coffee on attention-related performance confirms that the current dose was sufficient to exert behavioural effects, and previous studies have reported beneficial effects of even smaller doses (<50 mg) in both rested and sleep-deprived subjects (E.g., Smit & Rogers, 2000; for review, see McLellan et al., 2016). A recent resting state fMRI study compared the effects of a regular coffee with that of the same amount of caffeine in hot water and showed that coffee, but not caffeine, increased brain connectivity in visual and executive control networks (Picó-Pérez et al., 2023). Previous research already demonstrated that both caffeine and the mere expectation of having consumed caffeine improved attention and psychomotor speed (Dawkins et al., 2011). In our study these possible placebo effects were controlled for, as the product-related questions showed that participants did not distinguish between regular and decaf coffee in terms of liking nor in treatment guessing behaviour. However, the anticipation of coffee in both intervention arms might have partly diluted our effects. Moreover, although the current 'dose' in the form of one cup of coffee is more representative of caffeine intake in real-life than a single dose of e.g., 200mg caffeine, repeated coffee consumption across extended periods of time could have corresponded better with the caffeine intake of pilots during actual flight operations and therefore might have shown more distinctive effects (Kamimori et al., 2015). Furthermore, it is important to note that the administered caffeine dose in our study may have been insufficient to overcome potential withdrawal effects from participants' habitual caffeine consumption. The study participants arrived at the lab in a caffeine-deprived state, and caffeine withdrawal could have negatively influenced their performance on the tasks.

In exploring the multi-task environment of the MATB-II, our aim was to emulate a real-life working scenario for pilots. However, it is noteworthy that our assessment did not encompass vigilance- and attention-related performance within the actual operational context, such as during a flight. Actual flights impose heightened demands, increased responsibilities, and extended durations exceeding 30 minutes, factors that are likely to influence the impact of caffeine on vigilance. Notably, the elevated signal rates in laboratory-based vigilance and monitoring tasks have been criticized for not truly reflecting real-world tasks (Mackie, 1987).

Comparatively, the signal rate in the MATB-II SYSMON task, though lower than that of the PVT (~4 events per minute versus ~20 events per minute, respectively), may still be considered potentially high in terms of ecological validity (Molloy & Parasuraman, 1996). It is essential to acknowledge that even in a prolonged, low signal rate, multi-task vigilance environment, the inherent nature of multi-tasking, with its task-switching requirements and associated switch costs, makes it inherently more stimulating than a low-signal PVT (Monsell, 2003). Consequently, the observed effects in the MATB-II, stemming from its multi-task environment, are likely to have more relevance to real-life situations than those observed in the PVT, even when considering lower signal rates over an extended session.

Furthermore, it is crucial to recognize the potential for the MATB-II to provide a more realistic outcome in comparison to traditional tests like the PVT and AOD. The simplicity of the PVT and AOD, acknowledged in our introduction as potentially limiting their ability to mimic real-life scenarios, raises questions about whether these established tests might exaggerate the effects of stimulants or fatigue countermeasures. Therefore, it is plausible that the MATB-II, with its multi-dimensional tasks and real-world relevance, accurately identifies minimal to non-existent effects, as exemplified in our study involving the impact of one cup of coffee on morning performance. Future investigations should delve into potential trade-offs among

MATB-II sub-tasks and their difficulty (Kong et al., 2022), especially in relation to task duration, to further refine the test's applicability to realistic real-life situations. This acknowledgment opens up avenues for considering the MATB-II as a more relevant and nuanced assessment tool.

It is important to acknowledge that the current study, conducted exclusively in the morning, deviates from the intended representation of professionals working outside regular office hours, as initially highlighted in the introduction. This deviation may introduce a potential limitation, as the absence of testing during post-office hours fails to capture the nuances of cognitive performance influenced by factors like sleep debt and heightened fatigue at the beginning of a workday. Recognizing this constraint is crucial, as it may impact the generalizability of the study findings to a broader context of professionals facing non-traditional work schedules. Addressing this limitation encourages future research to explore cognitive performance in scenarios aligning with the diverse work patterns observed in contemporary professional settings.

4.2 Implications for practice and future research

The findings of our study indicate that there is no significant advantage of MATB-II over PVT in measuring intervention effects on vigilance performance. However, MATB-II is also not inferior to PVT since both tasks failed to demonstrate intervention effects. On the one hand, this might suggest that the intervention employed was not sufficiently effective in producing measurable changes in vigilance. On the other hand, intervention effects were observed on the oddball task as well as on physiological measures and experienced MATB-II workload, demonstrating the effectiveness of the intervention on at least these measures. Nevertheless, MATB-II was less effective than the oddball task in capturing intervention effects on selective attention. Together, these results suggest that MATB-II in its current form may be better suited as a stressor manipulator within a multitask environment, primarily focusing on measuring physiological responses and the impact of interventions on fatigue and workload, rather than as a tool for assessing intervention effects on cognitive performance.

Moreover, our findings highlight the impact of task repetition on vigilance performance. The possible overshadowing effect of task repetition should therefore be taken into account when analysing and interpreting results of intervention studies, especially using a crossover design, or when performing repeated tests in observational studies (as in e.g., van Dongen et al., 2003; Sallinen et al., 2020; Gander et al., 2014). When investigating vigilance, workload and fatigue, task order within a session is also important to take into account. Here, task order was kept constant between participants, but are not likely to explain the differential task repetition effects in the vigilance tasks, as the attention-related tasks were following each other in a similar order (from complex, MATB-II, to simple task environment), but were not differentially affected by task repetition.

Attention-related performance was not affected by task repetition, but the current implementation of the oddball task in MATB-II seems less suitable for measuring intervention effects. This could potentially be improved by adjusting the task settings. For instance, including a response requirement upon hearing the name of the aircraft would force participants to immediately distinguish between relevant ('hit') and irrelevant call signs (requiring a 'correct rejection'), from which a more sensitive outcome measure can be derived. Turning this into a speeded judgment will help capturing the effects of interventions on reaction time and information processing speed. Although task customization is difficult in MATB-II (as the source code is not freely available), the recently launched open-source variant *OpenMATB* (Cegarra et al., 2020) may uncover this possibility.

Our findings also show that eye blink and heart rate measurements mimicked subjective (self-report) measures of sleepiness. This has great potential for testing intervention effects in real-life, as the use of such technologies allows for an assessment of sleepiness and vigilance without relying on explicit responses from professionals (and hence interrupting their professional duties). Before such non-obtrusive, objective measurements can be used on a larger scale, however, more research and development is needed to optimize this technological equipment for application in field studies involving real-life working situations. By demonstrating both task repetition effects as well as intervention effects during the MATB-II, our results do show that - as an intermediate step towards field studies - the MATB-II can serve as an appropriate test environment for assessing such implicit, physiological measures of vigilance and fatigue under various circumstances.

5 Conclusion

In summary, while the current iteration of MATB-II may not be optimal for evaluating intervention effects on cognitive performance, including vigilance and selective attention, it exhibits promise in alternative domains. Specifically, MATB-II proves valuable for monitoring physiological responses and fatigue levels, making it a suitable tool for investigating intervention effects in a multitask setting. Incorporating recommended adjustments, such as the inclusion of speeded judgment tasks and correct rejections, could enhance MATB-II's capacity to measure intervention effects on cognitive performance, particularly selective attention. Despite the need for further refinement, MATB-II holds potential as a reliable intervention tool. Continued research and enhancements may position it to significantly contribute to intervention studies focused on cognitive performance.

Moreover, it is important to consider that the MATB-II may present a more realistic outcome compared to traditional tests like the PVT and AOD. As highlighted in our introduction, concerns exist regarding the simplicity of these widely-used tests in mimicking real-life scenarios. The potential for exaggeration of the effects of stimulants or other fatigue countermeasures in PVT and AOD underscores the relevance of MATB-II as a test that could accurately identify minimal to non-existent effects, as exemplified in our study focusing on the impact of one cup of coffee on morning performance. This recognition positions MATB-II as a potentially more realistic and valuable tool for cognitive performance assessment.

6 Declaration & data availability statement

The authors have no competing interests to declare that are relevant to the content of this report.
All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Maykel van Miltenburg, Lea Riesenbeck, Jane Sieters, Ruud van Stiphout, Garmt Dijksterhuis & Geertje van Bergen. The first draft of the manuscript was written by Geertje van Bergen, Maykel van Miltenburg, Alwin van Drongelen & Esther Aarts. All authors commented on previous versions of the manuscript, and all authors read and approved the final manuscript.

The data that support the findings of this study are not openly available due to reasons of sensitivity and are available from the corresponding author upon reasonable request. Data are located in controlled access data storage at Wageningen University & Research

7 Literature

Adams, J. A. (1987). Criticisms of Vigilance Research: A Discussion. *Human Factors*, 29(6), 737–740.

Akerstedt T, & Gillberg M. (1990). Subjective and objective sleepiness in the active individual. *International Journal of Neuroscience*, 52(1-2), 29–37.

Al-Shargie F, Tariq U, Mir H, Alawar H, Babiloni F, Al-Nashash H. (2019). Vigilance Decrement and Enhancement Techniques: A Review. *Brain Sciences*, 26 (9), 178. https://doi.org/10.3390/brainsci9080178.

Baker, W. J., & Theologus, G. C. (1972). Effects of caffeine on visual monitoring. *Journal of Applied Psychology*, 56(5), 422–427.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3). https://doi.org/10.1016/j.jml.2012.11.001.

Barone, J. J., & Roberts, H. R. (1996). Caffeine consumption. *Food and Chemical Toxicology*, 34(1), 119–129.

Caldwell Jr, J. A., Caldwell, J. L., Brown, D. L., & Smith, J. K. (2004). The effects of 37 hours of continuous wakefulness on the physiological arousal, cognitive performance, self-reported mood, and simulator flight performance of F-117A pilots. *Military Psychology*, 16(3), 163–181.

Carlozzi, N. E., Horner, M. D., Kose, S., Yamanaka, K., Mishory, A., Mu, Q., ... & George, M. S. (2010). Personality and reaction time after sleep deprivation. *Current Psychology*, 29(1), 24–33.

Casagrande, M. (2011). Flight Safety: Microsleeps, vigilance and stress in civil and military pilots. *Italian Journal of Aerospace Medicine*, 4, 44–50.

Cegarra, J., Valéry, B., Avril, E., Calmettes, C., & Navarro, J. (2020). OpenMATB: A Multi-Attribute Task Battery promoting task customization, software extensibility and experiment replicability. *Behavioral Research Methods*, 52, 1980–1990. https://doi.org/10.3758/s13428-020-01364-w

Comstock, J. R., & Arnegard, R. J. (1992). The Multi-Attribute Task battery for human operator workload and strategic behaviour research. National Aeronautics and Space Administration, Langley Research Center, NASA TM-104174.

Cooper Jr, R. K., Lawson, S. C., Tonkin, S. S., Ziegler, A. M., Temple, J. L., & Hawk Jr, L. W. (2021). Caffeine enhances sustained attention among adolescents. *Experimental and Clinical Psychopharmacology*, 29(1), 82.

Davies, D. R., & Parasuraman, R. (1982). *The psychology of vigilance*. Academic Press.

Dawkins, L., Shahzad, F. Z., Ahmed, S. S., & Edmonds, C. J. (2011). Expectation of having consumed caffeine can improve performance and mood. *Appetite*, 57(3), 597-600.

Dinges, D. F., & Powell, J. W. (1985). Microcomputer analyses of performance on a portable, simple visual RT task during sustained operation. *Behavior Research Methods, Instrument & Computers*, 17(6), 652–655.

Doan, B. K., Hickey, P. A., Lieberman, H. R., & Fischer, J. R. (2006). Caffeinated tube food effect on pilot performance during a 9-hour, simulated nighttime U-2 mission. *Aviation, Space, and Environmental Medicine*, 77(10), 1034–1040.

Federal Aviation Administration (June 7, 2010). Basics of Aviation Fatigue (Advisory Circular 120-100). Retrieved from: https://www.faa.gov/documentLibrary/media/Advisory_Circular/AC%20120-100.pdf

Fine, B. J., Kobrick, J. L., Lieberman, H. R., Marlowe, B., Riley, R. H., Tharion, W. J. (1994). Effects of caffeine or diphenydramine on visual vigilance. *Psychopharmacology*, 114, 233–238.

Frewer, L. J., & Lader, M. (1991). The effects of caffeine on two computerized tests of attention and vigilance. *Human Psychopharmacology: Clinical and Experimental*, 6(2), 119-128.

Gander, P. H., Mulrine, H. M., van den Berg, M. J., Smith, A. A. T., Signal, T. L., Wu, L. J., & Belenky, G. (2014). Pilot Fatigue: Relationships with Departure and Arrival Times, Flight Duration, and Direction. *Aviation, Space, and Environmental Medicine*, 85(8), 833–840. https://doi.org/10.3357/ASEM.3963.2014

Green, P. J., Kirby, R., & Suls, J. (1996). The effects of caffeine on blood pressure and heart rate: a review. *Annals of Behavioral Medicine*, 18(3), 201-216.

Hart, S. G. & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in Psychology*, 52, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

Hermann, C.S., & Knight, R.T. (2001). Mechanisms of human attention: event-related potentials and oscillations. *Neuroscience and Biobehavioral Reviews*, 25, 465–476. https://doi.org/10.1016/S0149-7634(01)00027-6

Horne, J. A., & Reyner, L. A. (1996). Counteracting driver sleepiness: effects of napping, caffeine, and placebo. *Psychophysiology*, 33(3):306-9. https://doi.org/10.1111/j.1469-8986.1996.tb00428.x.

Kamimori, G. H., McLellan, T. M., Tate, C. M., Voss, D. M., Niro, P., & Lieberman, H. R. (2015). Caffeine improves reaction time, vigilance and logical reasoning during extended periods with restricted opportunities for sleep. *Psychopharmacology*, 232(12), 2031–2042. https://doi.org/10.1007/s00213-014-3834-5.

Kibler, A. W. (1965). The Relevance of Vigilance Research to Aerospace Monitoring Tasks. *Human Factors*, 7(2), 93–99.

Koelega, H. S. (1993). Stimulant drugs and vigilance performance: a review. *Psychopharmacology*, 111(1), 1–16.

Kong, Y., Posada-Quintero, H. F., Gever, D., Bonacci, L., Chon, K. H., & Bolkhovsky, J. (2022). Multi-Attribute Task Battery configuration to effectively assess pilot performance deterioration during prolonged wakefulness. *Informatics in Medicine Unlocked*, *28*, 100822. https://doi.org/10.1016/j.imu.2021.100822

Kourtidou-Papadeli, C., Papadelis, C., Louizos, A.-L., Guiba-Tziampiri, O. (2002). Maximum cognitive performance and physiological time trend measurements after caffeine intake. *Cognitive Brain Research*, 13, 407-415. https://doi.org/10.1016/S0926-6410(01)00133-1.

Lanini, J., Galduróz, J. C., & Pompéia, S. (2015). Acute personalized habitual caffeine doses improve attention and have selective effects when considering the fractionation of executive functions. *Human Psychopharmacology*, 31(1), 29–43.

Lim, J., & Dinges, D. (2008). Sleep deprivation and vigilant attention. *Annals of the New York Academy of Sciences*, 1129(1), 305–322.

Loke, W. H. (1988). Effects of caffeine on mood and memory. *Physiology & Behavior*, 44(3), 367-372.

Lorist, M., Snel, J., Kok, A., & Mulder, G. (1994). Influence of caffeine on selective attention in well-rested and fatigued subjects. *Psychophysiology*, 31(6), 525–534. https://doi.org/10.1111/j.1469-8986.1994.tb02345.x.

Mackie, R. R. (1987). Vigilance research – Are we ready for the countermeasures? *Human Factors*, 29, 707-723.

McLellan, T., Caldwell, J. A., & Lieberman, H. R. (2016). A review of caffeine's effects on cognitive, physical and occupational performance. *Neuroscience and Biobehavioral Reviews*, 71, 294–312.

Michael, N., Johns, M., Owen, C., & Patterson, J. (2008). Effects of caffeine on alertness as measured by infrared reflectance oculography. *Psychopharmacology*, 200, 255-260.

Molloy, R., & Parasuraman, R. (1996). Monitoring an automated system for a single failure: Vigilance and task complexity effects. *Human Factors*, 38(2), 311–322.

Narita, Y., & Inouye, K. (2015). Chapter 21 - Chlorogenic Acids from Coffee. In *Coffee in Health and Disease Prevention* (pp. 189-199). Academic Press. https://doi.org/10.1016/C2012-0-06959-1

Navarro, J., Heuveline, L., Avril, E., & Cegarra, J. (2018). Influence of human-machine interactions and task demand on automation selection and use. *Ergonomics*, 61(12), 1601–1612. https://doi.org/10.1080/00140139.2018.1501517

Nehlig, A. (2018). Interindividual differences in caffeine metabolism and factors driving caffeine consumption. *Pharmacological Reviews*, 70(2), 384-411.

Oken, B. S., Salinsky, M. C., & Elsas, S. M. (2006). Vigilance, alertness, or sustained attention: physiological basis and measurement. *Clinical Neurophysiology*, 117(9), 1885–1901.

Pan, J., Takeshita, T., & Morimoto, K. (2000). Acute Caffeine Effect on Repeatedly Measured P300. *Environmental Health and Preventive Medicine*, 5, 13-17.

Papadelis, C., Kourtidou-Papadeli, C., Vlachogiannis, E., Skepastianos, P., Bamidis, P., Maglaveras, N., & Pappas, K. (2003). Effects of mental workload and caffeine on catecholamines and blood pressure compared to performance variations. *Brain and Cognition*, 51, 143-154. https://doi.org/10.1016/S0278-2626(02)00530-4.

Picó-Pérez, M., Magalhães, R., Esteves, M., Vieira, R., Castanho, T. C., Amorim, L., ... & Sousa, N. (2023). Coffee consumption decreases the connectivity of the posterior Default Mode Network (DMN) at rest. *Frontiers in Behavioral Neuroscience*, 17, 1176382.

Rogers, P. J., & Smith, J. E. (2011). Caffeine, mood and cognition. In *Lifetime nutritional influences on cognition, behaviour and psychiatric illness* (pp. 251–271). Oxford: Woodhead Publishing.

Sallinen, M., van Dijk, H., Aeschbach, D., Maij, A., & Åkerstedt, T. (2020). A Large-Scale European Union Study of Aircrew Fatigue During Long Night and Disruptive Duties. *Aerospace Medicine and Human Performance*, 91(8), 628–635. https://doi.org/10.3357/AMHP.5561.2020

Santiago-Espada, Y., Myer, R. R., Latorella, K. A., & Comstock Jr, J. R. (2011). The multi-attribute task battery II (MATB-II) software for human performance and workload research: A user's guide. NASA/TM–2011-217164.

Smit, H. J., & Rogers, P. J. (2000). Effects of low doses of caffeine on cognitive performance, mood and thirst in low and higher caffeine consumers. *Psychopharmacology*, 152(2), 167–173. https://doi.org/10.1007/s002130000050

Smith, M. E., & Gevins, A. (2005, May). Neurophysiologic monitoring of mental workload and fatigue during operation of a flight simulator. In *Biomonitoring for Physiological and Cognitive Performance during Military Operations* (Vol. 5797, pp. 116–126). International Society for Optics and Photonics.

Van Dongen, H. P. A., Maislin, G., Mullington, J. M., & Dinges, D. F. (2003). The Cumulative Cost of Additional Wakefulness: Dose-Response Effects on Neurobehavioral Functions and Sleep Physiology From Chronic Sleep Restriction and Total Sleep Deprivation. *Sleep*, 26(2), 117–126. https://doi.org/10.1093/sleep/26.2.117

Wiener, E. L., Curry, R. E., & Faustina, M. L. (1984). Vigilance and Task Load: In Search of the Inverted U. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 26(2), 215–222. https://doi.org/10.1177/001872088402600208

8 Supplemental materials

8.1 Blink characteristics

Blink rates and blink duration variability confirmed a gradual increase in sleepiness as the experiment progressed. Moreover, a significant Session x Treatment order interaction for blink duration variability indicated that participants were sleepier during the second session, but only in the regular-first group. Blink rates show a similar pattern, but the Session x Treatment Order interaction was not significant (p=.27).

Table S1. ***Result summary of repeated-measures ANOVAs of blink duration variability and blink rate: F-statistics, significance levels (*p<.05; **p<.01; ***p<.001) and effect sizes (η2P) (p-values < 0.05 in bold).***

	ALL		**Regular-first**		**Decaf-first**	
	F	η^2_P	F	η^2_P	F	η^2_P
Standard deviation blink duration						
Time	**41.6*****	.59	**35.9*****	.72	**13.9*****	.48
Session	**5.51***	.16	**28.5*****	.67	.68	.04
Treatment Order	1.29	.04				
Time x Session	.85	.03	.55	.04	.34	.02
Time x Treatment Order	.58	.02				
Session x Treatment Order	**13.5*****	.32				
Time x Session x Treatment Order	.02	.01				
Mean blink rate						
Time	**61.1*****	.68	**38.1*****	.73	**23.9*****	.62
Session	2.42	.08	1.94	.12	.37	.02
Treatment Order	.38	.01				
Time x Session	2.43	.08	2.0	.13	.54	.04
Time x Treatment Order	1.87	.06				
Session x Treatment Order	1.26	.04				
Time x Session x Treatment Order	.49	.02				

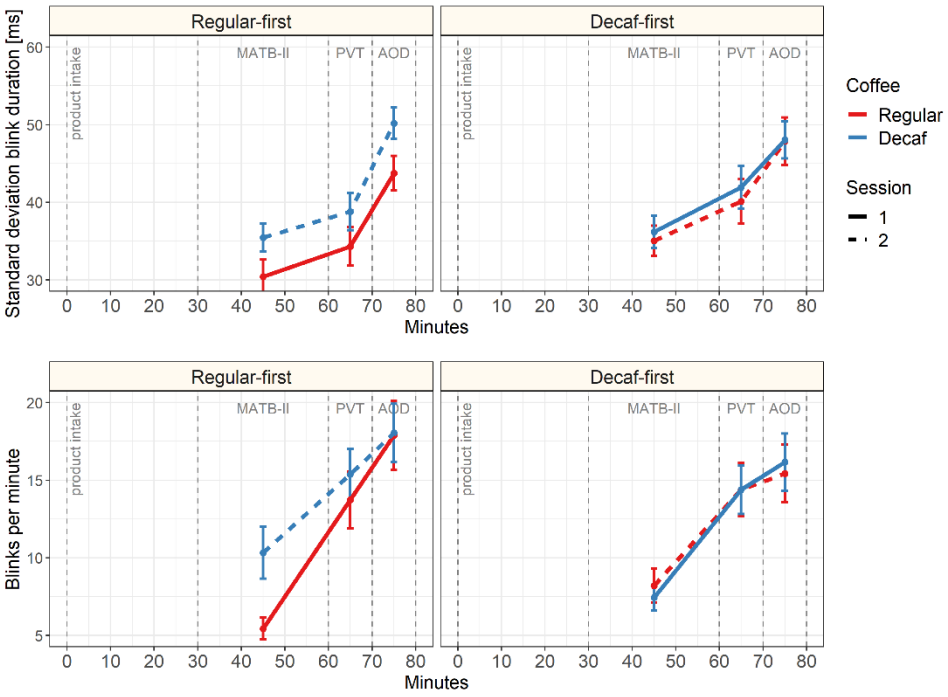

Figure S1 ***Blink duration variability (top panel) and blink rate (bottom) over time in regular coffee session (red lines) and decaf coffee session (blue lines) for the Regular-first (left) and the Decaf-first group (right). Solid lines reflect session 1, dashed lines session 2. Error bars show ± standard error of the mean (SEM).***

To explore the potential of nature to improve the quality of life

Wageningen Food & Biobased Research
Bornse Weilanden 9
6708 WG Wageningen
The Netherlands
E info.wfbr@wur.nl
wur.eu/wfbr

Report 2610

The mission of Wageningen University & Research is "To explore the potential of nature to improve the quality of life". Under the banner Wageningen University & Research, Wageningen University and the specialised research institutes of the Wageningen Research Foundation have joined forces in contributing to finding solutions to important questions in the domain of healthy food and living environment. With its roughly 30 branches, 7,700 employees (7,000 fte), 2,500 PhD and EngD candidates, 13,100 students and over 150,000 participants to WUR's Life Long Learning, Wageningen University & Research is one of the leading organisations in its domain. The unique Wageningen approach lies in its integrated approach to issues and the collaboration between different disciplines.