Investigating the (ir)reducibility and within- and between-subject correlates of intra-individual variability

Marlou Nadine Perquin

A thesis submitted to Cardiff University in fulfilment of the requirements for the degree of Doctor of Philosophy

August 2019

Declaration and Statements

Declaration

This work has not been submitted in substance for any other degree or award at this or any other university or place of learning, nor is being submitted concurrently in candidature for any degree or other award.

Signed: _____ (candidate) Date: _____

Statement 1

This thesis is being submitted in partial fulfilment of the requirements for the degree of PhD.

Signed: _____ (candidate) Date: _____

Statement 2

This thesis is the result of my own independent work/investigation, except where otherwise stated, and the thesis has not been edited by a third party beyond what is permitted by Cardiff University's Policy on the Use of Third Party Editors by Research Degree Students. Other sources are acknowledged by explicit references. The views expressed are my own.

Signed: _____ (candidate) Date: _____

Statement 3

I hereby give consent for my thesis, if accepted, to be available online in the University's Open Access repository and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed: _____ (candidate) Date: _____

Table of contents

Summary

Intra-individual variability is a prominent characteristic of our behaviour. A large part of this variability is *endogenous* – arising from fluctuations in our own inner states. In the Introduction, I identify two distinct literatures: 1) the *intuitive perspective,* which describes variability as a consequence of meta-cognitive fluctuations, and 2) the *intrinsic perspective,* which describes variability as a necessary feature of our nervous system. In this thesis, I compare these two literatures across four chapters.

In Chapter 1, I examined variability in the oculomotor system during rest, and found that variability is repeatable within. In Chapter 2, I found similar intra-individual reliability in variability on a rhythmic manual task, and in the temporal properties of variability. Furthermore, temporal structures correlated positively with variability, but did not correlate with subjective attentional state. In both chapters, variability did not correlate with ADHD, mind wandering, and impulsivity questionnaires.

In Chapter 3, I examined the relationships between variability, meta-cognition, and underlying neural activity. Results showed that participants were more variable on the task prior to off-task compared to on-task reports. Furthermore, neural states underlying attentional state reports showed overlap with those underlying behavioural variability. However, effect sizes were weak – implying that variability and meta-cognition are poor markers of each other. In Chapter 4, I tested a common intuition that people have some access to their fluctuating inner states which they can use to improve their performance. I found evidence against this assumption in both an ecological (darts) and two psychophysical tasks.

All in all, while the intuitive framework typically assumes a strong and possibly direct link between meta-cognition and behavioural variability, my current findings indicate that this link is clearly weak.

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my supervisor, Aline Bompas, for her expertise, wisdom, dedication, and support; for all the long scientific discussions we had; for always providing critical feedback to improve my work; for empowering me as a female scientist; and for encouraging me in my career.

Secondly, I would like to thank my secondary supervisors, Petroc Sumner and Jiaxiang Zhang, and my other co-authors, Craig Hedge and Christoph Teufel. Thank you for your knowledge and input. I would also like to express my gratitude to Gavin Perry and Krish Singh, for all their help in the MEG project. I am also very grateful to James Kolasinski, for always being supportive and understanding.

Furthermore, I would like to thank the students I supervised, for their work during the summer, and the project students I co-supervised with Aline, for collecting a great data set. Jess, Rachel, Laura F., Laura D., Joel, Ira, Nia, Natasha, Humera, and Sarah: Thank you. Particularly, I would like to thank Jess, for all her hard work on the darts experiment, even after her project was over.

On the non-scientific front, many thanks go to my family, for encouraging me to go to Cardiff to pursue this PhD, despite the dark times we were going through at that moment. Mama, papa, Lisa, Jeroen: dank voor jullie liefde en steun. Special thanks go to Puck and Thijn, for being the best niece and nephew an aunt could wish for. I also owe a debt of gratitude to Jamey, of whom I feel I have known her all my life. I would also like to thank Bárbara, Annelies, and Laura for our Scielliance, and for loving me as I am. Baba, I am glad for finishing this journey together with you, and I want to thank you for always understanding me. Many thanks go to Nicole,

for all long and distracting lunches we had, and for being the first friend I made in Cardiff.

Finally, I want to express my thanks to my friends, family, and colleagues, both in Cardiff and in the Netherlands, for all their support; and for offering some welcome distractions.

Impact

The results of Chapter 1 have been published as a peer-reviewed journal article, and have been presented at a scientific conference:

- **Perquin, M.N**. & Bompas, A. (2019). Reliability and correlates of intra-individual variability in the oculomotor system. *Journal of Eye Movement Research, 12*(6). doi: 10.16910/jemr.12.6.11
- **Perquin, M.N.** & Bompas, A. (2019). *Reliability and correlates of intra individual variability in the oculomotor system*. British Ocular Motor Group Annual Meeting, Cardiff.

The results of Chapter 2 have been made avaible as a pre-print on Biorvix:

- **Perquin, M.N**. & Bompas, A. (2019). Examining temporal structures in reaction time. *Biorxiv*. doi: 10.1101/817916

The results of Chapter 3 are in press as a peer-reviewed journal article, and have been presented at a scientific conference:

- **Perquin, M.N.**, Yang, J., Teufel, C., Sumner, P., Hedge, C. & Bompas, A. (in press). Inability to improve performance with control shows limited access to inner states. *Journal of Experimental Psychology: General.*
- **Perquin, M.N.** & Bompas, A. (2017). *The effects of predictability and higher task-control upon perception and action.* European Conference of Visual Perception, Berlin.

The results of Chapter 4 have been presented at two scientific conferences:

- **Perquin, M.N.**, Perry, G., Singh, K. & Bompas, A. (2019). *MEG correlates of mind wandering*. MEG UK, Cardiff.
- **Perquin, M.N.**, Perry, G., Singh, K. & Bompas, A. (2019). *MEG correlates of mind wandering and behavioural variability.* GW4 Early Career Neuroscientist Day, Exeter.

Introduction

Let us imagine you have a job at the museum, and you're tasked with counting the number of people who visit the special collection space. To ensure maximum accuracy, you press a clicker as soon as you see a person enter the room. However, even when you are highly motivated to do your task well, you will make some mistakes throughout the day; occasionally, you may forget to press the clicker, or accidentally press it twice. Moreover, the speed with which you click each time will be very inconsistent from time to time; even when your eyes are fixated on the entrance, sometimes it may take you only 200 milliseconds to press, while other times it may take a couple of seconds.

Now, the Board of the museum has decided to buy a robot to automate the counting process, to ensure that the staff can focus on helping the visitors. The robot is equipped with a high-tech sensor, and any time it 'sees' a person enters the room, it 'clicks' on its internal counter. Unlike you, the robot's behaviour will be extremely consistent over time; given that there are no defects, the robot will not make any mistakes in counting, and the speed of each execution will not differ by more than a couple of milliseconds.

The robot's performance on the task is near-constant over time, while yours shows high variability. Such variability may be referred to as '*intra-individual variability'* – variability within the same person – and is a ubiquitous phenomenon in all actions that we perform. This intra-individual variability and its properties are the common threads running through the current thesis. Over four experimental chapters, I will examine: 1) how variability manifests itself within individuals over different time points and different circumstances (*intra-individual correlates*), 2) to what extent variability may be used to differentiate between different people (*inter-*

individual correlates), and 3) to what extent variability is at all *reducible* (in other words, how closely can our performance resemble the robot's near-constant performance).

Exogenous and endogenous variability

So what is it that makes your performance so different from the robot's? At first glance, this question may have an obvious answer. The robot will be unaffected by any distractors or changes happening in the room. You on the other hand will be faced with a multitude of distractors – such as visitors asking you questions or children being loud. When a new person enters the room while you are answering a visitor's question about the special collection, it will likely take you much longer to press the clicker than if you are alone in the space and just waiting for people to walk in. Such examples relate to *'exogenous intra-individual variability'* – variability caused by changes in your external environment.

Still, these exogenous factors can only explain part of the variability. Let us imagine that the robot is send off to the factory for its annual check-up. During its absence, you are once again tasked with counting the visitors – but this time, you decide to set yourself a challenge and get as close to the robot's performance as possible. Close to the special collection's entrance, you build a small sound-proof booth which allows you to see the entrance, while people outside the booth are not able to see you. The only thing you bring with you is the clicker, and you keep watching the entrance the entire time you are in there – waiting to press once for each visitor. As such, your booth is completely devoid of any external distractions. However, even in these circumstances, your performance will not be similar to the robot's near-constant performance; you will still show a large variability in both speed and accuracy.

As this variability occurs without any changes in external factors, it should instead be caused by your own internal system. As such, this example relates to *'endogenous intra-individual variability'* – variability caused by changes in your own internal states. It is this type of variability that is of interest in the current thesis.

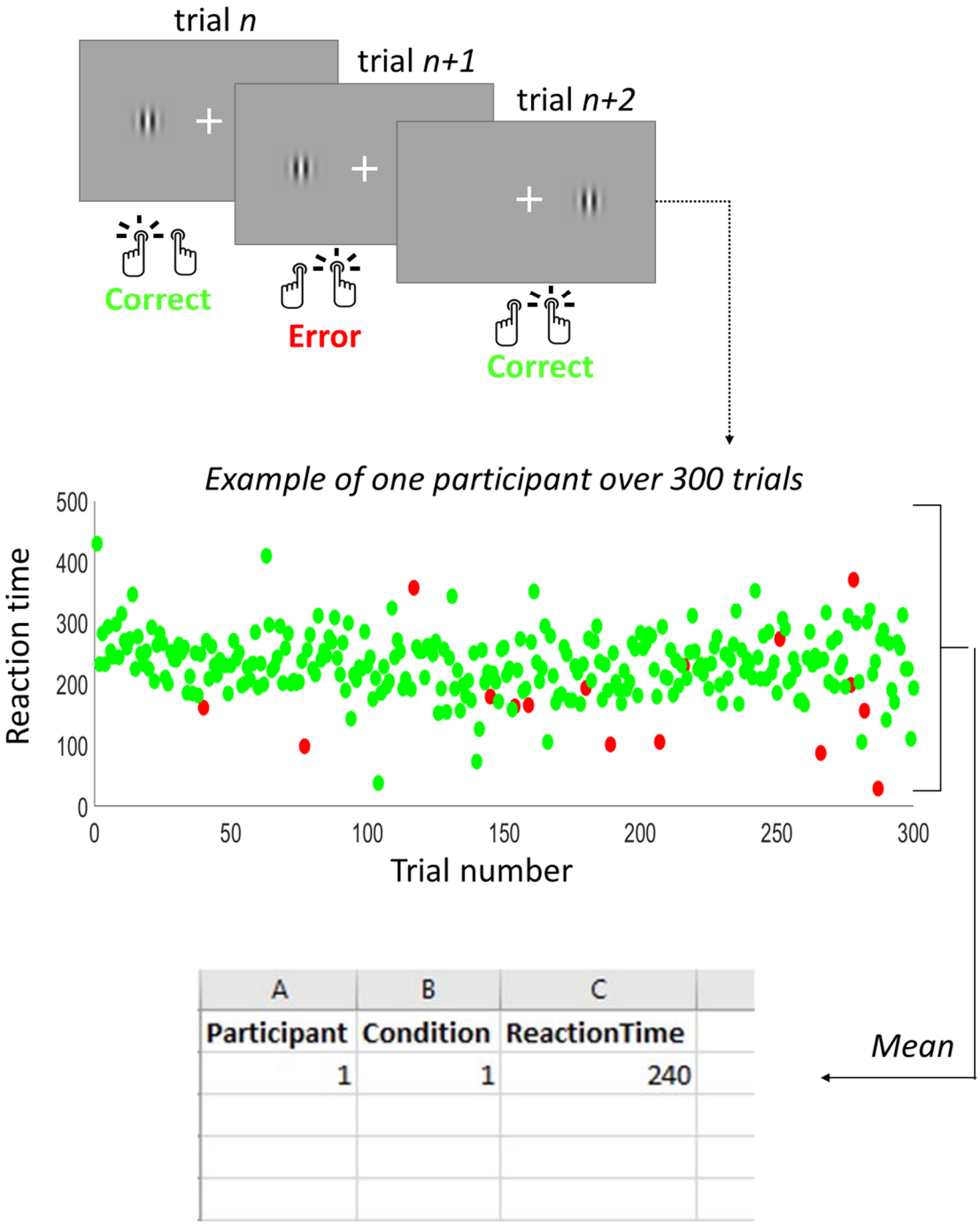


Figure 1. Example of the experimental data of one participant on a simple response task of about 15 minutes (300 trials). On every trial, the participant is presented with a high-contrast target and asked to indicate the location with a left-button or right-button press. The participant shows high fluctuations in speed over time, and also occasionally makes errors – even though the task is very simple to perform, the target is clearly visible each time, and there are no external distractions in the testing room. In typical analyses, this variability is 'averaged out' by calculating one mean over all the trials.

Studying variability

The existence of endogenous variability is clearly manifested in experimental data from typical neuro-cognitive experiments, in which there is equipment to measure variability precisely. While the working conditions of the small sound-proof booth in the museum may sound silly, these conditions closely resemble how participants are tested in psychological experiments: Testing labs are specifically built to be devoid from any external distractions, and participants are typically asked to perform a very limited set of actions over and over again. Even when the task is very basic, participants show high variability in their response time and accuracy (see Figure 1 for an example). In spite of the omnipresence of variability in behavioural data, it is often not considered. Instead, the variability is seen as 'measurement error' and 'averaged out' by calculating one (conditional) mean over all the trials. However, over the recent years, variability has gained more recognition as an interesting phenomenon in itself, as it may be informative in studying individual differences and/or behaviour.

Behavioural variability appears to be a consistent individual property across different types of tasks and modalities, as well as across different time points (Andrews & Coppola, 1999; Boot, Becic & Kramer, 2009; Castelhano & Henderson, 2008; Hultsch, MacDonald & Dixon, 2002; Poynter, Barber, Inman & Wiggins, 2013; Rayner, Li, Williams, Cave & Well, 2007; Saville et al., 2011; Saville et al., 2012; but see Salthouse, 2012) – meaning that if someone is for instance highly variable on one type of task, they are more likely to also be highly variable in another type of task. In this way, variability could be described as a 'personal trait'. Between individuals, this trait of variability has been negatively correlated with working memory capacity (Schmiedek, Oberauer, Wilhelm, Süβ & Wittmann, 2007) and intelligence (Schmiedek et al., 2007). Furthermore, variability may increase with cognitive aging (Hultsch et al., 2002; Hultsch, MacDonald, Hunter, Levy-Bencheton & Strauss, 2000). Related to this, increased variability has been associated with a multiplicity of neuropsychological disorders and diseases, such as ADHD (see Kofler et al., 2013 for a meta-analysis; see Tamm et al., 2012 for a review), Alzheimer's Disease (Tales et al., 2012; Tse, Balota, Yap, Duchek & MacCabe,

2010), as well as schizophrenia, depression, and borderline disorder (Kaiser et al., 2008). As such, behavioural variability could be a marker of general system dysfunctioning.

Furthermore, when calculating the participant-mean over the entire set of trials, it is commonly assumed that the residuals on these trials are independent from each other. However, in most experimental data, this assumption is false. Instead, performance shows temporal dependencies or 'clusters', such that good performance on trial *n* is followed by good performance on trial *n+1*, while poor performance on *n* is followed by poor performance on trial *n+1* – a phenomenon known as *'autocorrelation'.* These dependencies have been shown both in response speed (Gilden, 2001; Wagenmakers, Farrell & Ratcliff, 2004) and accuracy (Gilden, 2001; Monto, Palva, Voipio, & Palva, 2008), and may in some cases be larger than the effects of the experimental manipulations (Baayen & Milin, 2010; Gilden, 2001).

This means that with the processing of averaging, we do not just lose information on the variability, but also on the temporal structures. Likewise, it has been suggested that these structures may be consistent within individuals (Torre, Balasubramaniam, Rheaume, Lemoine & Zalznik, 2011), and that they show individual differences (Gilden & Hancock, 2007; Madison, 2004; Torre et al., 2011; Simola, Zhigalov, Morales-Muños, Palva & Palva, 2017). Still, the nature of these temporal structures as well as their benefit for studying individual differences remain largely unknown.

Two perspectives on endogenous variability

To summarise, so far we have seen that: 1) intra-individual variability is a ubiquitous phenomenon in human behaviour, 2) this variability and its underlying temporal structures may be interesting for studying individual differences, and 3) they may largely arise from *endogenous* factors. The question remains what constitutes these endogenous factors. Within the literature on intra-individual variability, we may distinguish two global sources of literature on this. From one source of literature, variability is perceived as a negative consequence of attentional lapses – for the

sake of this thesis, I shall refer to this as the *'intuitive perspective'*. From the second source of literature though, variability results from an intrinsic and necessary property of our nervous system, arising from a multitude of sources – I shall refer to this as the *'intrinsic perspective'*. Both perspectives are discussed below in more detail.

The intuitive perspective

To understand the intuitive perspective, let us go back once more to the thought experiment on counting visitors in the museum, and ask again: What is it that makes your performance so different from the robot's – even after accounting for exogenous factors? Once more, this question may have an obvious answer: As a person, you experience a large array of 'internal' or 'metacognitive' states, such as fatigue, boredom, inattention, motivation, mood, sleepiness, alertness, even emptiness. You will also face distractions from your own thoughts; thinking about issues that do not relate at all to the task you are currently performing, such as about a conversation during lunch break, tasks you will need to perform later, or personal matters. Such 'task-unrelated thoughts' have been referred to as 'mind wandering'. On any day, you will experience meta-cognitive and off-task fluctuations in these states, and on some days, a particular state may be more dominant. The museum's robot has none of these states programmed in; its internal state remains constant throughout.

As both your performance and inner states fluctuate over time, it may be reasonable to assume they relate to each other. Empirical evidence comes from studies on mind wandering, which have found positive correlations between subjective experiences of 'being mentally off-task' and behavioural variability (Laflamme, Seli & Smilek, 2018; Seli, Cheyne & Smilek, 2013; Thomson, Seli, Besner & Smilek, 2014). Furthermore, it has been suggested that mindfulness meditation training – which is specifically aimed increasing 'on-task' over 'off-task' focus – leads to reduced variability (Brown & Ryan, 2003; Morrison, Goolsarran, Rogers & Jha, 2014; Mrazek, Franklin, Phillips, Baird & Schooler, 2013; Wells, 2005; Zeidan, Johnson, Diamond, David & Goolkasian, 2010). These findings seem to suggest that our metacognitive experiences indeed bear a relationship with the

fluctuations in our performance. Indeed, this view matches our typical intuitions about our performance. When it takes you a long time to press the clicker after a new visitor walked in, you may feel this happened because 'you were not paying attention'; because 'you could use a coffee'; because 'you were drifting off'; because 'your mind was wandering off' – all of which express the same sentiment that your performance was reduced due to suboptimal attention. Even more so, you may try to 'pay more attention from now on', in order to increase your performance.

However, the intuitive appeal of this perspective may also be dangerous: The relationship between attention and performance may be taken for granted, not because it *explains* the exact mechanistic origins of variability, but because it matches our folk psychological theories. Throughout this thesis, a recurring theme is how constructs like 'attention' and 'mind wandering' are easily invoked as explanations for variability – even when the exact underlying mechanisms remain unclear, concepts are ill-defined, and effect sizes are very small.

The intrinsic perspective

A second perspective, however, is that no matter how much attention we pay to our task, we will never even come close to the near-constant performance of the robot. Rather, the variability is an intrinsic property of our system – not just at a behavioural level, but at every level of our central nervous system. For instance, even at the level of a single neuron, random noise may contribute to whether the action potential will be fired (Ermentrout, Galán & Urban, 2008; Faisal, Selen & Wolpert, 2008). While this perspective seems pessimistic at first glance ("We will never be as good as the robot"), it should not necessarily be taken as such. Instead, from this perspective, variability may be a key component that allows our system to function. Furthermore, it may facilitate behaviours that the robot is not capable off, such as exploration and novel behaviour (Shahan & Chase, 2002; see Sternad, 2018 for a review).

As such, the fluctuations in performance may arise from fluctuations in a multitude of sources. These fluctuations may occur on varying time scales, both on shorter term (see Bompas, Sumner, Muthukumaraswamy, Singh & Gilchrist, 2015; Busch, Dubois & VanRullen, 2009; van Dijk, Schoffelen, Oostenveld & Jensen,

2008; Drewes & VanRullen, 2011; Ergenoglu et al., 2004; de Graaf et al., 2015; Hanslmayr et al., 2007; Rihs, Michel & Thut, 2007; Romei, Gross & Thut, 2010; Romei, Rihs, Brodbeck & Thut, 2008; Thut, Nietzel, Brandt & Pascual-Leone, 2006; VanRullen, Busch, Drewes & Dubois, 2011; for alpha, beta, and gamma oscillatory power in magneto- and electroencephalography; MEG/EEG), and on longer terms (see Weissman, Roberts, Visscher & Woldorff, 2006 for ~.05 Hz BOLD activity in the Default Mode Network; see Monto et al., 2008 for 0.01– 0.1 Hz EEG fluctuations). Furthermore, they may not just only relate to the brain, but also to other bodily states, such as heart rate (Salomon et al., 2016) and slow rhythms (~.05 Hz) in the stomach (Richter, Babo-Rebelo, Schwartz & Tallon-Baudry, 2017).

As these fluctuations appear from different brain and bodily states, and occur on multiple time scales, it is unclear to what extent the sources of variability are consciously accessible to us – and relatedly, it is unclear to what extent we may be able to reduce our variability at all.

Overview of the current thesis

In the current thesis, I aim to investigate the nature and origins of endogenous intra-individual variability in more detail. In particular, I will examine: 1) how variability manifests itself within and between individuals – or in other words, examine whether some individuals show variability that is closer to the robot's than others, whether this is a reliable, repeatable trait, 2) to what extent it relates to concepts as attention and mind wandering – or in other words, examine whether we indeed vary because we are affected by metacognitive states that the robot does not have, and 3) to what extent it may be at all reducible – or in other words, how closely we can match the robot's performance at all.

Overview of the chapters

In Chapter 1, I will examine endogenous variability in the oculomotor system in resting-state-based paradigms. This type of paradigm allows for the study of

variability in the absence of any changes in the external environment. As such, I will investigate how ‘pure’ endogenous variability manifests itself within individuals, and particularly, whether it constitutes a reliable individual trait over different points in time (repeatability) and over different conditions (generalisation). Furthermore, I examine to what extent these oculomotor variability measures are beneficial for distinguishing between different individuals based on self-assessed personality traits.

In Chapter 2, I will examine endogenous variability in a tapping-based task (the Metronome Task, Seli et al., 2013). The advantage of this task is that it requires only simple motor action that stays constant throughout all the trials. As such, it is useful in capturing endogenous variability specifically, while still providing a measure of performance on each trial. This will allow me not only to investigate the within-individual repeatability of endogenous variability, but also that of its temporal structures. Throughout the task, participants are also quasi-randomly enquired about their subjective attentional state. Likewise, I will examine the intra-individual repeatability of these states over time. Furthermore, I am interested to investigate to what extent the temporal structures can inform us about behaviour and attention, and whether they can be used to distinguish between self-assessed personality traits.

In Chapter 3, I will use MEG to examine the relationships between objective measures of performance, subjective reports of perceived performance, subjective reports of mind wandering, and preceding oscillatory power. Previous EEG studies on mind wandering and preceding neural states have been scarce and contradictory. These studies have largely not differentiated mind wandering from other forms of mental off-taskness (such as mind blanking) and have taken the meta-cognitive information of the attentional state ratings for granted. I will introduce a second subjective rating, on which participants are asked to rate their performance – to get an idea of how different metacognitive ratings may compare to each other – and will ask them to classify their off-taskness states. Importantly, prior studies have typically been conducted from the intuitive perspective, and have assumed that behavioural variability and subjective attentional state reflect highly similar (if not identical) underlying processes. As such, they investigate the links between mind wandering and performance as well as between mind wandering and neural

states, but not the links between the three. I will specifically focus on this three-way link, aiming to investigate to what extent subjective attentional state ratings and behavioural variability are truly related to the same underlying processes.

The last experimental chapter is focused on to what extent variability can be reduced with more task-control. To draw a parallel, imagine that you agree with your boss that visitors can only enter the special collection space when you are feeling fully 'on-task'. From the intuitive perspective, you have the means to access this information, and the agreement should subsequently make your performance much more similar to the robot's near-constant performance. From the intrinsic perspective, this agreement may not make much difference to your performance. While this example may sound nonsensical within our current thought experiment, Chapter 3 shows some practical implementations of this type of control in certain sports (e.g., darts, shooting). I will then test whether such control can be used to reduce variability both in a darts-based and in two computer-based tasks. As such, this chapter allows me to test the intuitive and intrinsic perspective against each other.

Overview of the experimental data

I will investigate the above-described questions with data from four different experiments. For presentation purposes, the order in and combination with which the experiments were conducted are not identical to how I describe in the current thesis. Here I will give a short overview of the conducted experiments, in the order in which they were performed (see Figure 2 for full overview).

For the first experiment ('Control experiment'), I tested 39 participants in a four-day experiment to test the effect of task-control upon variability (36 in action-oriented task; 39 in perception-oriented task). This data has been used for 'Experiment 2' in Chapter 4. Of these participants, 28 also participated in an oculomotor resting-state paradigm. Before each of the four behavioural sessions, their eye movements and pupil dilation were recorded for one minute while: 1) fixating on a dot in the centre of the screen, 2) fixating in the centre of the screen without a fixation dot, and 3) looking at the screen without specific instructions. At the end of the four behavioural sessions, these participants also filled in

questionnaires on ADHD, mind wandering, and impulsivity. The oculomotor data and questionnaires have been used for ‘Experiment 3’ in Chapter 1.

The second experiment (‘Darts experiment’) concerns the darts experiment (Experiment 1 in Chapter 4) and was intended as a follow-up on the psychophysics experiments, to verify that people also cannot act upon their internal performance-relevant states when they are highly motivated to do so.

For the third experiment (‘Mind wandering experiment’), 84 participants were tested in a single session. All of them started with an oculomotor resting-state task (continuous fixation for four minutes). Next, they performed the Metronome task for ~25 minutes. Straight after the task, they repeated the oculomotor resting-state task. Finally, they completed a number of questionnaires on ADHD tendencies, mind wandering, impulsivity, depression, anxiety, mindfulness, mood, and schizotypy. Of these participants, 25 then repeated the ~25 minutes Metronome Task.

Part of these questionnaires (ADHD, mind wandering, and impulsivity) as well as the oculomotor data have been used for ‘Experiment 1’ in Chapter 1. The same questionnaires plus the behavioural data have been used for Chapter 2. The other questionnaires have been analysed for final year projects from BSc students.

For the last experiment (‘MEG experiment’), 21 participants were tested in a two-day experiment in the MEG. On both days, they first took part in a resting-state paradigm (four minutes of fixation, measuring both oculomotor and MEG activity). Next, they did the Metronome Task for ~50 minutes. Straight after, they repeated the resting-state. At the end of the second day, they completed questionnaires on ADHD tendencies, mind wandering, and impulsivity. The oculomotor data and questionnaires have been used in ‘Experiment 2’ of Chapter 1. The behavioural and MEG data have been used in Chapter 3.

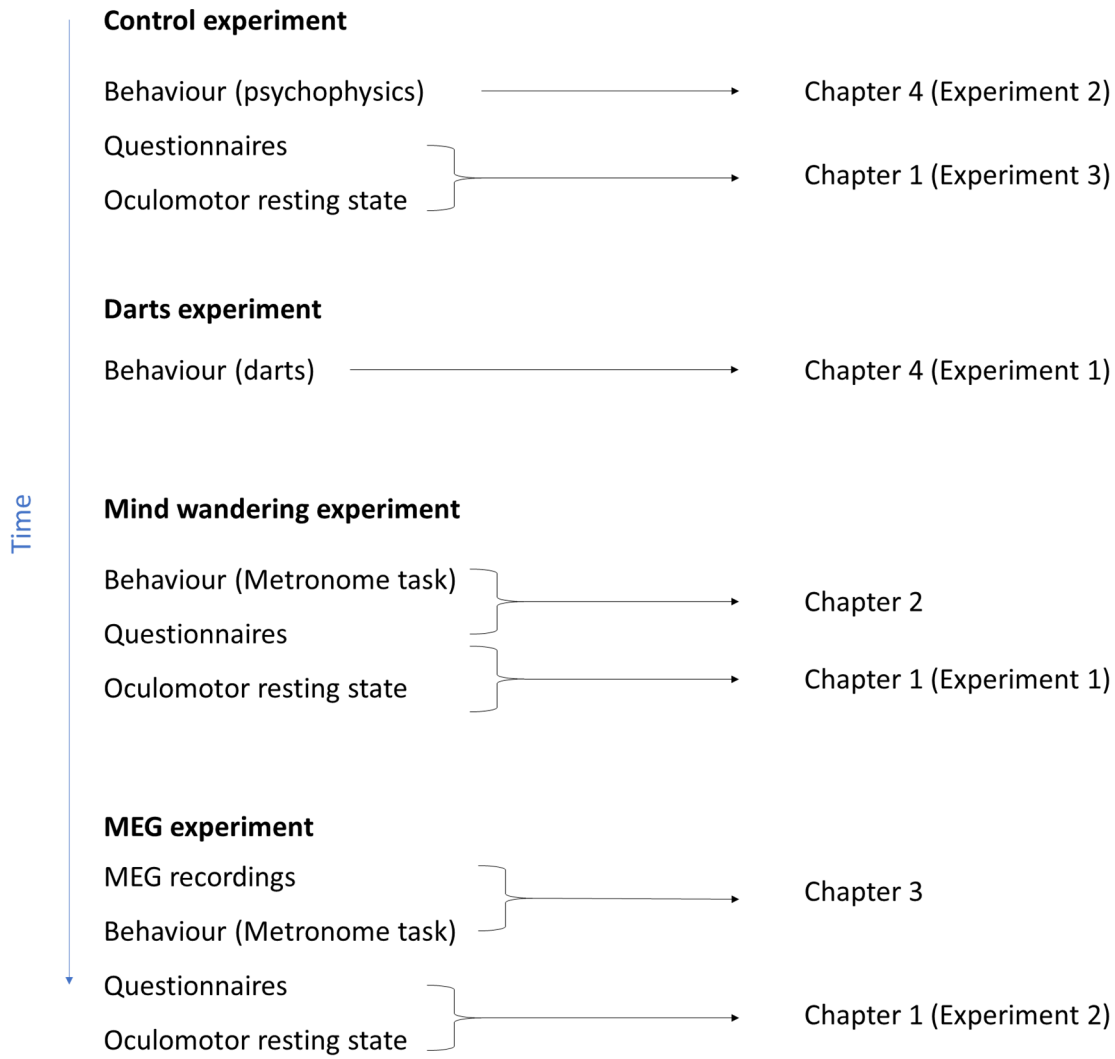


Figure 2. Timeline of the conducted experiments (in chronological order) on the left, and the corresponding chapter (plus experiment) in which their methods and results have been described.

Chapter 1

Reliability and correlates of intra-individual variability in the oculomotor system

Abstract

Even if all external circumstances are kept equal, the oculomotor system shows intra-individual variability over time, affecting measures such as microsaccade rate, blink rate, pupil size, and gaze position. Recently, some of these measures have been associated with ADHD on a between-subject level. However, it remains unclear to what extent these measures constitute stable individual traits. In the current study, we investigate the intra-individual reliability of these oculomotor features. Combining results over three experiments (> 100 healthy participants), we find that most measures show good intra-individual reliability over different time points (repeatability) as well as over different conditions (generalisation). However, we find evidence against any correlation with self-assessed ADHD tendencies, mind wandering, and impulsivity. As such, the oculomotor system shows reliable intra-individual reliability, but its benefit for investigating self-assessed individual differences in healthy subjects remains unclear. With our results, we highlight the importance of reliability and statistical power when studying between-subject differences.

Keywords: Eye movement; eye tracking; microsaccades; gaze; attention; reliability; intra-individual variability; individual differences; ADHD, mind wandering

Introduction

Imagine that you are working in your office, and one of your colleagues suddenly walks in: Your eyes will immediately change position from your work and will subsequently fixate on your colleague, and your pupil size will be modulated by the differences in light hitting your eye. These types of changes in eye position and pupil size may be described as 'exogenous' intra-individual variability – variability within an individual over time that is brought about by changes in the external environment. However, even when all external circumstances remain the same and one is solely fixating on a static dot, the eyes are still not 'perfectly stable'. Instead, there will still be small changes in eye position (i.e., 'fixational eye movements', see Rolfs, 2009 for a review) and in pupil size, as well as blinks. All of these changes may be described as 'endogenous' intra-individual variability – brought about by internal fluctuations.

It seems reasonable that endogenous variability differs between individuals. Supporting this, a recent paper found positive associations between endogenous oculomotor variability and Attention-Deficit and/or Hyperactivity Disorder (ADHD) tendencies (Panagiotidi, Overton & Stafford, 2017). However, the intra-individual reliability of endogenous oculomotor variability is still largely unknown – meaning it is unclear to what extent this variability constitutes a reliable individual trait. While there is evidence for intra-individual reliability in oculomotor measures over different types of tasks (Andrews & Coppola, 1999; Boot et al., 2009; Castelhano & Henderson, 2008; Poynter et al., 2013; Rayner et al., 2007), the reliability of oculomotor variability during rest has not been investigated. Such reliability is an important quality for any potential biomarker (Mayeux, 2004). The aim of the current paper is therefore twofold. First, we aim to examine whether variability in the

oculomotor system shows reliable intra-individual consistency over different time points and different conditions in 'resting state' circumstances, to investigate to what extent this endogenous variability may be a reliable individual property. Secondly, we aim to investigate potential interindividual differences, by testing whether oculomotor variability correlates with ADHD, mind wandering, and impulsivity.

Oculomotor functioning and variability

While 'saccades' refer to sudden, ballistic movements in eye position and 'fixations' refer to the maintenance of the eye position on a particular spot, microsaccades refer to small, sudden movements of the eye position during fixations (see Rolfs, 2009 for a review). Microsaccades are one of three types of fixational eye movements, the other two being drift and tremor. The movements of microsaccades have been described as 'jerk-like', small (typically being below 1-2° in amplitude), often as 'binocular' (i.e., occurring in both eyes simultaneously). Extracted microsaccades are characterised by a 'main sequence' – a very strong linear correlation between saccade amplitude and velocity across all extracted saccades. There have been several suggestions about the purpose of microsaccades (and fixational eye movements in general), relating to control over fixation position, prevention of perceptual fading, improvement of visual processing, (small-area) scanning of the environment, and acuity (see Rolfs, 2009; Martinez-Conde, Otero-Millan & Macknik, 2013 for reviews).

While microsaccades have been related to attention, this refers mostly to attentional cuing and 'covert attention' (i.e., foci of attention that are separate from the current eye position). Attentional cuing has been known to modulate both the direction and the occurrence of microsaccades, with the latter most commonly following a shape known as the 'microsaccade rate signature' – showing a sudden drop in microsaccades after cue onset, followed by a strong increase right after. Interestingly, this modulation of microsaccade rate seems influenceable by top-down expectations (Valsecchi, Betta & Turatto, 2007). However, the role of attentional cuing relates to exogenous variability, not to the manifestation of variability during rest – which would instead be related to fluctuations in internal states over time.

Intra-individual stability of oculomotor variability has been shown previously over different types of tasks, images, and display modalities (Andrews & Coppola, 1999; Boot et al., 2009; Castelhano & Henderson, 2008; Poynter et al., 2013; Rayner et al., 2007). For example, Castelhano & Henderson (2008) found consistency in individuals' oculomotor behaviour between images in different display formats, but also between faces and scenes. In these cases, it seems plausible that the intra-individual consistency appears because of individual consistency in viewing and processing information; individuals may have a 'default way' of information processing that is reflected in their oculomotor behaviour. This is supported by findings that oculomotor behaviour can be altered when participants are given different instructions and feedback (Boot et al., 2009), and explains why it shows cultural differences (Rayner et al., 2007).

The intra-individual correlations of endogenous variability have furthermore been studied in relation to task-based variability. Andrews and Coppola (1999) looked at fixation duration and saccade size across five conditions: a 'dark room' condition, in which participants' eye movements were continuously recorded for 100 seconds, two 'viewing' conditions, in which participants viewed simple and complex patterns, and two more 'cognitive' tasks, in which participants did visual search and reading. Oculomotor measures in the dark room condition showed positive intra-individual correlations to the viewing conditions, but not to cognitive conditions. Poynter et al. (2013) used a larger array of measures: For each participant, they extracted six measures of oculomotor activity (saccade amplitude, microsaccade rate and amplitude, and fixation rate, duration, and size) over four different tasks (a sustained fixation, scan-identify, search, and Stroop task). They found that each oculomotor measure correlated to itself between the different tasks within participants. However, their fixation task consisted of trials that were only three seconds long – meaning that the variability is still highly dependent on stimulus-onset, and that the task is not aimed at capturing (mostly) endogenous variability. Overall, none of these articles address the question of the current research directly – namely, to what extent endogenous variability itself is a reliable individual trait.

Oculomotor variability and ADHD symptomatology

In the search for a potential 'biomarker' of ADHD, two previous studies investigated the relation between ADHD and oculomotor variability. Fried et al. (2014) examined differences between adults with ADHD (both in an 'unmedicated' and 'medicated' session) and healthy controls (unmedicated in both sessions). Participants were asked to make a button press in response to targets but not to non-targets. In their 'unmedicated' session, participants with ADHD showed significantly higher microsaccade and blink rates compared to controls, both near stimulus onset and throughout the entire trial. However, these differences were not found in the 'medicated' session. No significant differences were found in pupil size mean or variability in either session. Panagiotidi et al. (2017) found a similar positive association between microsaccade rate and self-assessed ADHD tendencies within a healthy population, but did not investigate pupil size or blink rate.

It is important to note that these studies differ in a significant way. Fried et al. (2014) focused on task-based differences, which arise partly from external circumstances. Healthy controls were able to fixate before target onset, meaning that they were able to control their eye movements to some extent when this was relevant for the task. Those control participants showed a large increase in blinks and microsaccades only after the target has disappeared from the screen. ADHD patients showed deficiencies in this functionality, which was accompanied with decreased task performance. However, Panagiotidi et al. (2017) took a more resting state-based approach, using 20 trials of 20 seconds each, in which participants were asked to fixate on a cross, without any additional task or stimuli. This type of paradigm, in which all circumstances are kept equal, captures mostly endogenous variability by default.

It may be tempting to attribute the observed effects to individual differences in a general concept of 'attention'. However, in the paradigms of Fried et al. (2014) and Panagiotidi et al. (2017), 'attention' may manifest in different ways – the latter relates to internal fluctuations over time, while the former paradigm makes use of covert attention. As described in the above section on *Oculomotor functioning and variability*, these reflect distinct phenomena, and as such, they may not necessarily have similar outcomes.

This is of importance, because ADHD may affect a multiplicity of neuropsychological domains, which means that even when certain behavioural deficiencies or differences are found, it can be hard to pinpoint through which mechanism(s) these arise. While Panagiotidi et al. (2017) did use an ADHD questionnaire with two subscales – Inattention and Impulsivity/Hyperactivity, reflecting the two main subtypes of ADHD – they only analysed the total scores, because both scales correlated highly to the total scores (r-values of .81 and .90 respectively). However, these types of high correlations between subscales and total scores are to be expected, as questionnaires tend to measure one construct, and the total scores reflect nothing more than the sum of the parts. As the correlation between the subscales was only moderate (r = .46), the subscales show sufficient non-shared variance (78.8%) to investigate their separate contributions. Analysing the subscales separately may still reveal potential differences between them, particularly when it is unclear what exact mechanism causes the correlation.

Current research

In the current research, we examine the resting state paradigm for eye movements in more detail, to see if it produces reliable markers within individuals over different time points (repeatability) and over different conditions (generalisation). In particular, we will be looking at microsaccade rate, pupil size variability, blink rate, and gaze variability. We also aim to further explore the relationship of oculomotor variability to self-assessed ADHD symptomatology.

Impulsivity is one of the main characteristics of ADHD, and previous literature has associated self-assessed ADHD tendencies with impulsivity (Berg, Latzman, Bliwise & Lilienfield, 2015; Miller, Derefinko, Lynam, Milich & Fillmore, 2010; although some facets of impulsivity may be more important than others). ADHD has also been associated with increased mind wandering (Shaw & Giambra, 1993; Seli, Smallwood, Cheyne & Smilek, 2015). Shaw and Giambra (1993) furthermore investigated mind wandering in undiagnosed college students and found that participants who scored in the lowest tier of self-assessed ADHD symptoms during childhood were less prone to mind wandering than participants who scored in the highest tier. Possibly, this reflects a decreased tendency to keep top-down focus

with increased ADHD tendencies. To get further insight into the mechanisms underlying potential individual differences in oculomotor variability, we therefore also included self-assessed measures of mind wandering and impulsivity. We aim to replicate positive associations of these two measures with self-assessed ADHD, as well as investigate their relationship to oculomotor variability.

Figure 1 show an overview of our three aims. To examine these aims, we combined data of three behavioural experiments (129 participants in total). In Experiment 1 and 2, participants took part in a four minutes long resting state paradigm and repeated this half an hour to 50 minutes later after they completed a computerised task. In Experiment 3, participants took part in a resting state paradigm in three different conditions repeatedly over four different days, with each resting state being one minute long. In all three experiments, participants filled in questionnaires on ADHD tendencies, mind wandering tendencies, and impulsivity. This allowed us to investigate: 1) the intra-individual reliability of oculomotor behaviour, 2) the between-subject correlations on the questionnaires, and 3) the between-subject correlations between oculomotor behaviour and the questionnaires. Because the predictions for all three questions are highly similar across experiments, they are discussed together below, and analyses were combined whenever possible.

Aim 1. Intra-individual reliability of oculomotor behaviour

If variability of oculomotor functioning is to make a good marker for personality traits, it should show reliability within individuals. We therefore examined the intra-individual reliability of the markers (variability in gaze, pupil size variability, and blink rate in all three experiments, plus microsaccade rate in Experiments 1 and 2) over different points in time on the same day (Experiment 1 and 2) and over different conditions and different days (Experiment 3).

To examine this, mean scores were calculated on each of the different measures for every participant, separately for each resting state (reflecting time/condition). If a measure shows intra-individual reliability, it should correlate

highly with itself over the different resting states. Because of the differences in design, the intra-individual reliability was examined separately for each experiment.

Aim 2. Between-subject correlations between ADHD, mind wandering, and impulsivity

After completing the resting state paradigms, participants filled in questionnaires on ADHD, mind wandering, and impulsivity. Based on previous literature, we would expect a positive correlation between self-assessed ADHD tendencies and self-assessed mind wandering (Shaw & Giambra, 1993; Seli et al., 2015). Furthermore, we expect a positive correlation between ADHD and impulsivity, similarly based on previous literature (Berg et al., 2015; Miller et al., 2010). Data were combined for all three experiments.

Aim 3. Between-subject correlations between questionnaires and oculomotor behaviour

Next, one overall mean was calculated for every participant, separately for each oculomotor measure, collapsed over all time points and conditions. These means were correlated to the ADHD scores, to test if ADHD tendencies are associated with higher oculomotor variability. Furthermore, the scores were correlated to the two subscales of the ADHD questionnaire scores (Inattention and Impulsivity/Hyperactivity), as well as to the impulsivity and mind wandering questionnaire scores. The correlations were calculated on the combined data from all three experiments.

If a potential relationship between ADHD and oculomotor variability is caused mainly by a lack of attentional task maintenance, one could expect similar correlations of oculomotor variability to mind wandering and inattention. Alternatively, higher correlations to impulsivity and hyperactivity may reflect that the relationship is driven by a lack of inhibition.

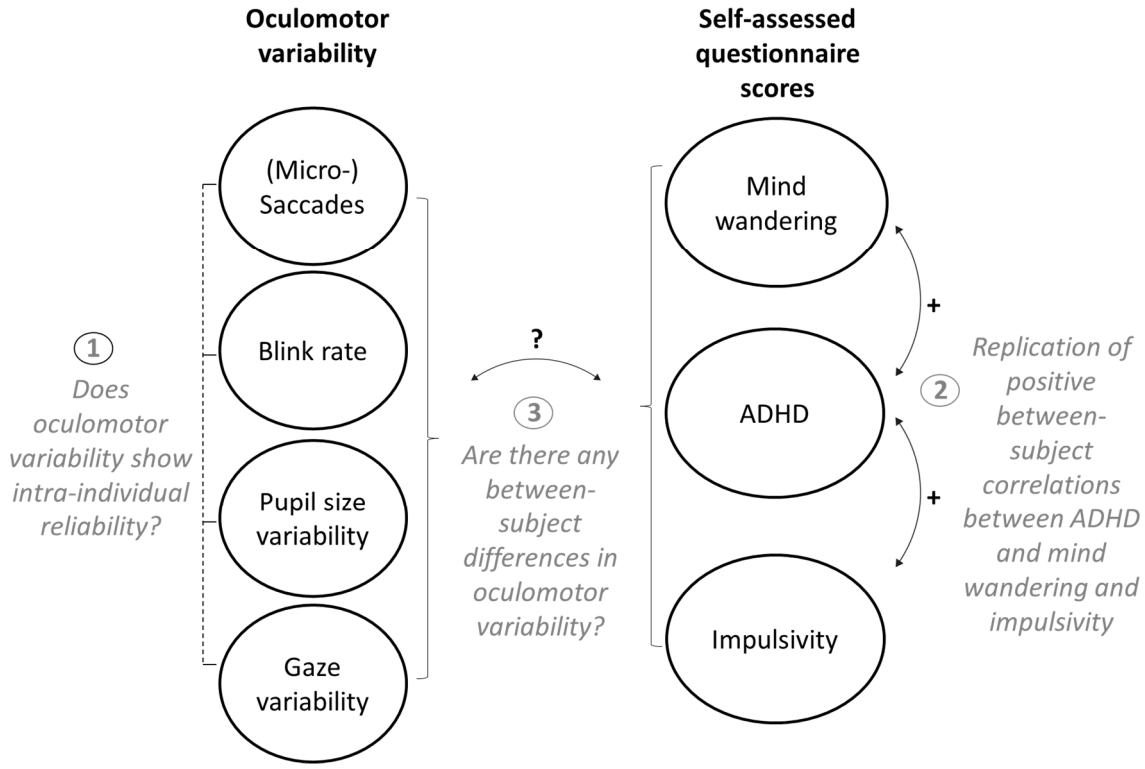


Figure 1. Graphical representation of the oculomotor measures and self-assessed personality traits, with the three aims of the current study.

Methods

Participants

In total, data of 129 participants was collected. All of them had normal or corrected-to-normal vision. The studies were approved by the local ethics commission.

Experiment 1

Eighty-one participants (66 female, fourteen male, one other, aged between 18-25) contributed in exchange of course credits. Of them, 73 had valid eye tracking data. For three of these remaining 73, the second session was not included because they had more than 33% missing samples.

Experiment 2

Twenty-one participants (eighteen female, 21-40 old, M_{age} = 26.3) contributed in exchange of a monetary reward. All had valid eye-tracking data. Two of them only took part in one test day, due to technical issues. For another three participants, the second session on the first day was excluded, and for one participant, the second session of the second day was excluded, because more than 33% samples were missing.

Experiment 3

Twenty-eight participants (eighteen female, 18-36 years old, M_{age} = 25.5) contributed in exchange of a monetary reward, and twenty-six of them had valid eye tracking data. Of these twenty-six participants, one participant had only three sessions, and another had only two sessions. Furthermore, another eleven (out of 303 remaining) sessions from five different participants were not included because more than 33% missing samples were missing.

Materials

The resting state paradigms were generated with MATLAB (The Mathworks, Inc.) and Psychtoolbox-3 (Brainard, 1997; Kleiner, Brainard & Pelli, 2007; Pelli, 1997). The background of the paradigms was set at light-grey, and the fixation point was white. An Eyelink 1000 (SR Research) was used in each of the experiments for eye data recording. Each experiment started with calibration and validation with the eye tracker (five-dot calibration in Experiment 1, nine-dot calibration in Experiment 2 and 3). Participants were seated in a chinrest to limit head movement.

The Adult ADHD Self-Report Scale (ASRS-v1.1; Kessler et al., 2005) was administered to measure ADHD tendencies. The ASRS-v1.1 consists of 18 items with a 5-point scale from 0 (*"Never")* to 4 (*"Very often")* and has a high reliability (with Cronbach's α ranging from .88 to .94; Adler et al., 2006; 2012). The ASRS-v1.1 can be divided into two subscales – Inattention and Hyperactivity / impulsivity

- reflecting the two main subtypes of ADHD (Kessler et al., 2005; Reuter, Kirsch & Hennig, 2006).

Furthermore, the Daydreaming Frequency Scale (DFS; Singer & Antrobus, 1963) was administered to measure mind wandering in daily life. The DFS is a subscale of the Imaginal Processes Inventory and measures the amount of daydreaming and off-task mind wandering in daily life. It consists of 12 items, each with a 5-point scale. It has a high internal consistency (Cronbach's α = .91) and a high test-retest reliability (.76 with an interval of maximum one year; Giambra, 1980).

To measure impulsivity, participants completed the UPPS-P Impulsive Behaviour Scale (Lynam, Smith, Whiteside & Cyders, 2006; Whiteside & Lynam, 2001). The UPPS-P consists of 59 items, with a scale ranging from 1 (*"agree strongly"*) to 4 (*"disagree strongly"*), divided over five subscales: positive urgency, negative urgency, (lack of) premeditation, (lack of) perseverance, and sensation seeking.

Experiment 1

The paradigms were generated with a Viglen Genie PC and displayed on an ASUS VG248 monitor with a resolution of 1920 by 1080 and a refresh rate of 144 Hz. Eye movements and pupil dilation were recorded binocularly at 500 Hz.

Experiment 2

The resting state paradigms were generated on a HP Z230 Workstation PC and an LG 24GM77 monitor with a resolution of 1920 by 1080 and a refresh rate of 120 Hz. The paradigms were displayed on a projector screen. Eye movements and pupil dilation were recorded binocularly at 500 Hz.

Experiment 3

The resting state paradigms were generated with a Bits# Stimulus Processor video-graphic card (Cambridge Research Systems) and a Viglen VIG80S PC, and were

displayed on an hp p1230 monitor with a resolution of 1280 by 1024 and a refresh rate of 85Hz. Eye movements and pupil dilation were recorded monocularly at 1000 Hz.

Design

Experiment 1 and 2

Resting state eye movements and pupil dilation were recorded before and after a behavioural task – see Figure 2 for an overview. This gave (2 x 4) 8 minutes of resting state eye measures in total for each participant. ADHD tendencies, mind wandering tendencies, and impulsivity characteristics in daily life were measured with questionnaires.

Experiment 3

Resting state eye movements and pupil dilation were recorded in three different condition – see Figure 2 for an overview. In the 'Fixation plus instruction'-condition, participants were asked to fixate on a fixation dot that was displayed on the centre of the screen. In the 'No fixation, Instruction only'-condition, participants were shown a blank screen, and were asked to fixate on the centre of the screen. In the 'No fixation plus no instruction'-condition, participants were also shown a blank screen, but they were only asked to not turn away from the screen, with no further fixation-related instructions. This procedure was repeated over four days – resulting in (1 x 3 x 4) 12 minutes of resting state measures for each participant in total. ADHD tendencies, mind wandering tendencies, and impulsivity characteristics in daily life were measured with questionnaires. Again, ADHD tendencies, mind wandering tendencies, and impulsivity characteristics in daily life were measured with questionnaires.

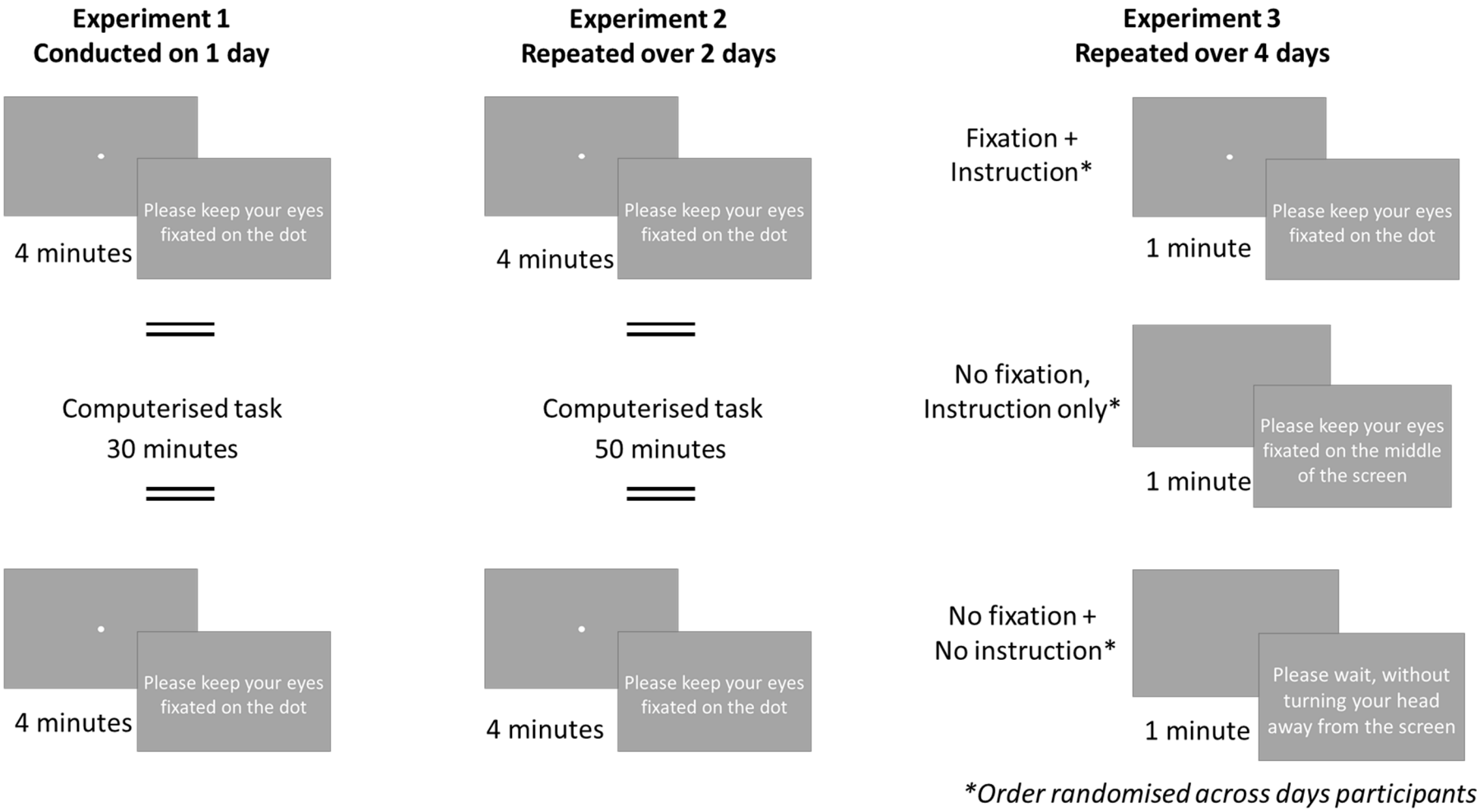


Figure 2. Overview of the resting state eye movement paradigms of all three experiments.

Procedure

Experiment 1

Participants came to the lab for a session of about 1.5 hours. They were seated at a distance of 615 cm from the screen. Eyes were tracked binocularly during the resting state for four minutes (*time 1*). Next, participants performed a computerised task, lasting about 30 minutes (data not analysed in the current paper). Right after finishing this task, the resting state paradigm was conducted again (*time 2*). Lastly, participants filled in nine questionnaires: the DFS, ASRS-v1.1, and UPPS-P, as well as the Beck Anxiety Inventory Second edition (Beck & Steer, 1993), Beck Depression Inventory Second edition (Beck, Steer & Brown, 1996), Short form Wisconsin Schizotypy scales (Winterstein et al., 2011), Five-facet Mindfulness Questionnaire (Baer, Smith, Hopkins, Krietemeyer & Toney, 2008), Toronto mindfulness scale (Lau et al., 2006), and Positive and Negative Affect Schedule (Watson, Clark & Tellegen, 1988). Only the first three questionnaires were analysed in the current study.

Experiment 2

Participants came to the lab for two sessions, each about 1.5 hours. They were seated at a distance of 1185 cm to the screen. Eyes were tracked binocularly for four minutes (*time 1*). Next, they performed a computerised task of about 50 minutes (data not analysed in the current paper), and afterwards they conducted the resting state paradigm again (*time 2*). Lastly, participants completed the DFS, ASRS-v1.1, and UPPS-P.

Experiment 3

The experiment consisted of four sessions of about an hour. Participants were seated at a distance of 104 cm to the screen. Eyes were tracked monocularly in the three different conditions. Each condition lasted 60 seconds. Instructions were shown for two seconds. For each participant, the order of the conditions was random

on each of the four sessions. After completing the resting state eye movements paradigm, participants completed a 30 to 45 minutes computerised task (data not analysed in the current paper). On the last day, they filled in the DFS, ASRS-v1.1, and UPPS-P.

Data preparation and analysis

Oculomotor measures

Blinks were defined as missing tracking data, with a maximum of 1000 ms. The total number of blinks throughout each session was counted, and a blink rate per second was subsequently calculated. Pupil size variability was calculated by dividing the standard deviation of the pupil size throughout each session by the mean pupil size – reflecting the coefficient of variation (CV). Gaze variability was calculated separately for the x- and y-screen dimension by calculating the standard deviation of position in degrees throughout the entire session (these standard deviations were not normalised by the mean, as the mean degrees in the middle of the screen is approximately zero). To minimise noise, 20 ms were excluded both before and after missing samples from the calculation of the pupil size and gaze variability.

Binocular microsaccade detection (Experiment 1 and 2 only) was done with the algorithm of Engbert and Kliegl (2003), using the Microsaccade Toolbox for R (Engbert, Mergenthaler & Trukenbrod, 2015). The λ value was set to five. To reduce noise in the detection process, saccades were defined as being at least three samples long. Furthermore, a period of 100 ms both prior and following blinks was excluded. Missing/excluded samples were subsequently interpolated. To avoid the false detection of post-saccadic oscillations as microsaccades, a window of 20 ms following each saccade was excluded. Saccades with amplitudes above 2° or with peak velocities above 200°/s were excluded from subsequent analyses. To sanity check the microsaccades, saccade amplitude was correlated with velocity over all participants and over both time points (i.e., main sequence). These were highly correlated to each other for both Experiment 1 ($r = .88$, $BF_{10} = \infty$, $p < .001$) and for Experiment 2 ($r = .86$, $BF_{10} = \infty$, $p < .001$). The mean microsaccade rate was 1.1 per

second (SD = .43) for Experiment 1 and 1.58 (SD = .47) for Experiment 2, which is within the typical rate of 1-2 per second (Ciuffreda & Tannen, 1995).

Distributions of the oculomotor measures were highly skewed on the group level. This may bias the results of the correlation analyses, particularly for Experiment 2 and 3, which have smaller sample sizes. For consistency, all analyses were conducted on the natural logarithm of the measures.

All Bayesian statistics throughout the current research were conducted in JASP (JASP Team, 2017), using the default options of equal prior probabilities for each model and 10000 Monte Carlo simulation iterations.

Questionnaires

Scores on items of the questionnaires were reversed when necessary. Missing responses were substituted with the median (but note that the number of missing responses was neglectable, .26%). Next, the total score was calculated for each of questionnaire. Individual item scores were used to check the questionnaires internal consistency (Cronbach's α; Cronbach, 1951) – see Table 1 for an overview.

Table 1. Overview of the Daydreaming Frequency Scale (DFS), the Adult ADHD Self-Report Scale (ASRS-v1.1), and the UPPS-P Impulsive Behaviour Scale (UPPS-P). Shown are the mean scores and standard deviations (SD) over all the participants, as well as the internal consistency (Cronbach's α) for each questionnaire, for each sample separately as well as for the combined data. Also shown are the minimum and maximum possible scores of each questionnaire.

Questionnaire	Sample	Mean score	SD	Cronbach's α	Possible range
DFS	Exp 1	39.3	9.8	.93	12-60
	Exp 2	39.3	8.7	.92	
	Exp 3	37.7	9.1	.93	
	Combined	39.0	9.4	.92	
ASRS-v1.1	Exp 1	33.4	8.5	.81	0-72
	Exp 2	28.5	5.7	.62	
	Exp 3	25.5	7.3	.76	
	Combined	30.6	8.6	.89	
UPPS-P	Exp 1	138.8	23.9	.93	59-236
	Exp 2	119.7	20.3	.93	
	Exp 3	122.8	18.6	.68	
	Combined	132.3	23.7	.92	

Results aim 1. Intra-individual reliability of oculomotor variability measures

Experiment 1. Reliability over time

Two means were calculated for each measure (microsaccade rate, blink rate, pupil size variability, gaze-x variability, and gaze-y variability): One for time point 1 (pre-task) and one for time point 2 (post-task). Bayesian Pearson pairs were then conducted on each of the measures to test intra-individual reliability over time. Figure 3 shows the within-subject correlational plots over the two time points for the logged measures of gaze variability in the horizontal and vertical dimension, pupil size variability, blink rate, and microsaccade rate – with correlation coefficients and logged Bayes Factors (BF_{10}) on top.

The BF_{10} reflect the likelihood of the data for the alternative hypothesis (in this case, the presence of a correlation) over the null-hypothesis (in this case, the absence of a correlation), and can take a value between zero to infinity.[1] For example, for gaze variability in the horizontal dimension, the $\log(BF_{10})$ between time 1 and 2 is 17.7 – meaning that the likelihood of the data is ($\exp(17.7)$ =) 48642102 times larger under the alternative than under the null-hypothesis. This can be interpreted as extremely high evidence for the presence over the absence of a correlation between the two time points. The other four measures show similarly extreme Bayes Factors. Each of the measures show high and positive r-values, indicating that they show intra-individual consistency. Thus, oculomotor shows reliability when measured half an hour apart.

[1] Note that BF01 (null over alternative hypothesis) can be derived from BF10 (alternative over null) by taking its inverse.

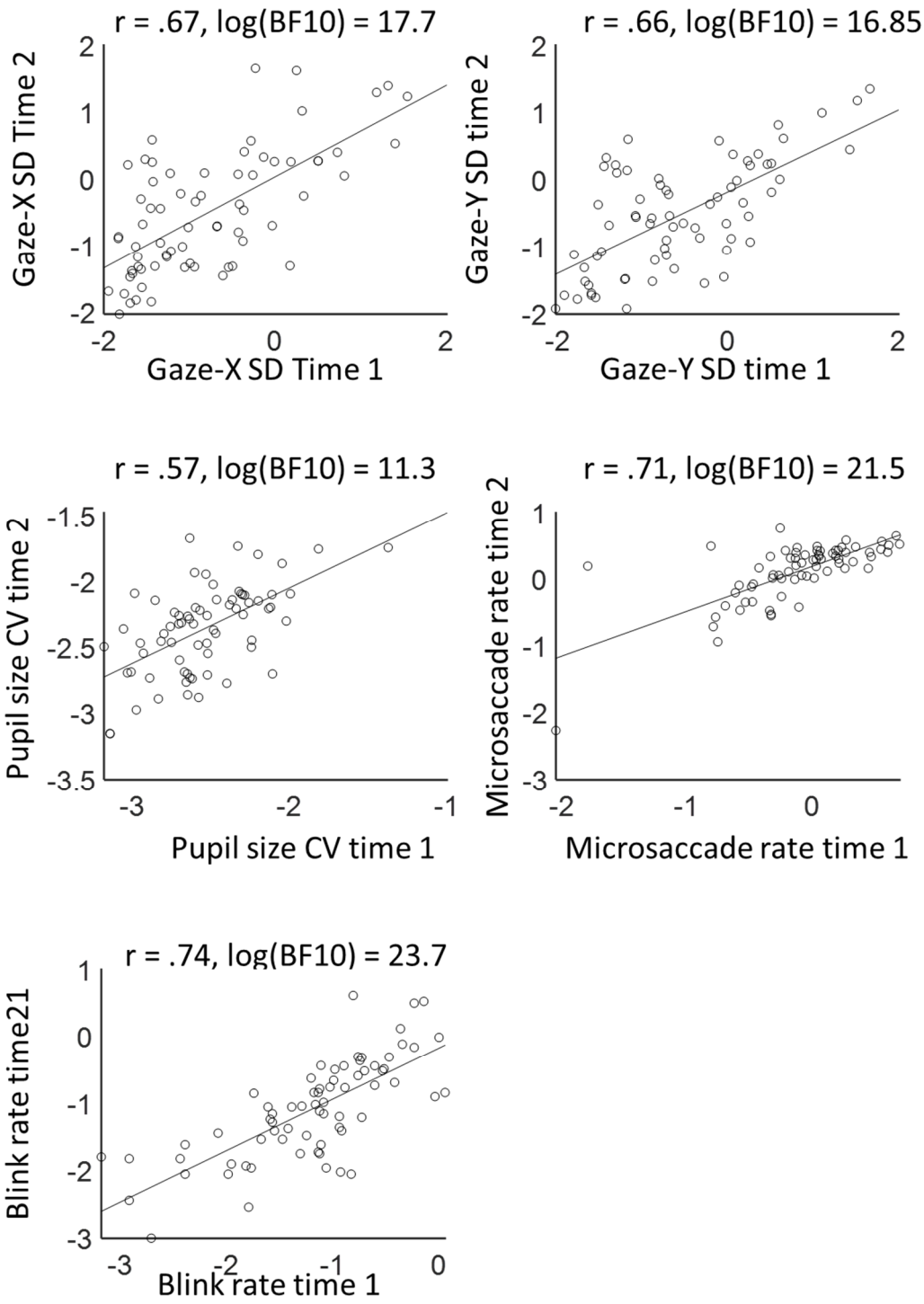


Figure 3. Correlations between time point 1 (pre-task) and time point 2 (post-task) for each of the five oculomotor measures from Experiment 1: Gaze variability (standard deviation; SD) in the horizontal dimension, gaze SD in the vertical dimension, pupil size coefficient of variability (CV), blink rate per second, and microsaccade rate per second (Ms). All five measures show a high correlation coefficient and accompanying high Bayes Factor, indicating that the measures show

intra-individual reliability over time. Note that both the measures and the Bayes Factors are logged.

Experiment 2. Reliability over time and days

Combined, Experiments 2 and 3 have 21 correlation pairs for each oculomotor measure, each testing the reliability over different time points and days. Rather than having to plot each correlation separately and then trying to assess the global patterns, the distributions of these correlations are shown in violin plots (Figure 4). This way of representing the data allows for an immediate overall picture of the correlations. The vertical dimension of these violin plots indicates the entire range of correlation coefficients (top panel) and accompanying Bayes Factors (bottom panel), while the horizontal dimension indicates the density. Each condition is also plotted (coloured triangles and asterisks), with the white dot representing the median value.

To test the intra-individual reliability over time in Experiment 2, four means were calculated for each measure (microsaccade rate, blink rate, pupil size variability, gaze-x variability, and gaze-y variability): One for time 1 (pre-task) and one for time 2 (post-task), both for day 1 and day 2. For both days, Bayesian Pearson pairs were conducted between time 1 and time 2 on each measure – giving two replications of the analysis of Experiment 1 (shown in Figure 4 in light-blue triangles). Again, we found evidence in favour of correlations between time 1 and 2 for pupil size variability, blink rate, and microsaccade rate (with all six BF_{10} above 1, and only one of them in the indeterminate range), with corresponding r-values all being moderate to high. These findings again indicate good intra-individual reliability of the measures – especially when considering the much smaller sample size of this experiment. These results replicate the findings from Experiment 1 with almost twice as much time in between the two time points. However, we no longer found evidence for intra-individual reliability in gaze variability, especially in the horizontal dimension: All four BF_{10} were in the indeterminate range, with three of them being below 1.

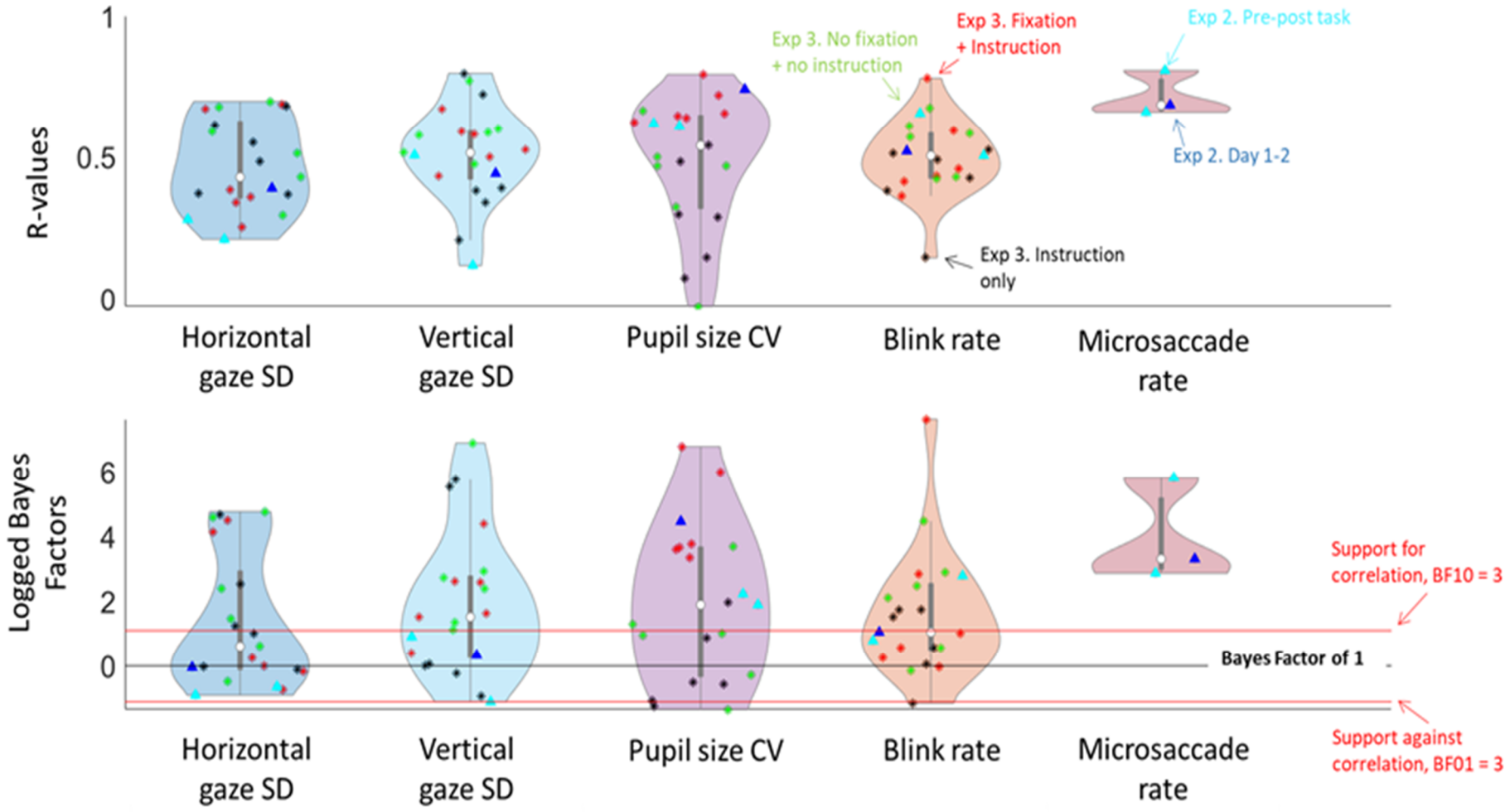

R-values
1
0.5
0
Exp 3. No fixation + no instruction
Exp 3. Fixation + Instruction
Exp 2. Pre-post task
Exp 2. Day 1-2
Exp 3. Instruction only
Horizontal gaze SD
Vertical gaze SD
Pupil size CV
Blink rate
Microsaccade rate
Logged Bayes Factors
6
4
2
0
Support for correlation, BF10 = 3
Bayes Factor of 1
Support against correlation, BF01 = 3
Horizontal gaze SD
Vertical gaze SD
Pupil size CV
Blink rate
Microsaccade rate

Figure 4. Distributions of the correlation coefficients (top panel) and accompanying logged Bayes Factors (bottom panel) of the correlation analyses on within-subject reliability for each of the five oculomotor measures. Values denoted with a triangle represent the correlations for Experiment 2, with light-blue triangles representing the correlations between different time points (pre and post task), and dark-blue triangles representing the correlation between days. Values denoted with an asterisk represent the correlations for Experiment 3, with red, black, and green representing the different conditions ('Fixation plus instruction', 'No fixation, instruction only', and 'No fixation plus no instruction' respectively). In the top panel, higher values on the y-axis indicate higher correlation coefficients. In the bottom panel, values above the upper red line indicate evidence in favour of the existence of correlations over time, while values below the lower red line (log(BF) < -1) indicate evidence against correlation over time. Values falling between the two red lines are interpreted as indeterminate. Overall, reliability seems low for variability in gaze position, particularly in the horizontal dimension, but the other measures show good reliability.

Next, means over time points were averaged, resulting in two means for each measure: One for day 1, and one for day 2. Bayesian Pearson pairs were conducted on each of the measures between day 1 and day 2 to test intra-individual reliability on a longer time-span. Figure 4 shows the correlation coefficient and Bayes Factor for each measure (dark-blue triangles). The correlations between days show similar patterns to the ones between time points: Gaze variability appears least reliable, while pupil size variability, blink rate, and microsaccade rate show good reliability.

Interim-discussion: How long should a resting state session be?

Overall, oculomotor variability showed good intra-individual reliability over time, both before and after a task of 30/50 minutes (Experiment 1 and 2 respectively), as well as over days (Experiment 2) – although variability in gaze position appeared to be the least reliable measure. It should be noted that the differences we found between individuals are substantial – for example, in Experiment 1, for gaze variability in the

horizontal dimension at time 1, the most variable participant has an SD that is 32 times larger than the least variable participant.

Findings for both experiments were based on a resting state of four minutes. The next question may be how long a resting state session should minimally take before it could be considered to produce reliable measures. To answer this question, we analysed the data of Experiment 1 – looking at variability in gaze and in pupil size over the course of the resting state. First, for each measure, the Pearson r-value between time 1 and time 2 was calculated on every cumulative second. This results in 240 r-values – with the first r-value being based on one second of data, and the last r-value being based on four minutes of data. This trajectory reflects how the consistency between the two time points develops as more data is collected (red line on Figure 5).

Next, we adopted a subsampling approach, using a simplified version of Schonbrodt and Perugini's (2013) approach. From the entire pool of data of four minutes, one chunk of data was randomly selected for both time points, and the r-value between them was calculated. This subsampling was done 1000 times for each cumulative second, represented on Figure 5 by the grey circles, with the mean represented by the black line. This means that, for example, at time = 1 sec, there are 1000 different r-values, each based on one continuous randomly selected second in the entire pool of data. Next, at time = 2 sec, there are also 1000 different r-values, each based on two continuous randomly selected seconds in the data. As such, we end up with 1000 r-values at each cumulative second. Because of this method, the r-values converge to one point as the subsamples are based on more data – resulting in very small margins of error at the right side of the x-axis. Still, the mean trajectory of the subsampled r-values combined with the trajectory of the 'actual' r-values can give an idea of the minimal necessary length for an oculomotor resting state.

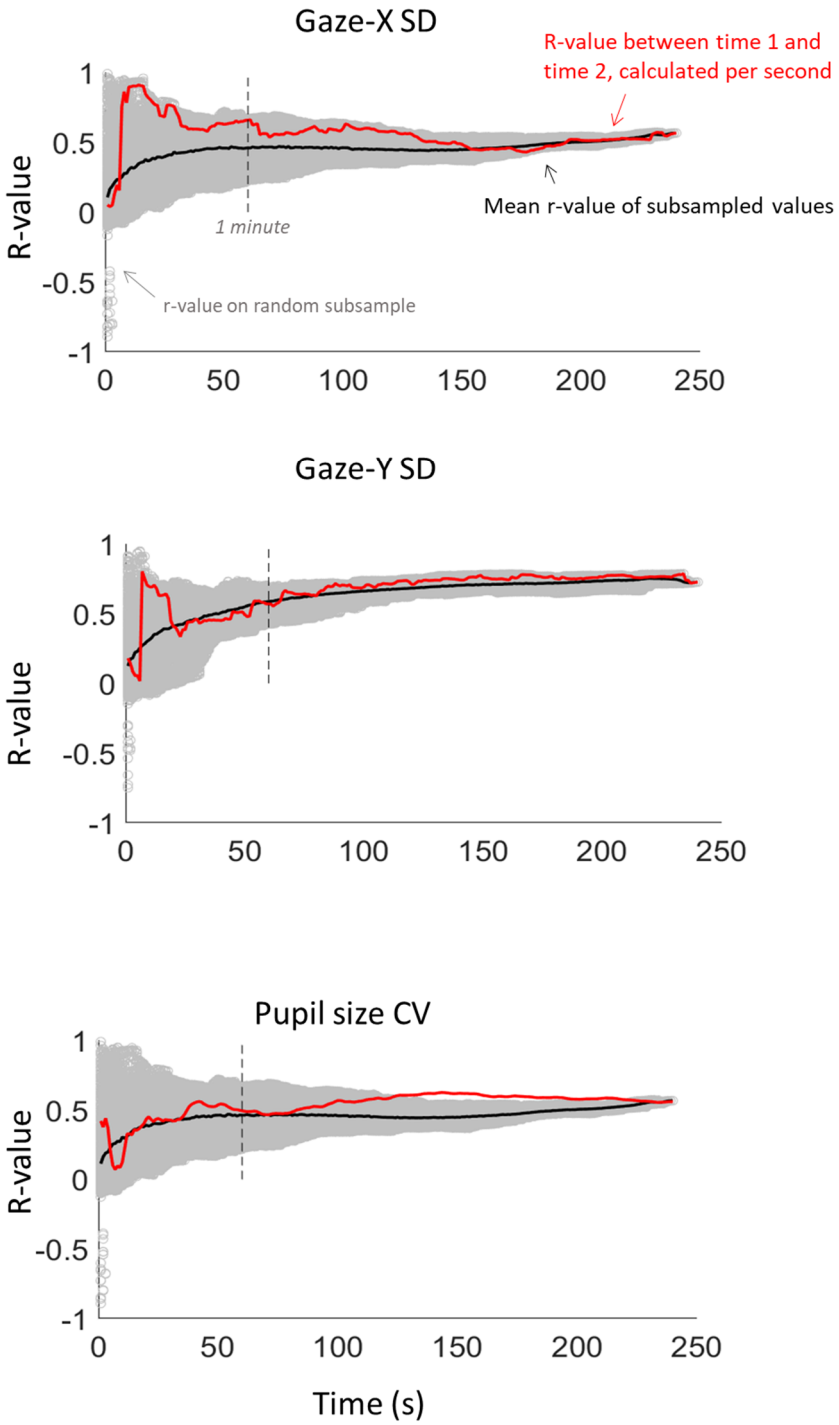


Figure 5. Intra-individual reliability of Experiment 1 over the course of the resting state for our three continuous measures: gaze variability (in the horizontal and vertical dimension) and pupil size variability. The r-value between time points 1 and 2 was calculated at each cumulative second (red), thus reflecting the trajectory over time. Next, for each cumulative second, estimates of the r-value were calculated on 1000 random subsamples. These estimates are shown in light-grey circles, with the mean of these subsamples shown in black.

Looking at Figure 5, it seems that reliability is lower and more volatile when it is based on less than a minute of data. After one minute, the reliability stabilises, and does not seem to improve any further after two minutes. Based on these outcomes, we recommend that an oculomotor resting state session is no shorter than one minute, but that it may not be necessary to collect more than two minutes of continuous data. However, this conclusion is based merely solely on the gaze position and pupil size recordings, and not on blink and microsaccade rates (which occur at a much slower time scale).

In Experiment 3, we were not only interested in the intra-individual reliability of oculomotor variability over different days (repeatability), but also in the extent to which the oculomotor variability would generalise over different types of 'oculomotor resting states'. For this, we used the same resting state version as in Experiment 1 and 2, as well as a free viewing version (in which participants did not have to fixate on anything, and were free to look anywhere on the screen), and an 'intermediate' version (in which participants were asked to fixate on the middle of the screen, but were not provided with a fixation dot). Because participants were asked to participate in each condition on four different days (resulting in twelve resting states per participant), we made the sessions shorter – using one minute per resting state instead of four. As shown above, this is long enough to produce reliable estimates.

Experiment 3: Reliability over days and conditions

For each of the measures (blink rate, pupil size variability, horizontal gaze variability, and vertical gaze variability), means were calculated separately for each condition and each day (thus resulting in twelve means for each measure). Bayesian Pearson correlations were conducted for each measure between the means over the different days, separately for each condition (resulting in eighteen correlation pairs for each) – to test the reliability of the oculomotor measures over time. Figure 4 shows these correlation coefficients and Bayes Factors (asterisks) for each of the three conditions (with 'Fixation plus instruction' in red, 'No fixation, instruction only' in black, and 'No fixation plus no instruction' in light-green). The overall pattern is similar to that of Experiment 2. Gaze variability in the horizontal dimension seems least reliable: Bayes Factors mostly show indeterminate evidence *against* a

correlation. Pupil size variability and blink rate show the most evidence for good reliability: While the Bayes Factors show a very wide range (with some below 1, but others logged values around 6-7), the overall distribution favours the existence of correlations over the absence of correlations. Again, correlation coefficients for these two measures were mostly moderate to high, with both median values around .5. Our 'intermediate' condition, in which participants were asked to fixate at the middle of a blank screen, appeared to produce the least reliable measures.

Over all three experiments, we thus found reliability in oculomotor measures over time, from relatively short ranges (30 to 50 minutes) up to multiple days apart. Next, we were interested in to what extent the oculomotor measures were generalisable over different types of resting states. To examine this, means were averaged over days, resulting in three means for each measure, each reflecting one condition. Bayesian Pearson correlations were conducted on the means of the three conditions – to investigate the reliability of the measures over different conditions. Figure 6 shows the correlation plots between the conditions for each measure, with Table 2 showing the accompanying correlation coefficients and Bayes Factors. All correlations had a Bayes Factor above 1, with eight of them ranging from moderate to extreme. Overall, the measures again show moderate to high reliability, although it is the poorest for gaze variability in the horizontal dimension. Blink rate seems to have the highest reliability over conditions.

Intra-class correlation

The intra-class correlation can estimate the reliability of a larger group of measures, to reflect to what extent they measure the same underlying phenomenon – and as such, can reflect the 'correlation' between more than two measures. To estimate the intra-class correlation, a two-way random model was conducted on each measure. The measure of consistency was estimated, as this is most similar to our Pearson correlation analyses. Table 3 shows the correlation coefficients for the average measure, to reflect the overall consistency of the resting states. The analysis was run both on each condition separately as well, to get an estimate of reliability over days, and collapsed over conditions and days, to get an estimate of the overall reliability of the paradigm.

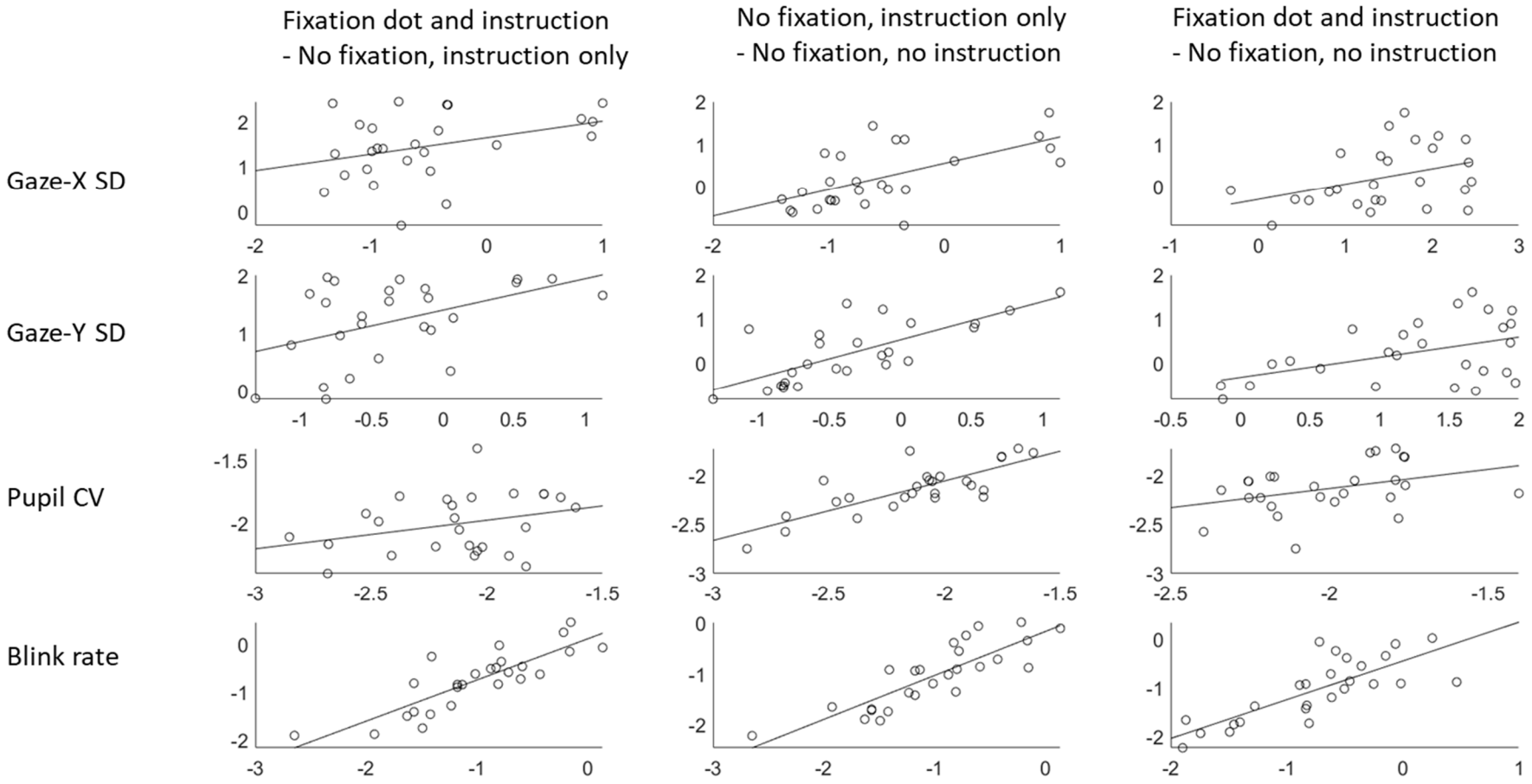


Figure 6. Correlation plots between the three different conditions ('Fixation plus instruction', 'No fixation, instruction only', and 'No fixation plus no instruction') on each of the four oculomotor measures from Experiment 3. Overall, evidence favours the existence of correlations – suggesting good intra-individual reliability of oculomotor variability over the different conditions. Note that the measures are logged.

Table 2. Overview of the intra-individual reliability across conditions for each of the four measures from Experiment 3. For each pair of conditions and each measure, the correlation coefficient is shown, with the accompanying BF_{10} in brackets.

Measure	Fixation + Instruction vs Instruction only	Fixation + Instruction vs No fixation + Instruction	Instruction only vs No fixation + Instruction
Gaze-X	.36 (1.12)	.63 (60.05)	.37 (1.28)
Gaze-Y	.47 (3.66)	.73 (1241.77)	.45 (3.10)
Pupil size	.40 (1.67)	.83 (77689)	.43 (2.41)
Blink rate	.84 (241807)	.85 (288824)	.79 (10224)

Table 3. Overview of the intra-class correlation coefficients of the average measure for each of the three conditions from Experiment 3, separately for each of the four measures, as well as the coefficients per measure over all conditions and days combined.

Measure	Fixation + Instruction	Instruction only	No fixation + Instruction	All
Gaze-X SD	.74	.77	.83	.85
Gaze-Y SD	.75	.74	.85	.87
Pupil size CV	.88	.65	.76	.88
Blink rate	.80	.65	.81	.91

All three conditions showed moderate (.5-.75) to good (.75-.9) reliability (see Koo & Li, 2016 for guidelines), although results again indicate that the 'Instruction only' condition produces the least reliable results. When collapsing over all days and all conditions, reliability is even higher, ranging from good to excellent (.9-1). Overall, the conditions seem to measure the same underlying construct – reflecting good

intra-individual reliability of oculomotor measures. Interestingly, the coefficients are all at least in the good range, even variability in gaze position – as such, diverging from the results of the individual Pearson correlations. However, the Pearson correlations can only reflect the consistency between two single measures, while our intra-class correlations reflect the consistency over all the different days averaged together. This suggests that over all the days combined, the oculomotor variability still shows within-subject consistency.

Results aim 2. Between-subject correlations between ADHD, mind wandering, and impulsivity

Bayesian Person correlations were conducted on the questionnaire scores. Figure 6 shows the between-subject correlational plots with their corresponding Pearson r coefficients and Bayes Factors. Looking at the between-subject correlations between ADHD tendencies, mind wandering (DFS), and impulsivity (UPPS-P), we found that ADHD tendencies were highly correlated to impulsivity and mind wandering tendencies. Both of these findings thus provide extreme evidence for replication of previous literature.

There was also some evidence for a correlation between mind wandering and impulsivity, but the evidence was in a much lower range and the accompanying correlation coefficient was similarly low, Pearson r = .23, BF_{10} = 3.8. It seems plausible that this correlation is caused by a confounding effect of ADHD tendencies. To statistically control for ADHD tendencies, a Bayesian Linear Regression was performed in which impulsivity scores were regressed on mind wandering tendencies (alternative Model M_1) and compared to a null-model that included the ADHD tendencies as model term (model M_0; see Wetzels & Wagenmakers, 2012 for more details on this method). Bayesian evidence favoured M_0 over M_1, BF_{01} = 7.7, indicating that the relationship between impulsivity and mind wandering disappears when controlling for ADHD tendencies.

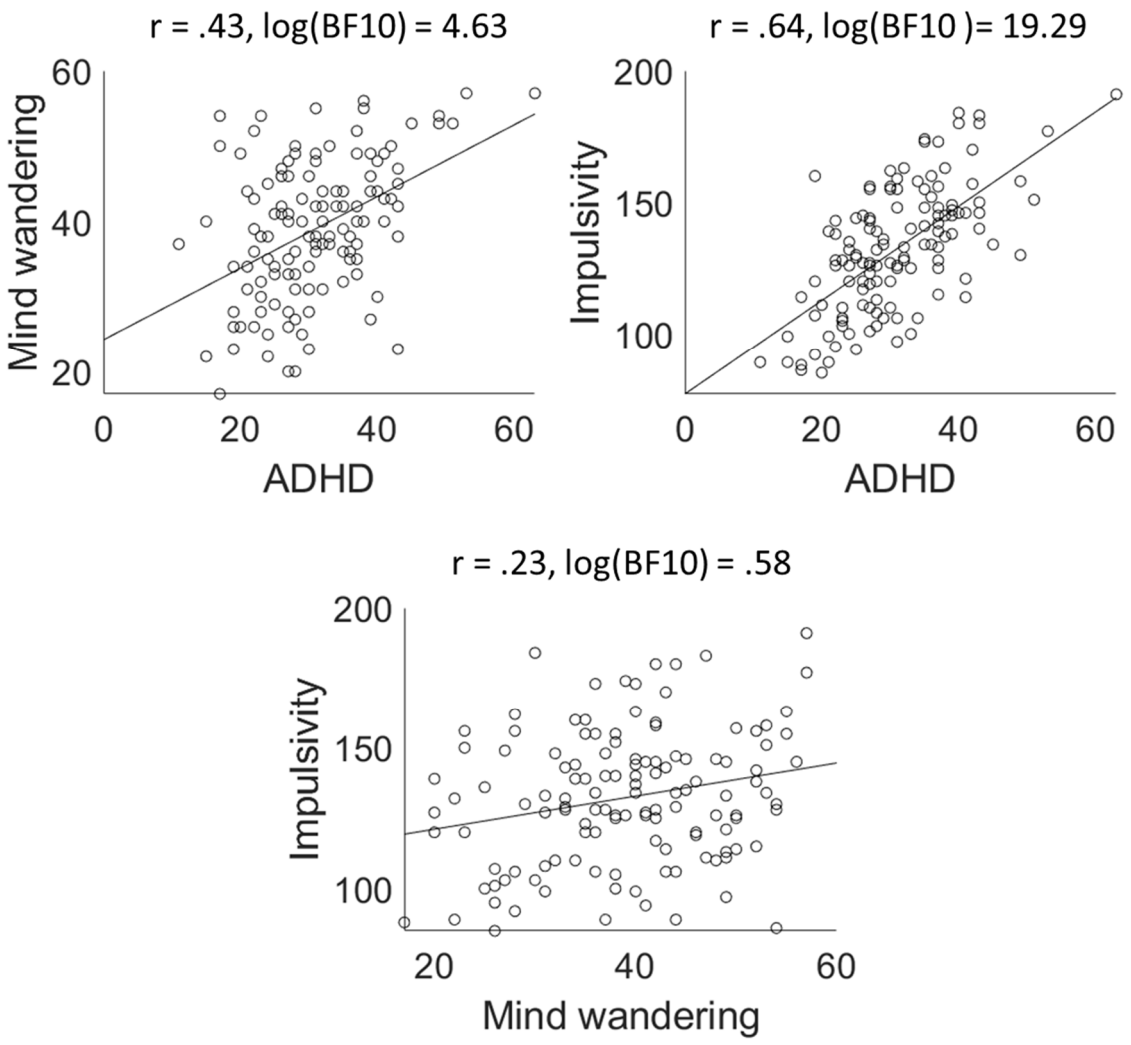


Figure 7. Correlational plots between self-assessed ADHD tendencies, mind wandering tendencies, and impulsivity, with accompanying Pearson r and Bayes Factor values. ADHD tendencies are positively correlated to both mind wandering and impulsivity – replicating previous literature.

Results aim 3. No between-subject correlations between questionnaires and oculomotor behaviour

For each participant, one mean was calculated for each measure, collapsed over all potential points of time, days, and conditions. Bayesian Pearson correlations were conducted between these oculomotor measures and the questionnaire scores. Out

of fifteen analyses, ten showed moderate evidence *against* a correlation, and five were in the indeterminate range (three of them with $BF_{10} < 1$, and the other two with $BF_{10} > 1$). Looking at the two correlations with $BF_{10} > 1$ (though in the indeterminate range), the r-values were low (only 4.4 and 4.8% explained variance).

To examine if any correlations would be more pronounced when looking at the subscales instead of the total scores of ADHD, the inattention and impulsivity/hyperactivity scores were correlated with the oculomotor measures. Pupil size variability correlated with the inattention subscale (r = .24, BF10 = 3.75), but not with impulsivity/hyperactivity (r = .13, BF10 = .31) – indicating that participants with more inattention-related ADHD tendencies showed more variability in pupil size. However, the explained variance was again low (5.8%).

Table 4. Pearson r-values (BF10) between the three questionnaires and the measures of oculomotor variability, combined over the three experiments.

Measure	ADHD	Mind wandering	Impulsivity
Gaze-X SD	-.15 (.41)	-.15 (.38)	.02 (.12)
Gaze-Y SD	-.10 (.20)	-.21 (1.61)	.02 (.12)
Pupil size SD	.22 (2.11)	.08 (.16)	.15 (.43)
Blink rate	.11 (.24)	-.09 (.18)	.12 (.27)
Microsacccade rate	.10 (.21)	-.02 (.13)	.08 (.17)

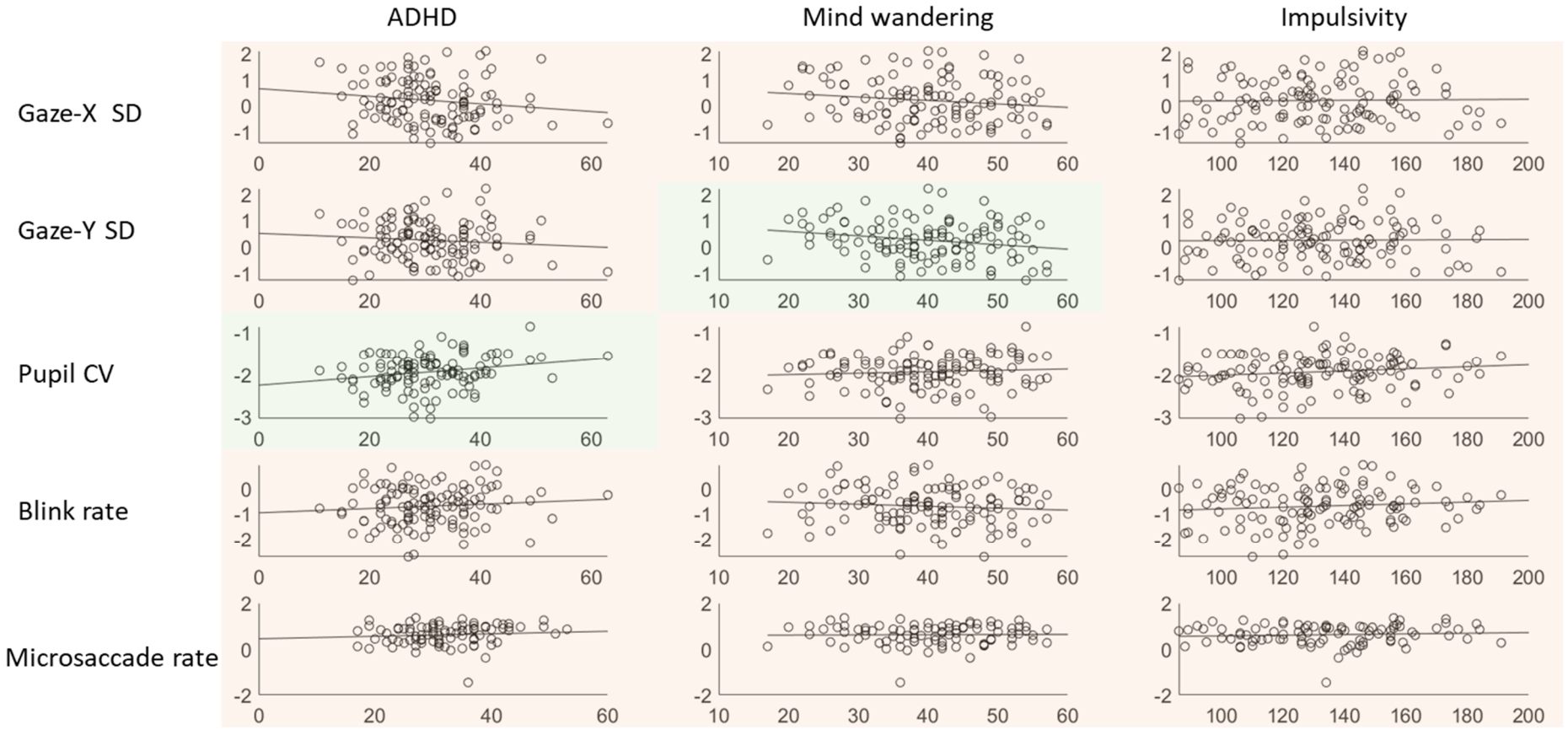


Figure 8. Correlation plots between the oculomotor measures and the self-assessed personality traits. Green shading indicates that the corresponding Bayes Factor is above 1 (indicating evidence in favour of a correlation between the conditions on that measure), while red shading indicates a Bayes Factor below 1 (indicating evidence against a correlation). Note that the oculomotor measures are logged.

Interim-discussion 2: Measuring gaze variability

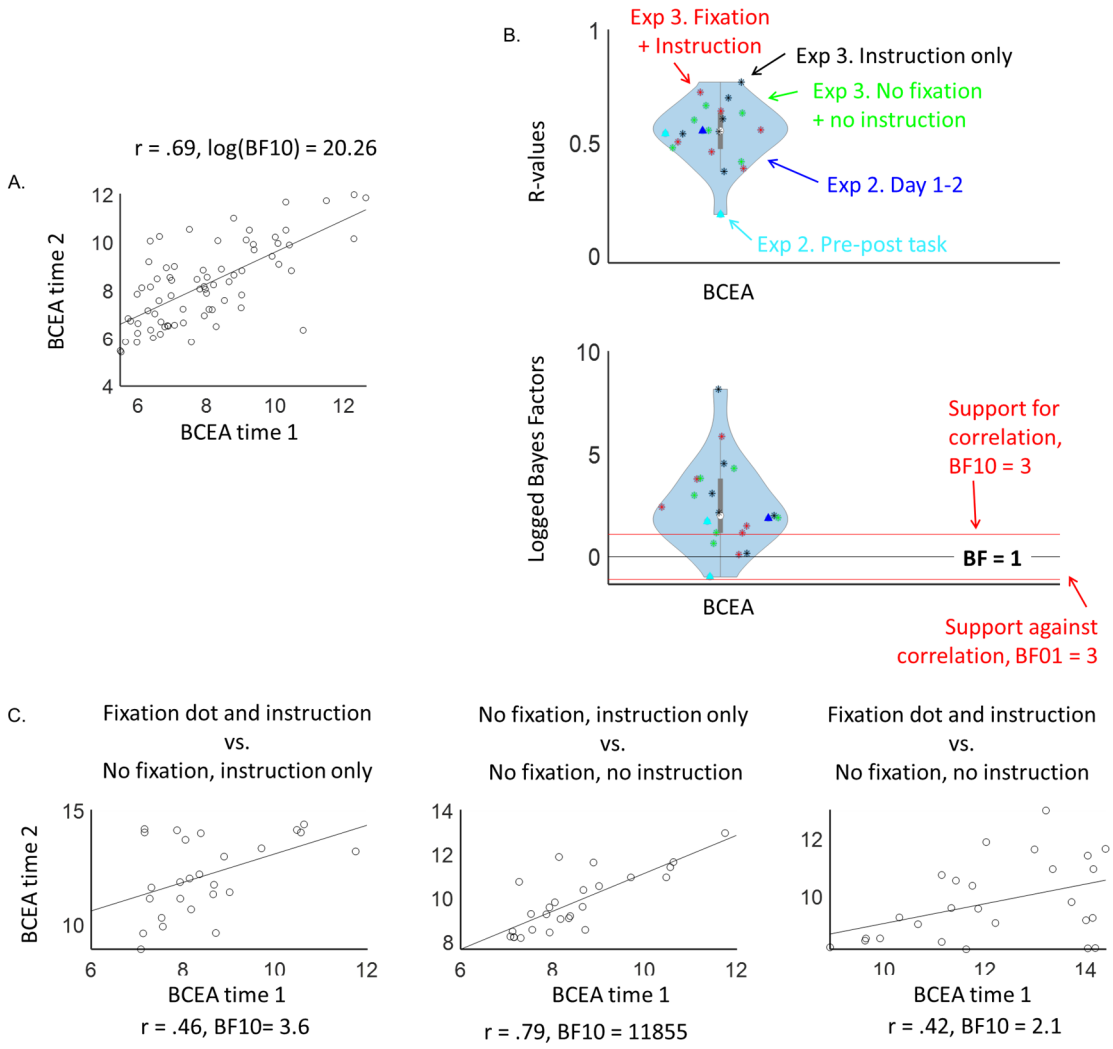


*Figure 9. Intra-individual reliability for the Bivariate Contour Ellipse Area (BCEA) for **A.** Experiment 1, showing the reliability between time 1 and time 2, **B.** Experiment 2 and 3, showing the reliability between different time points and between different days, and **C.** Experiment 3, showing the reliability between different conditions. Con*

Over the three experiments, gaze variability was consistently the weakest measure, particularly in the horizontal dimension. In the current analyses, the gaze-position over the horizontal and vertical dimensions were examined separately, as we had no prior knowledge on whether they would show similar reliability. However, within

the oculomotor literature, it is not unusual to quantify fixation stability with one combined measure. One measure is the 'Bivariate Contour Ellipse Area' (BCEA; Steinman, 1965), representing the ellipse-shaped area in which P % of fixations occur – as such, representing some sort of 'two-dimensional standard deviation'.

We calculated the BCEA with P = 68%, and reran the analyses on these values, for the data of Experiment 1 (Figure 9A), Experiment 2 (Figure 9B), and Experiment 3 (Figure 9C). Overall, it seems that both the repeatability and the generalisation of BCEA are comparable to the reliability of gaze variability in the vertical axis only – indicating that combining variability over the two axes does not add substantially more reliability.

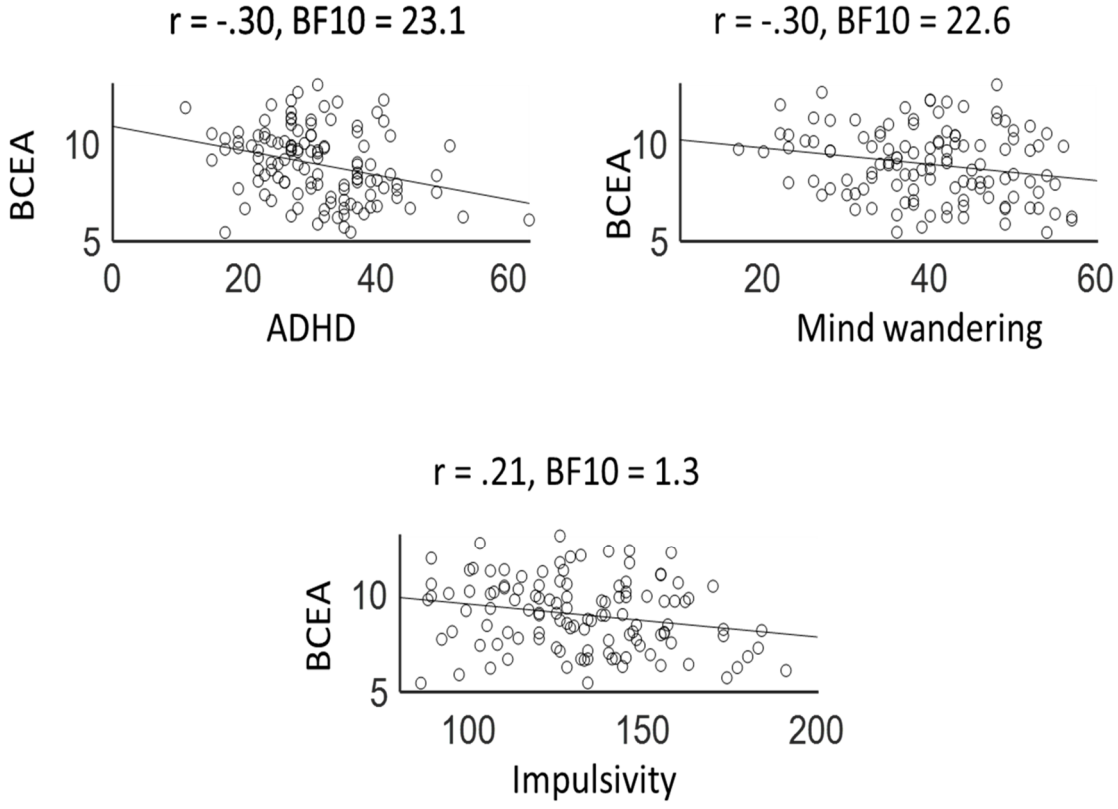


Figure 8. Correlation plots between the oculomotor measures and the self-assessed personality traits, with corresponding r-values and Bayes Factors on top. Results indicate negative correlations between: 1) BCEA and ADHD tendencies, and 2) BCEA and mind wandering tendencies.

However, there was evidence for a correlation between BCEA and ADHD tendencies, as well as between BCEA and mind wandering tendencies, though to our surprise, the directions were negative (see Figure 10). Remarkably, when examining the BCEA distribution, we noticed the values have an extremely large range between participants (minimum value = 234, maximum value = 457257). Visual inspection of the data revealed the larger BCEA values appeared to be driven by partial blinks – i.e., sudden large jumps in eye position, without complete loss of signal – suggesting that participants with more ADHD tendencies made fewer partial blinks. The meaning and significance of this remains open to interpretation. One possibility may be that people with low ADHD tendencies are better at suppressing blinks. However, note that the partial blinks have not been quantified, meaning this remains speculative

Discussion

In the current research, we aimed to: 1) examine to what extent endogenous oculomotor variability constitutes a reliable individual trait, 2) replicate positive associations of self-assessed ADHD tendencies to mind wandering and impulsivity, and 3) investigate potential relationships between personality traits and endogenous oculomotor variability. We combined datasets from three experiments including 'oculomotor resting states' as well as a set of questionnaires. We found that oculomotor variability indeed shows consistency within individuals, both over time (repeatability) and over different conditions (generalisation). Of the five measures that we used (variability in both horizontal and vertical dimension, pupil size variability, blink rate, and microsaccade rate), each showed consistency to some extent – with blink and microsaccade rate appearing to be the most consistent measures, and gaze variability being the weakest..

We also found positive correlations between the self-assessed personality traits, replicating previous associations of ADHD to mind wandering (Shaw & Giambra, 1993; Seli et al., 2015) and impulsivity (Berg et al., 2015; Miller et al., 2010). However, these personality traits did not show convincing correlations with

oculomotor variability. Overall, we mostly found Bayesian evidence against correlations – and for the few correlations that were weakly supported, the effect sizes were small. This suggests that within our sample of healthy participants, oculomotor variability did not prove a useful measure for corroborating self-assessed personality traits. Overall, gaze variability is driven by a multitude of sources, including saccades, drift, and tremor, but also by phenomena such as partial blinks. Because of this, gaze variability may have less specificity than the other measures, and thus, less validity as a measure. While reliability and validity are theoretically different constructs, in practice, they often go hand in hand.

Reliability of oculomotor variability

Recently, the World Federation of Societies of Biological Psychiatry and the World Federation of ADHD have identified the need for dedicated biomarkers of ADHD (Thome et al., 2012). Oculomotor measures seem appealing: They are easily accessible in terms of money and time, in sharp contrast with typical neuroimaging methods. Indeed, endogenous oculomotor variability has been proposed as a potential biomarker for ADHD (Panagiotidi et al., 2017). However, it is crucial for any biomarker to show intra-individual reliability (Mayeux, 2004).

Intra-individual stability of oculomotor variability during task has been shown in previous research (Andrews & Coppola, 1999; Boot et al., 2009; Castelhano & Henderson, 2008; Poynter et al., 2013; Rayner et al., 2007). Furthermore, there is evidence that oculomotor variability in viewing tasks shows intra-individual correlations with oculomotor variability in the absence of any visual stimulation (‘dark room condition’; Andrews & Coppola, 1999). Because oculomotor variability is measured within tasks, it reflects a mixture of exogenous and endogenous variability. This means that findings can be (partly) driven by individual differences in information processing and strategies. Previous studies have found similar intra-individual reliabilities in reaction time variability across time during task and across different tasks (Hultsch et al., 2002; Saville et al., 2011; Saville et al., 2012; but see Salthouse, 2012). In these contexts, it appears difficult to exactly quantify which part of the variability arises due to exogenous variability, and which part arises due to endogenous variability. To our knowledge, our design is the first to investigate the

intra-individual stability in ‘pure’ endogenous variability in oculomotor behaviour – captured by continuous measurement under an absence of changes in the external environment.

When comparing the reliability over time, correlation coefficients (and accompanying Bayesian evidence) were highest in Experiment 1 – in which the two measures were closest together in time – and lowest in Experiment 3 – in which measures were typically separated by multiple days. Still, correlation coefficients showed reasonable intra-individual consistency even in Experiment 3. Of course, the individual correlation pairs will be affected by chance. This is evidenced by the distribution plots in Figure 5, that shows the range of found correlation coefficients is large. However, overall, the distributions favoured moderate to high correlations, with median r-values being around .5 (with the exception of gaze variability). Furthermore, intra-class correlation coefficients showed good to excellent consistency for each of the measures – revealing that, overall, the measures over days appear to measure the same underlying construct. Based on our subsampling analysis on the data of Experiment 1, we can recommend that these sessions should be between 1-2 minutes long, with longer recording sessions being only necessary when the sample is small.

It is important to note that our findings show that oculomotor behaviour is consistent within individuals over time – likely reflecting individual traits. This means that individuals who are highly variable at time 1 typically are also highly variable at time 2, while individuals who show low variability at time 1 typically also have low variability at time 2. However, this does not mean that the measures are exactly the same at time 1 and time 2; they are still subject to variability. For example, looking at the reliability over time in Experiment 2 and 3, we typically explain 25% of the total variance (and for Experiment 1, explained variance ranged from 32 to 55%).

Statistical power and sample size

By combining data from multiple experiments, we were able to study individual differences in oculomotor variability in a large sample. Still, a number of our Bayesian analyses on individual differences produced Bayes Factors in the indeterminate range. If anything, this highlights the importance of using large

samples in these types of individual differences studies, especially when considering that effect sizes will typically be small (see Gignac & Szodorai, 2016 for a meta-analysis of effect sizes in Psychology). This may be more apparent with Bayesian analyses, in which evidence is gradual rather than a binary 'significant versus non-significant' decision. However, it is also important in traditional 'null hypothesis significance testing' (NHST), especially when considering that running underpowered studies actually increases the chance of making Type I errors (Bakker, van Dijk & Wicherts, 2012), while simultaneously making it difficult to interpret non-significant results.

It has been known for a long time that psychological research studies have been largely underpowered due to the use of small sample sizes (Cohen, 1962; Sedlmeier & Gigerenzer, 1989), but that sample sizes have not increased (although this seems partly field-dependent, see Marszalek, Barber, Kohlhart & Holmes, 2011). Researchers in Psychology appear to have incorrect intuitions about statistical power and sample sizes, and rely on rules of thumb and on (incorrect) practices from the literature (Bakker, Hartgerink, Wichters & van der Maas, 2016). To give an example, to obtain the traditionally recommended power of 80% for a correlation with r-value of .3 in NHST, a power analysis shows that a sample of minimally 85 participants is required. In the last years, the topic has gained increasing traction (e.g., Bakker et al., 2012; Button et al., 2013). It should be noted that increasing sample size is not the most obvious and necessary step in all types of studies (Rouder & Haaf, 2018; Smith & Little, 2018), but appears key when studying associations between measures that are prone to large heterogeneity.

However, although the power analysis is a common framework to think about power, sample size is not the only determinant of statistical power (Asendorpf et al., 2013; McCelland, 2000). Among others, one can obtain higher power by testing hypotheses that are well-grounded in theory, avoiding redundancy in predictor variables, increasing differences in experimental conditions, minimising measurement noise during data collection, and using appropriate statistical analyses. The ratio of 'true variance' to error variance can further be improved by using reliable measurements, and by collecting enough data points. For instance, using a high-quality eye tracker with a fast refresh rate leads to better oculomotor data, and subsequently to better estimates and higher statistical power.

Our results on individual differences diverge from previous literature, which found a positive association between ADHD and microsaccade rate in a healthy population (Panagiotidi et al., 2017). Comparing our study with theirs, we used the same eye tracker system and refresh rate, as well as the same microsaccade detection algorithm (Engbert & Kliegl, 2003) and the same analysis (Pearson r correlation). However, our study had a higher sample size (our correlation between ADHD tendencies and microsaccade rate included 94 participants, compared to 38 in theirs), as well as more data points (a minimum of 8 minutes in ours compared to ~6.5 minutes). This means that the absence of a replication in our results is not caused by a lack of power.

Individual differences in oculomotor variability

Potentially, these different findings may be explained by differences in design. In our experiments, oculomotor variability was recorded over a continuous 'resting state' session, while in Panagiotidi et al. (2017) participants were asked to fixate for only 20 seconds in a row over 20 different trials. After each trial, they were given a break, and could decide themselves when to continue with the next trial. Participants may control their eye movements in a different manner when they are aware they have a sufficient break in between. One possibility is that their found individual differences are driven mostly by an increasing deficiency to switch back and forth between trial and break – reflecting difficulties in executive functioning (which has been related to ADHD – see Willcutt, Doyle, Nigg, Faraone & Pennington, 2005 for a meta-analysis). To answer this, it would be necessary to investigate individual differences in how oculomotor variability evolves over the time course of individual trials and of the experiment as a whole – rather than looking only at a mean saccadic rate over all trials – to get more insight into possible mechanisms underlying these potential individual differences.

Our results also diverge from Fried et al. (2014), who found increases in microsaccade and blink rates, but not in pupil size variability, in a clinical ADHD population compared to healthy controls. However, again there are profound differences between these two studies: In Fried et al. (2014), participants performed a rapid action selection task with trials of 2 seconds long (Fried et al., 2014), that

featured a visual stimulus in each trial – and as such, meant to capture exogenous variability. In this case, the individual differences aim to reflect functional, task-based deficiencies in ADHD patients. As such, their task may be more sensitive to capturing such individual differences.

Overall, our findings show that the benefit of measuring endogenous oculomotor variability as an objective surrogate to self-assessed traits in the healthy population is unclear. When we did find correlations, the effects were small. Of course, we do not want to deny the importance of finding biomarkers for ADHD, nor the importance of studying behavioural correlates of ADHD. However, within the context of our findings, the benefit of measuring oculomotor activity as an 'objective biomarker of ADHD' seems unclear when short and simple questionnaires lead to much larger effect sizes.

One important point to bring up is the severity of symptoms. Our experiments were conducted on healthy participants, and as a result of that, there were not many individuals at the high end of the spectrum. If any individual differences exist, they will be more pronounced when comparing extremer cases. In healthy and academic samples, these more extreme cases will be difficult to find by chance, particularly in small samples – and even if they are found (e.g., in Panagiotidi et al., 2017, who obtained some extreme scores in their N = 38), interpretations should remain conservative. More definitive conclusions would require larger sample sizes on the healthy population, or oversampling for extreme scores. Furthermore, the comparison between clinical cases and healthy controls (e.g., Fried et al., 2014) is more sensitive by default. Even if oculomotor measures do not appear beneficial for differentiating between healthy individuals from a healthy population, they may still prove useful to distinguish clinical (or extreme) cases of ADHD, further characterise the dysfunctional circuitry underlying the disorder or assess the possible benefits of medication.

Mechanisms underlying potential individual differences

Within the context of our study, we have discussed possible associations between oculomotor variability and ADHD. This may imply that oculomotor variability is inherently detrimental. Of course, this would be a false assumption; oculomotor

variability inherently reflects the functioning of our oculomotor system. Fixational eye movements have been proven to be important for our vision (see Rolfs, 2009; Martinez-Conde et al., 2013 for reviews).

Within the context of oculomotor resting states, when participants are instructed to keep fixation, higher variability may be perceived as 'worse performance'. This means that on the one hand, as ADHD symptomology has been associated with decreased task performance in other types of tasks (see Kofler et al., 2013 for a meta-analysis; see Tamm et al., 2012 for a review), it could also be associated with decreased 'fixation performance'. On the other hand, because fixational eye movements are a healthy phenomenon during fixation, they may be reduced in clinical conditions. This highlights the importance of indicating which mechanisms would drive potential individual differences in variability. Instead, task-based oculomotor variability, in which certain eye movement patterns may be considered as beneficial or detrimental for the task, may be better suited to study these individual differences.

Oculomotor measures: extraction and correlations

In the current analysis, we used a cut-off of two degrees in amplitude; only microsaccades below this cut-off were counted for the microsaccade rate (similar to Fried et al., 2014; Panagiotidi et al., 2017). However, despite this cut-off being a traditional standard in the literature, it remains somewhat arbitrary. Saccades and microsaccades may represent a continuum, rather than two opposing categories (Otero-Millan, Troncoso, Macknik, Serrano-Pedraza & Martinez-Conde, 2008; Otero-Millan, Macknik, Langston & Martinez-Conde, 2013). We therefore reran our (micro-) saccades analyses without an amplitude cut-off. This measure may capture more of the total variability that participants exhibited. However, without this cut-off, results remained highly similar, and conclusions did not change.

It should be noted that we also used a cut-off for the extraction of blinks: Blinks were computed as missing samples with a maximum of one second – to differentiate blinks from periods of task disengagement (e.g., a participant falling asleep). Similarly, when rerunning our blink-related analyses without the upper-bound cut-off, our findings did not change.

To extract the microsaccades, we used the binocular detection algorithm of Engbert and Kliegl (2003). One feature of this algorithm is that the threshold for detecting a microsaccade is computed for each trial, to adjust for differing amounts of noise between different trials. However, our tasks do not contain any traditional trials, but continuous measurements of 1-4 minutes. This may affect the computation detection threshold due to untypical variability within the 'trial', resulting in too lenient thresholds. Still, our microsaccade rate is well in line with previously reported rates using shorter trials. Furthermore, we also used the measures of gaze variability, which may capture the microsaccades as well at the other types of fixational eye movements – thus reflecting an overall capacity to fixate.

Previous research has also looked at the associations between task-based oculomotor measures, and found that the six measures (saccade amplitude, microsaccade rate and amplitude, and fixation rate, duration, and size) that they used could be all be captured by one single factor in a Factor Analysis (Poynter et al., 2013) – they interpret this factor as *"Individuals' eye-movement behavior profiles"*. In our data, this was not the case. Seven out of ten pairs of measures showed evidence against correlation, with support only for some low correlations of pupil size variability with microsaccade and blink rate (r-values of .31 and .24 respectively). The only exception of gaze variability and the horizontal and the vertical dimension, which unsurprisingly are highly similar (r = .82), as they are intended to measure the same construct. Overall, our measures thus shared little to no variance and cannot be captured by one underlying construct. The differences in analysed measures may explain why Poynter et al. (2013) found one underlying construct in their measures, while we did not. Poynter et al. (2013) used three measures related to fixation, and three measures related to (micro-) saccades. Our measures are quite different from Poynter et al. (2013), with only microsaccade rate overlapping, and seem more divergent from each other.

Conclusion

In the current study, we found that oculomotor variability shows good correlation within individuals both over time and over different conditions. Particularly microsaccade rate, blink rate, and variability of pupil diameter show good reliability – meaning that these measures have the potential to be used as biomarkers. Of course, this begs the question of *what for* they can be used as biomarkers. Our results showed that the between-subject correlations to self-assessed ADHD, mind wandering, and impulsivity were all either absent or very small. In contrast, the questionnaires themselves correlated well with each other. Considering the low costs and ease of questionnaires compared to oculomotor data, the benefit of the latter in differentiating between personality traits remains unclear. Still, it is possible that oculomotor measures may serve a function complementing questionnaires or show stronger validity, for instance in predicting important outcomes.. Future research should focus on linking the resting-state oculomotor measures to task-related deficiencies in ADHD or differences in brain structure or integrity, as in these cases, oculomotor measures may serve as an easy and cheap substitute.

Chapter 2

Examining temporal structures in reaction time

Abstract

Human performance shows substantial endogenous variability over time. Such variability has been known to show temporal structures: Performance from action to action is not independent, but shows correlation with itself over time. While the existence of such dependencies has been frequently reported on, its measurement and interpretation come with a number of controversies, and its potential benefit for studying individual differences remains unclear. Two recent studies have linked temporal structures to individual differences in task performance, but with contrasting results. In the current study, we aim to investigate the intra-individual repeatability of these temporal structures in endogenous performance on the Metronome Task (25 participants, tested in two sessions ~45 minutes apart). Secondly, we examine the inter-individual correlates of the temporal structures (83 participants), specifically looking at: 1) task performance, 2) meta-cognitive ratings of attentional state, and 3) self-assessed personality traits (ADHD tendencies, mind wandering, and impulsivity). Rather than using one analysis method (as is common in the literature), we consistently compare all the frequently-used analysis methods – allowing us to investigate the structures without an a prior assumption on the underlying time scales. Results indicate that autocorrelation at lag 1 and Power Spectra Density slope showed the most intra-individual repeatability, while

autoregressive fractionally integrated moving-average model – ARFIMA(1,d,1) – parameters showed the least, and evidence for Detrended Fluctuation Analysis was indeterminate with a moderate effect size at best. Overall, the autocorrelation at lag 1 seemed the best measure for studying individual differences, due to its high reliability and ease of use. Furthermore, with exception of the ARFIMA parameters, temporal structure was correlated with performance but not with subjective attentional state or self-assessed personality traits between participants.

Keywords: Tapping; intra-individual variability; reaction time variability; ADHD; mind wandering; impulsivity

Introduction

For any action that one repeatedly executes over time, the iterations will show a large amount of variability in their time of execution. Such 'intra-individual variability' – variability within the same individual – manifests itself prominently during cognitive testing: In experimental tasks, participants are commonly instructed to repeat the same actions over a large amount of trials. Even in very simple reaction time (RT) tasks, participants' performance over the trials shows large fluctuations over time (see Figure 1A, top-left panel for an example of the RT series from one participant over 1000 trials). It is common practice to assume that trials are independent from each other, and to subsequently calculate one (conditional) mean per participant for analysis.

However, in practice, trials over time are not independent from each other, and by the process of averaging, information on any underlying temporal structures in the RT data gets lost. Still, as RTs across trials can be described as a 'time series' (i.e., a series that has a 'natural' temporal order), they can be quantified with time series analyses. These types of analyses confirm that RT shows temporal structures (see for instance Gilden, 2001; Van Orden, Holden & Turvey, 2003; Wagenmakers

et al., 2004), but their benefit for investigating individual differences is less clear. It has been suggested that these structures may be consistent within individuals (Torre et al., 2011), and that they may differ between individuals (i.e., that some individuals show more dependency over time than others; Gilden & Hancock, 2007; Madison, 2004; Simola et al., 2017; Torre et al., 2011).

However, it remains controversial: 1) how high these dependencies are, 2) at what time scale they occur, 3) how they should be measured, and 4) which neuro-cognitive mechanisms they reveal. The current research investigates these intra- and inter-individual properties of temporal dependency in more detail. In particular, we take a methodological approach: While most studies have focused on one analysis method, we consistently compare the different methods – allowing for an inspection of the temporal structures without an a priori assumption on the underlying time scales. Furthermore, we assess the relationship between temporal structures and attention.

Investigating temporal structures

Time series analysis methods

Below, we discuss the methods that have been commonly used to study temporal structure in behaviour: 1) autocorrelation, 2) Power Spectrum Density (PSD), 3) autoregressive fractionally integrated moving-average (ARFIMA) models, and 4) Detrended Fluctuation Analysis (DFA). To aid interpretation of these analyses, we compare white, brown, and pink noise – each of which is characterised by its own typical time structure (see Figure 1 for an overview).

Autocorrelation and spectral power density. White noise is generated by a completely random process, meaning that by definition, the observations are fully independent from each other: Every observation *n* on time series *x* consists solely of a random error term ε:

$$x_n = \varepsilon_n$$

This independency is reflected in the autocorrelation, which quantifies the correlation of a time series with itself over a specified lag (Box, Jenkins, Reinsel & Ljung, 2016). The autocorrelation *ρ* at lag *k* is calculated by:

$$\rho_k = \frac{E[(x_n - \mu)(X_{n+k} - \mu)]}{\sqrt{E[(x_n - \mu)^2]E[(x_{n+k} - \mu)^2]}}$$

with X_n reflecting observation *n* in time series *x,* μ reflecting the mean of the complete time series, and *E* reflecting the expected value. For a white noise series, the autocorrelation is not significantly different from zero at any lag (see Figure 1 for the correlogram).

Temporal structures can also be analysed in the frequency domain. By Fourier-transforming the time series and calculating the squared amplitude, one can obtain the power spectrum of the series[2] – or alternatively, the power spectrum can be calculated with a Fourier transform on the autocorrelation function (Box et al., 2016). The frequency *f* and power *S(f)* are directly proportional to each other:

$$S(f) \propto \frac{1}{f^{\alpha}}$$

To estimate α, both the frequency and power are log-transformed, and a linear regression line is fit in this log-log space. The linear slope indicates the α value. For a white noise series, the power spectrum (and corresponding slope) is flat and around zero.

Contrary to white noise, brown noise shows high dependency over time. This type of time series is also known as a 'random walk', as each observation *n* is the combination of the preceding observation *n-1* plus random error:

$$x_n = x_{n-1} + \varepsilon_t$$

For a random walk, the autocorrelation at lag 1 is high (near one), and shows a very slow decay over the subsequent lags – theoretically never reaching zero. In the frequency domain, α takes on a value of 2 – indicating a steep linear slope.

[2] Note that for reaction time data, the frequency is calculated by taking the inverse of the trial numbers (Gilden, 2001).

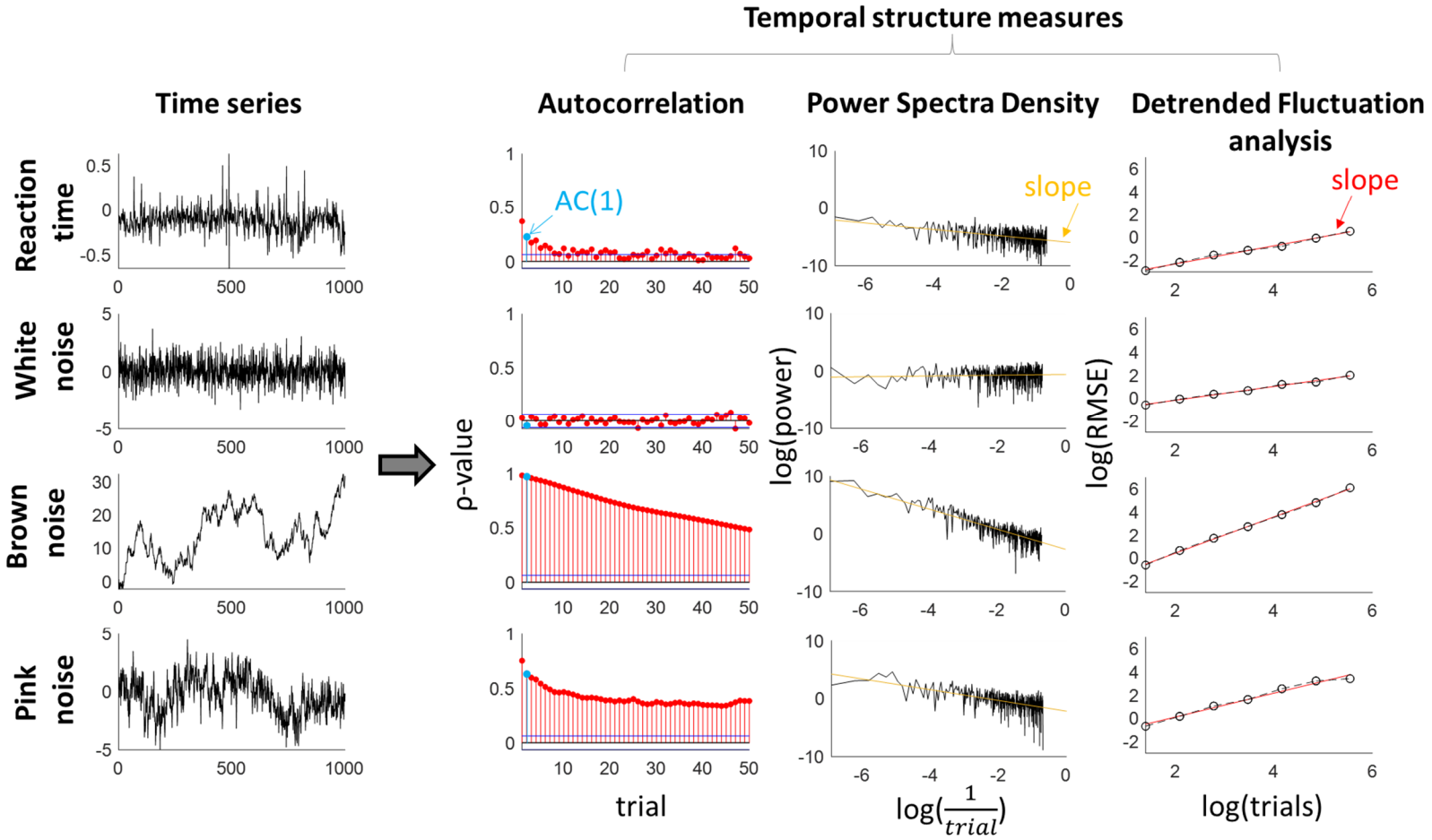


Figure 1. Examples of time series data over 1000 samples (left) and their corresponding temporal structures from lag 0 to lag 50. Shown are a reaction time series (showing small but clear temporal dependency), a white noise series (no temporal dependency), a brown noise series (i.e., a random walk, very high temporal dependency), and pink noise (high temporal dependency with slow decay).

While white noise shows no temporal dependency and brown noise shows high temporal dependencies, pink noise lies in-between the two. Pink noise is also known as '*1/f noise'*, as it characterised by an α of one (although α between .5-1.5 is typically considered as '1/f noise'). Pink noise shows relatively high autocorrelation at short lags, which slowly but gradually decreases to zero over the larger lags.

Although PSD has been popular for analysing RT, the method has an important limitation: While it is aimed specifically at measuring *long-range* dependencies (e.g., over the entire RT series), in practice, it has difficulties differentiating long- from short-term structures. When statistically assessing spectral slopes, they are typically compared to zero – i.e., to a null-hypothesis stating the time series has no dependency at all. If the null-hypothesis can be rejected, the structure is assumed to be long-term. As such, the hypothesis of 'short-term dependencies' is not considered at all. This is particularly problematic when considering that, although theoretically, short-term dependencies should be represented in shallow slopes, in practice they can resemble pink noise (Wagenmakers et al., 2004). Furthermore, while the regression line is linearly fit, the appropriateness of this fit is not tested.

Detrended Fluctuation Analysis. Another method to analyse temporal structure in RT is Detrended Fluctuation Analysis (DFA; Peng, Havlin, Stanley, & Goldberger). First, the time series *x* of total length *N* is integrated into *y(k)* by calculating the cumulative sum of each observation *n* relative to the mean of the time series *μ*:

$$y(k) = \sum_{n=1}^{k} (x_n - \mu)$$

Next, *y(k)* is divided into *b* number of windows $y_b(k)$. Each $y_b(k)$ value is detrended by the linear trend of that window. On the detrended values, the root mean square error – also called 'average fluctuation' *F* – can be calculated as a function of *b* with:

$$F(b) = \sqrt{\frac{1}{N}\sum_{k=1}^{N}[y(k) - y_b(k)]^2}$$

Both *F(b)* and *b* are log-transformed and linearly fit. Regression lope *α* is interpreted as amount of temporal dependency. A white noise series will be reflected in *α* = .5, pink noise in *α* = ~1, brown noise in *α* = ~2-3, and anticorrelated series in *α* < .5.

One advantage of this method is that it does not require the time series to be stationary. Secondly, unlike SPD, the fit of the linear regression line in log-log space can be visually inspected, and the minimum and maximum number of observations in each box can be adjusted for better fit.

ARFIMA models. The use of autoregressive fractionally integrated moving-average (ARFIMA) models has also been suggested (Wagenmakers et al., 2004; Torre, Delignières & Lemoine, 2007) – particularly as a means to quantify both long- and short-term processes. The ARFIMA model is an extension of the ARIMA model, which consists of a combination of three processes.

The first process of the model is the autoregressive process (AR), which aims to capture short-term dependencies. The model takes on an 'order' *p*, reflecting how many AR parameters are being estimated. For an AR model of order *p,* AR(p), observation *n* in time series *x* is predicted by its preceding observations x_{n-1} to x_{n-p}, with φ_1 to φ_p reflecting the weight for each observation. The model also includes an independently drawn error term ε_n. As such, the model can be described as:

$$x_n = \varphi_1 x_{n-1} + \varphi_2 x_{n-2} + \cdots + \varphi_p x_{n-p} + \varepsilon_n$$

The second process refers to the moving-average process (MA), which also captures short-term dependencies. For a MA model of order *q,* MA(q), observation *n* in time series *x* is predicted by a combination of random error ε_n and the error terms of the preceding observations, ε_{n-1} to ε_{n-q}, with θ_1 to θ_q reflecting the weight for each error term:

$$x_n = \varepsilon_n - \theta_1 \varepsilon_{n-1} - \theta_2 \varepsilon_{n-2} - \cdots - \theta_q \varepsilon_{q-1}$$

These two processes can be combined into a mixed ARMA(p,q) model:

$$x_n = \varphi_1 x_{n-1} + \varphi_2 x_{n-2} + \cdots + \varphi_p x_{n-p} + \varepsilon_n - \theta_1 \varepsilon_{n-1} - \theta_2 \varepsilon_{n-2} - \cdots - \theta_q \varepsilon_{q-1}$$

For example, an ARMA(1,1) model can be described as:

$$x_n = \varphi_1 x_{n-1} + \varepsilon_n - \theta_n \varepsilon_{n-1}$$

AR, MA, and ARMA are all meant for stationary time series – time series that have a constant mean and variance throughout (e.g., white noise). If a time series is not stationary (e.g., brown noise), an ARIMA(p,d,q) model should be used instead. This model includes a long-term process *d*, referring to the amount of times a time series should be 'differenced' to make it (approximately) stationary. In the process of differencing, each observation in the time series is subtracted from its subsequent observation. For instance, an ARIMA(1,1,1) model then takes the form of:

$$x_n = \varphi_1 (x_{n-1} - x_{n-2}) + \varepsilon_n - \theta_n \varepsilon_{n-1}$$

Importantly, in an ARIMA model, *d* refers to a discrete value. Most typical are d-values of 1 or 2, which are able to remove respectively linear and quadratic trends. Instead, in the ARFIMA model the series is instead 'fractionally differenced' – such that d can take on any value between -.5 and .5. Similarly, *d* in the ARFIMA model refers to a long-term process. One advantage of ARMA/ARFIMA is that they are nested models – meaning the best model can be selected using goodness-of-fit measures such as the Akaike Information Criterion (AIC; Akaike, 1974) and/or Bayesian Information Criterion (BIC; Schwarz, 1978). As such, one can fit both ARMA and ARFIMA on a time series, and test if the long-term parameter *d* sufficiently adds new information (Wagenmakers et al., 2004; Torre et al., 2007).

Overview. While of these four methodologies belong to the same class (*'fractal analyses"*) and their outcomes may be expected to show resemblances, they all serve a different function. The autocorrelation is able to measure dependency on the shortest measurable scale between directly-neighbouring datapoints (e.g., RT on trial *n* to RT on trial *n+1* – AC(1), as shown in Figure 1), and therefore comes with the most straightforward interpretation (Box et al., 2016). On the contrary, the PSD is meant for measuring long-term dependency, as it provides one measure (fitted slope) over a large range of data (e.g., the entire RT series, as shown in Figure 1; Box et al., 2016; Wagenmakers et al., 2004). The function of the DFA is highly similar to the PSD: It also provides one fitted slope over a large range of datapoints (see Figure 1; Peng et al., 1995). However, compared to PSD, it puts less weight on the short-term dependency to be more sensitive to long-term

dependency.While these three methods may all capture some form of temporal dependency, ARMA/ARFIMA is the only method that can statistically *test* for the benefit of adding a long-term dependency parameter – and may tell us whether long-term dependency is actually present in the series (Wagenmakers et al., 2004; Torre et al., 2007). Due to these differences in the measures, we therefore systematically compare all four.

Why might temporal structures be interesting?

Criticality. One common reason why the existence of temporal structures in behaviour has piqued interest is its link with criticality. Below, we describe the idea of criticality briefly (but see e.g., Beggs & Timme, 2012; Shew & Plenz, 2013 for detailed reviews).

In short, critical systems supposedly reflect an optimal balance between predictability and randomness. To get an idea of what this means within the field of neuroscience, imagine a population of neurons. On the one hand, if spiking activity between neighbouring neurons is completely independent, the system will be hypo-sensitive and activity will quickly go extinct. On the other hand, if neighbouring neurons are fully depended on each other, the system will be hyper-sensitive and activity will quickly spread everywhere. Such conditions respectively represent 'subcritical' and 'supercritical' systems. Critical systems would fall in between – thus allowing for activity to be send forward in the system without imploding. More precisely, this is called *'self-organised criticality*', as such organisation is endogenously driven. It has been argued that brains are systems that operate around the critical point (Beggs & Timme, 2012; Shew & Plenz, 2013).

Due to this element of dependency, a critical system should display some amount of correlation, with activity close together showing the strongest correlation, that decreases as it gets further apart. As the dependency is neither perfect nor random, this is most similar to the pink noise described above. Indeed, critical systems are thought to show such pink noise (or 1/f noise). Within the literature, this is often referred to as a 'power law' – meaning that in log-log space, activity shows up as a straight line throughout (as in Figure 1). This straight line reflects that the

relationship is ‘scale-free’: One can take any subpart of the spectrum and find the same straight line. Although not all critical systems adhere to the power law, and power laws can show up in non-critical systems, it is generally seen as a highly important characteristic of critical systems.

The link between criticality and temporal structures in behaviour is as following: Within the literature, it has been argued that behaviour over time shows 1/f noise too, and therefore, that cognition is a self-organised critical system (e.g., Gilden, 2001; Thorston & Gilden, 2005; Van Orden et al., 2003). However, whether or not the time structures are actually high enough to be considered pink noise is a controversial topic (see Farrell, Wagenmakers & Ratcliff, 2006; Wagenmakers et al., 2004; Wagenmakers, Farrell & Ratcliff, 2005; Wagenmakers, van der Maas & Farrell, 2012 for critiques) – and may be dependent on the used analysis method. Nonetheless, the interest in human cognition as a critical system partly explains why focus on the literature has solely been put on measuring temporal structure (as opposed to manipulating it, or examining its individual differences): The interest often starts and stops at the mere existence of pink noise in the data.

Predicting behaviour. Even if the temporal structures do not relate to criticality, one may still agree that RT carry a predictable and an error component. However, while prior studies have used time series analyses to quantify the structure in an existing series, it remains unknown to what extent these structures are informative for future behaviour. In other words, if one can find that behaviour on trial *n* is correlated to trial *n-1*, is it also possible to predict behaviour on yet-unobserved trial *n+1*?

Such ‘forecasting’ lies within the possibilities of the time series analysis, particularly of the ARFIMA models. These have been used to forecast weather or economic trends – but so far, have not been used to forecast behaviour. Aside from a theoretical interest, such behaviour forecasting may also have a practical use: Given that behavioural variability fluctuates over time and occasionally fluctuates to extremely poor responses (high RT or error), one may be able to prevent these poor responses by predicting them before they occur based on past behaviour. However,

the fruitfulness of this approach is partly dependent on the reliability of the temporal structures.

Attentional state. Just as our behaviour shows fluctuations over time, so do our meta-cognitive states. For example, throughout a task, we may feel more on-task on some moments and more off-task on others. It has been found that these fluctuations in subjective attentional state correlate to fluctuations in RT, such that variability is increased when one feels more off-task (Laflamme et al., 2018; Seli et al., 2013; Thomson et al., 2014). These findings seem to match common intuitions about our own functioning – namely, that we may show streaks of good performance during which we are extremely focused as well as streaks of poor performance in which we "*can't seem to get it right*". It has been argued that increased fluctuations from on-taskness to off-taskness are reflected in increased temporal structures (Irrmischer, van der Wal & Linkenkaer-Hansen, 2018; but see Wagenmakers et al., 2004).

If temporal structures are indeed related to attention, they may be different in people who show high Attention-Deficit and/or Hyperactivity Disorder (ADHD) tendencies. ADHD has previously been associated with higher RT variability (see Kofler et al., 2013 for a meta-analysis; see Tamm et al., 2012 for a review). It should be noted that increased variability in ADHD is not just associated with more attentional lapses, but also with a lack of response inhibition, the combination of which may lead to a pattern of extremely slow and extremely fast responses. Some previous work has examined temporal structures in performance of ADHD patients (e.g., Castellanos et al., 2005; Geurts et al., 2008; Johnson et al., 2007; see Karalunas, Huang-Pollock & Nigg, 2012; Karalunas, Geurts, Konrad, Bender & Nigg, 2014 for reviews; see Kofler et al., 2013 for a meta-analysis), but with different aims than in the current research (see Discussion for more details). These papers have hinted that individuals with ADHD show increased power spectra.

Individual differences

Two recent studies have aimed to link temporal structures to individual differences in performance. Firstly, Simola et al. (2017) extracted the DFA slopes from the RT of a Go/No-Go task. They found that participants with steeper DFA slopes made less commission errors on the task (r-value = -.35), but found no correlation with mean RT or standard deviation of RT.

Similarly, Irrmischer et al. (2018) extracted a DFA slope for each participant on target RT in a Continuous Temporal Expectations Task (CTET). The CTET is a sustained attention task, in which participants are presented with a series of stimuli – most of which have fixed temporal duration. Only if a stimulus is presented for longer than usual, participants have to make a response. However, they found a positive, high correlation between DFA slope and mean target RT (r = .72) – indicating that participants with high temporal dependencies performed worse on the task. Furthermore, Irrmischer et al. (2018) conducted a second experiment using the same task, with a mood manipulation with either a positive, neutral, or negative video – the latter of which has been thought to increase mind wandering (Smallwood, Fitzgerald, Miles & Philips, 2009). Participants in the negative condition showed increased RT and higher DFA slopes compared to the positive group. It should be noted that the study did not include a pre-manipulation performance baseline measure. In a third experiment, they investigated the temporal dependencies of subjective off-taskness probes during a meditation task. Subjects were probed quasi-randomly during a twelve-minute presentation, and were asked to rate their attentional state from 1 to 5. Participants who reported higher levels of off-taskness on average showed higher DFA slopes on these ratings (r = .66).

Despite the different findings, their interpretations rely on similar constructs. On the one hand, Simola et al. (2017) interpret their negative correlation between temporal structure and performance as evidence that brains which operate closer to the critical point show higher long-term correlations and allow for the mental flexibility needed to perform the Go/No-Go task. On the other hand, Irrmischer et al. (2017) interpret their positive correlation as evidence that brains which operate closer to the critical point show higher long-term correlations and allow for the successful dynamics of switching attention from task-related to task-unrelated

thoughts – and that the amount of attentional switching is affected by mood. Possibly, these differences could be explained by different task demands: in the Go/No-Go task, participants are required to make a response on 75% of the trials – and may therefore rely more heavily on response inhibition – while in the CTET, targets only appeared every fourth to tenth trial – and may therefore rely more heavily on sustained attention. Still, either task requires some of both elements, and it is difficult to see how the different task demands would lead to this particular pattern of results.

A few studies have looked at the intra-individual repeatability of temporal dependency. Torre et al. (2011) analysed the PSD slopes of 43 participants from a circle drawing and from a tapping task (seven sessions for each), but found no significant within-subject correlations between the two – suggesting that temporal structures are not consistent within participants over different tasks. Separately for each task, they computed a Cronbach's α on the seven sessions. However, these indicated moderate reliability at most (α = .61 for circle drawing and .56 for tapping). Simola et al. (2017) ran their task twice on each participant, changing only a task-irrelevant feature of the stimulus (colour), and found no significant differences in either the autocorrelation, the PSD, or the DFA between time 1 and 2. However, non-significant results are difficult to interpret, and this analysis does not give an estimate of the extent of repeatability.[3]

Current research

Here, we examine the properties of temporal dependencies in more detail, to see: 1) to what extent these structures repeat in individuals over time, 2) how these structures relate to objective and subjective task measures, and 3) how these structures relate to differences in self-assessed personality traits. Furthermore, we are interested in the effect of using different measures of temporal dependency, without making prior assumptions on the underlying structures.

[3] More generally speaking, it is also possible that two measures are significantly different from each other, but still show repeatability within participants – reflecting there is a difference in the group distributions, but the ranking of participants remain relatively stable.

To study these questions, we used the Metronome Task (MRT; Seli et al., 2013), in which participants are instructed to press in synchrony with a metronome. Throughout the task, participants are pseudo-randomly presented with thought probes on their attentional state. This task comes with a number of benefits for our current interests. First, it requires minimal dependency on the external environment while still obtaining behavioural measures – meaning it is particularly suited to assess *endogenous* fluctuations in performance. Secondly, tapping-based tasks (both with and without metronome) have been used extensively in the motor literature and show clear temporal structures (e.g., Delignières, Lemoine & Torre, 2004; Gilden, Thornton & Mallon, 1995; Lemoine, Torre & Delignières, 2009). Thirdly, the MRT provides on *online* measure of attentional state (i.e., measured during the experiment), which is known to correlate to fluctuations in RT.

It should be noted that the current study is the first to capture correlations between performance and temporal structures in a task that measures *endogenous* variability specifically, as the task does not involve different stimuli and responses. This makes our measures of performance straightforward to interpret. In contrast, tasks such as the Go/No-Go or the CTET require both withholding and responding. As such, they give multiple measures of performance, such as 'omission errors', 'commission errors', and 'RT to target stimuli'. To get a full picture of performance, these have to be interpreted in relation to each other. For instance, if a participant has a low amount of commission errors but also a large amount of omission errors, it is unclear how this would constitute 'better performance' or 'higher mental flexibility'. When investigating individual differences in performance, it is important to take all performance-relevant elements into account.

Intra-individual repeatability of temporal dependency

Before investigating if temporal dependencies are an interesting measure for individual differences, it is important to know if it shows consistency within individuals. We therefore examined the intra-individual repeatability of the temporal dependency measures (autocorrelation, PSD slope, DFA slope, and ARFIMA) in 25 participants over two sessions of the MRT (conducted about 40 minutes apart). To examine this, the measures of temporal dependency were calculated separately for

each session (time 1 vs time 2). If the measures show high repeatability, they should correlate highly with themselves over time. To further examine within-subject variability, we conducted the same analyses on performance and on the subjective attentional state ratings.

Correlates of temporal dependency

Furthermore, we were interested in the extent to which temporal structures relate to individual differences in: 1) task performance, 2) subjective reports of attentional state, and 3) (self-assessed) personality traits. To investigate this, we examined the temporal structure measures in 83 participants.

As the MRT requires the same response throughout, it obtains a basic measure of performance. As such, it allows us to study the relation between performance and temporal dependency in a straightforward manner. To examine this, the standard deviation on the full RT series was calculated, and subsequently correlated to the time series measures.

Furthermore, as participants are asked about their attentional state throughout the task, this allows us to study the associations between temporal dependencies and subjective attentional state ratings. For each participant, the average attentional state rating was calculated over all their probes. If off-taskness indeed enhances long-range dependencies, it should correlate positively to the temporal dependency measures. Participants also completed questionnaires on ADHD tendencies, mind wandering tendencies, and impulsivity. Scores on these questionnaires were correlated to the time series measures, to test if these self-assessed personality traits correlate with temporal structures in performance.

Methods

Participants

Eighty-four healthy participants (69 female, fourteen male, one other, aged between 18-25) participated for course credits. Of them, 25 randomly selected participants performed the behavioural task twice (session ~2 hours in total), and the rest performed the task once (session ~1.5 hours in total). The study was approved by the local ethics commission.

Materials

The behavioural paradigm was generated on a Viglen Genie PC with MATLAB version 8 (The Mathworks, Inc, Release 2015b) and Psychtoolbox-3 (Brainard, 1997; Kleiner et al., 2007; Pelli, 1997), and was displayed on an ASUS VG248 monitor with a resolution of 1920 by 1080 and a refresh rate of 144 Hz. The background was light-grey throughout the experiment, with the fixation point and text in white. During the MRT task(s), eye movements and pupil dilation were recorded with an Eyelink 1000 (SR Research), with participants seated with their head in a chinrest to limit motion (at 615 cm distance from the screen).

Questionnaires

To measure ADHD tendencies, participants completed the Adult ADHD Self-Report Scale (ASRS-v1.1; Kessler et al., 2005). This scale consists of eighteen items on a scale from 0 (*"Never")* to 4 (*"Very often"),* and is composed of two subscales: Inattention and Hyperactivity / impulsivity (Kessler et al., 2005; Reuter et al., 2006). Internal consistency of the ASRS-v1.1 is high (Cronbach's ranged .88-.94; Adler et al., 2006; 2012).

To measure mind wandering tendencies in daily life, participants completed the Daydreaming Frequency Scale (DFS; Singer & Antrobus, 1963), a subscale of the Imaginal Processes Inventory that consists of twelve 5-point items. The DFS

also has a high internal consistency, as well as high test-rest reliability (Cronbach's α = .91, test-retest reliability with interval of maximum one year = .76; Giambra, 1980).

Furthermore, participants filled in the UPPS-P Impulsive Behaviour Scale (Whiteside & Lynam, 2001; Lynam et al., 2006). This questionnaire consists of 59 items, ranged from 1 (*"agree strongly"*) to 4 (*"disagree strongly"*), and is composed of five subscales: positive urgency, negative urgency, (lack of) premeditation, (lack of) perseverance, and sensation seeking.

For the purpose of several other studies, all participants also filled in six other questionnaires, which were not analysed in the current study: the Beck Anxiety Inventory Second edition (Beck & Steer, 1993), Beck Depression Inventory Second edition (Beck et al., 1996), Short form Wisconsin Schizotypy scales (Winterstein et al., 2011), Five-facet Mindfulness Questionnaire (Baer et al., 2008), Toronto mindfulness scale (Lau et al., 2006), and Positive and Negative Affect Schedule (Watson et al., 1988).

Design

The Metronome Task (Seli et al., 2013) was used to obtain a RT series for each participant. From these series, we calculated for each participant: 1) the standard deviation of the RT, reflecting an overall measure of performance on the task, and 2) three measures of temporal dependency in the RT series. The MRT also measured participants subjective ratings of attentional state quasi-randomly throughout the experiment. Although the original MRT task offered only three levels of responses ("on task", "tuned out" and "zoned out"), we offered instead a scale from 1 (completely on task) to 9 (completely off task) in order to get a more gradual response. For participants who performed the MRT twice, these measures were extracted separately for both. ADHD tendencies, mind wandering tendencies, and impulsivity were self-assessed by means of questionnaires.

.

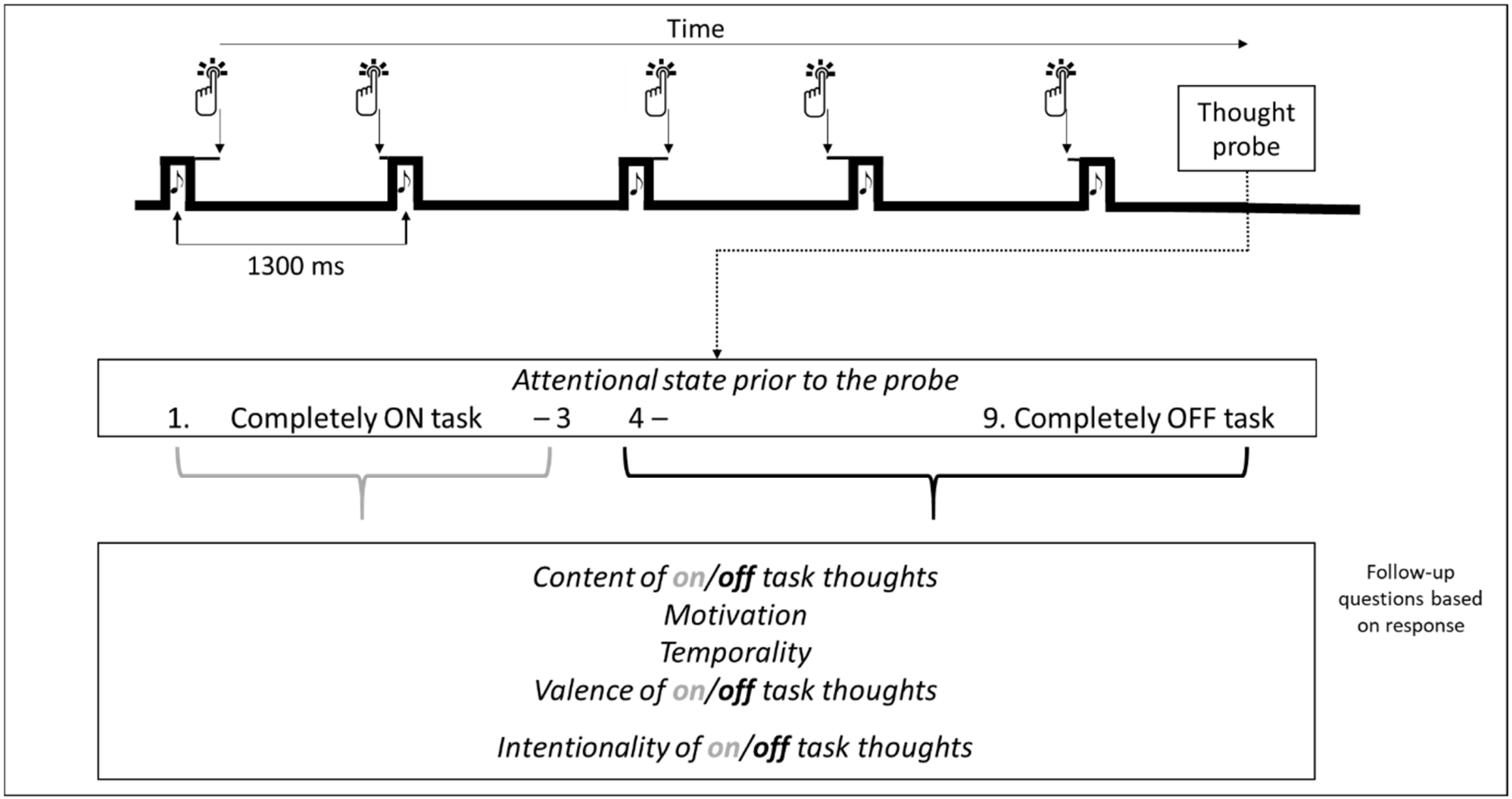


Figure 2. Overview of the task and thought probes.

Procedure

Participants came to the lab for one session. After eye tracker calibration, participants took part in a four-minute resting state session (eyes open), to get them into a common baseline state before starting the behavioural task. Next, they performed the MRT (~25 minutes). After the task, they performed another resting state session (eyes open), and then filled in nine questionnaires (DFS, ASRS, and UPPS-P, plus six additional questionnaires which were not analysed for the current study). In total, this took about 1.5 hours. Of the 83 participants, 2 of them then performed the MRT again, after watching one of two video clips of 3 and 5 min

Figure 1 shows an overview of the MRT task over time (Seli et al., 2013). Every trial lasted 1300 ms. In the middle of the trial (650 ms after onset), a short tone was presented to the participants (~75ms). The task of the participant is to press *on* this tone, such that perfect performance is indicated by a complete synchrony between tones and presses. Reaction time (RT) is measured as the relative time from the press to the tone (with RT = 0 being a perfect response). The RT series throughout the task were used to measure performance and temporal structures in performance.

Throughout the task, participants were presented with thought probes. The first question related to their subjective rating of attention just prior to the thought probe appeared ("Please record a response from 1 to 9 which characterises how on task you were just before this screen appeared", with 1 as 'completely ON task' and 9 as 'completely OFF task'). Participants were instructed that ratings between 1 and 3 indicated 'on task' states, and any higher rating indicated 'off task' states. Based on their rating, they then received five follow-up questions related to the content, temporality, valence, and intentionality of their thoughts, as well as to their motivation. The follow-up questions were not analysed in the current research. In total, the experimental phase of the MRT consisted of 21 blocks with 50 trials each (1050 trials in total), with one thought probe in every block. Probes were presented pseudo-randomly. To make sure the probes did not follow too closely after each other, they were never administered in the first five trials of a block.

The random lottery reward system (Cubitt, Starmer & Sugden, 1998) was used to motivate participants to keep up good performance throughout the task.

After the session, one trial *n* was randomly extracted, and if the standard deviation of trial *n* to trial *n-4* was below .075 (indicating consistent performance in that time window, with the cut-off based on pilot data), the participant received a reward of £5.

Before the experimental phase of the MRT, participants received a training block of 50 trials to learn the rhythm of the tone. At training trial 15, they were presented with a thought probe. After the training, the participant received feedback on their performance from the experimenter, to make sure they understood the task. Participants were also told how many of their trials would qualify for the reward, to provide them motivation to keep up good performance.

Data preparation and analysis

For each participant, the total percentage of omissions was calculated, and participants with more than 10% omissions were excluded from analyses (following the procedure of Seli et al., 2013). One participant was furthermore excluded for responding in anti-phase with the tone. This left us with 78 participants to analyse, with 21 having done both MRT sessions. For each of these remaining participants, the performance and temporal measures were calculated for each RT series separately – i.e., if participants performed the task twice, measures were calculated separately for both, to investigate the reliability between the two.

Two measures of overall performance were calculated on each RT series (RT_1, RT_2, …, RT_{1050}): the standard deviation of the RT (reflecting consistency), and the mean of the absolute RT (reflecting distance to the tone). As the distributions for both measures were highly skewed on the group level, both were log-transformed. Log(SD) and log($\overline{RT}$) were highly positively correlated to each other ($r = .81$, $\log(BF_{10}) = 38.1$). Subjective attentional state was measured as the mean and SD of the 21 ratings. We checked whether subjective off-taskness was correlated with variability (as per Laflamme et al., 2018; Seli et al., 2013). SD was calculated on the last five trials before each probe, and was correlated within participants to the attentional state ratings. Indeed, increased off-taskness was associated with increased variability on the group level (median Kendall's $\tau = .09$, $BF_{10} = 585$).

The autocorrelation at lag one (i.e., correlation trial *n* with trial *n+1*) was calculated for each participant in R (R Core Team, 2013) on their RT series with the *acf* function in the *Forecast* package (Hyndman et al., 2018; Hyndman & Khandakar, 2008). Furthermore, the power spectrum density was calculated on the RT series, using the inverse of the trial number as frequency (following Wagenmakers et al., 2004). As discussed in the introduction, white noise shows a flat power spectrum, centred around zero. However, when shuffling our RT data, the spectra did not behave like white noise. Instead, their intercept was dependent upon the overall variance in the series – which is particularly problematic when looking at intra- and inter-individual correlations. To correct for these differences, each RT series was randomly shuffled 100 times. The mean power spectra of these 100 iterations was subtracted from the original RT power spectrum. Next, the linear regression slope was calculated in log-log space – with the absolute value of the slope representing the α in $1/f^{\alpha}$.

DFA was performed on each RT series with the *Fractal* package (Constantine & Percival, 2017), following the procedure of Stadnitski (2012), over non-overlapping blocks sized from a minimum of 4 trials (as lower window sizes are not recommended for linear detrending; Peng et al., 1994) to 512 trials (maximum window size we were able to use). The linear regression slope was calculated in log-log space. Similarly, for each RT series, DFA was performed on 100 randomly shuffled series. The slope of the original RT series was corrected by the difference between the mean slope of the shuffled series and white noise (.5). To keep the inter-measure correlations as fair as possible, the same 100 shuffled RT series were used for the PSD and DFA analyses.

Lastly, an ARFIMA(1,d,1) model was performed on each of the RT series using the *Fracdiff* package (Fraley, Leisch, Maecler, Reisen & Lemonte, 2012), following the procedure of Wagenmakers et al. (2004), to extract the long-term parameter d, together with the two short-term parameters AR and MA.

Bayesian statistics were conducted in JASP (JASP Team, 2017), using equal prior probabilities for each model and 10000 Monte Carlo simulation iterations.

Results

Testing for the presence of temporal dependency

Before examining the intra- and inter-individual correlates of temporal structures in RT series, we first tested whether these series actually showed such temporal structures. Bayesian One Sample t-tests were conducted on the autocorrelations at lag one (AC1), the linear slopes of the PSD, and the long-term dependency parameter d – to test if they were statistically different from zero – and on the linear slopes of the DFA – to test if they were statistically different from .5. This was done separately for the participants from the first session (78 participants) and from the second session (21 participants). For each all measures, we found extreme evidence in favour of the alternative hypothesis –Table 1 for the log(BF_{10}) – except on the MA parameter on the first session (indeterminate evidence), and on the AR and MA parameters on the second session (moderate evidence against). All in all, these results suggest that the RT series indeed show temporal dependency.

Table 1. Logged Bayes' Factors in favour of the existence of temporal structures in the RT in the different measures: the autocorrelation at lag 1 (AC1), the linear fitted slope of the spectral power, the linear fitted slope on the detrended fluctuation analysis, and the ARFIMA(1,d,1) parameters (AR, MA, and d). .

	AC1*	**Spectral slope***	**DFA slope****	**AR***	**MA***	**d***
Session 1	73.7	79.2	78.5	4.8	0.7	57.3
Session 2	15.9	16.6	19.0	-1.6	-1.5	14.6

* $\log(BF_{value>0})$, ** $\log(BF_{value>.5})$

Intra-individual repeatability

To test the intra-individual repeatability of our MRT measures (RT variability, mean absolute RT, mean and SD of subjective attentional state, and temporal dependency of the RT series), Bayesian Pearson correlation pairs were computed for each measure between time one and two. Figure 3 shows the correlational plots with

corresponding r-values and Bayes Factors (BF_{10}) on top. The BF_{10} indicate the ratio of the likelihood of the data under the alternative hypothesis (e.g., the presence of a correlation) compared to the null-hypothesis. For example, for variability, the BF_{10} between time 1 and 2 is 48 – meaning that the likelihood for the data is 48 times larger under the alternative than under the null-hypothesis. This indicates strong evidence in favour of a correlation. On the other hand, the BF_{10} between time 1 and 2 for the d parameter is .77. This means that the data is $(\frac{1}{.77})$ 1.3 times more likely under the null than under the alternative hypothesis, indicating indeterminate evidence.

Overall, measures of performance (variability and mean absolute RT) and of subjective attentional state ratings (mean and variability) showed high repeatability over time. These findings indicate that participants were consistent both in their behaviour and subjective reporting over the two MRT sessions, ~45 minutes apart. Looking at the temporal structure measures, the autocorrelation at lag 1 and PSD were the most reliable (equally high as the performance measures), while the three ARFIMA parameters were the least reliable.

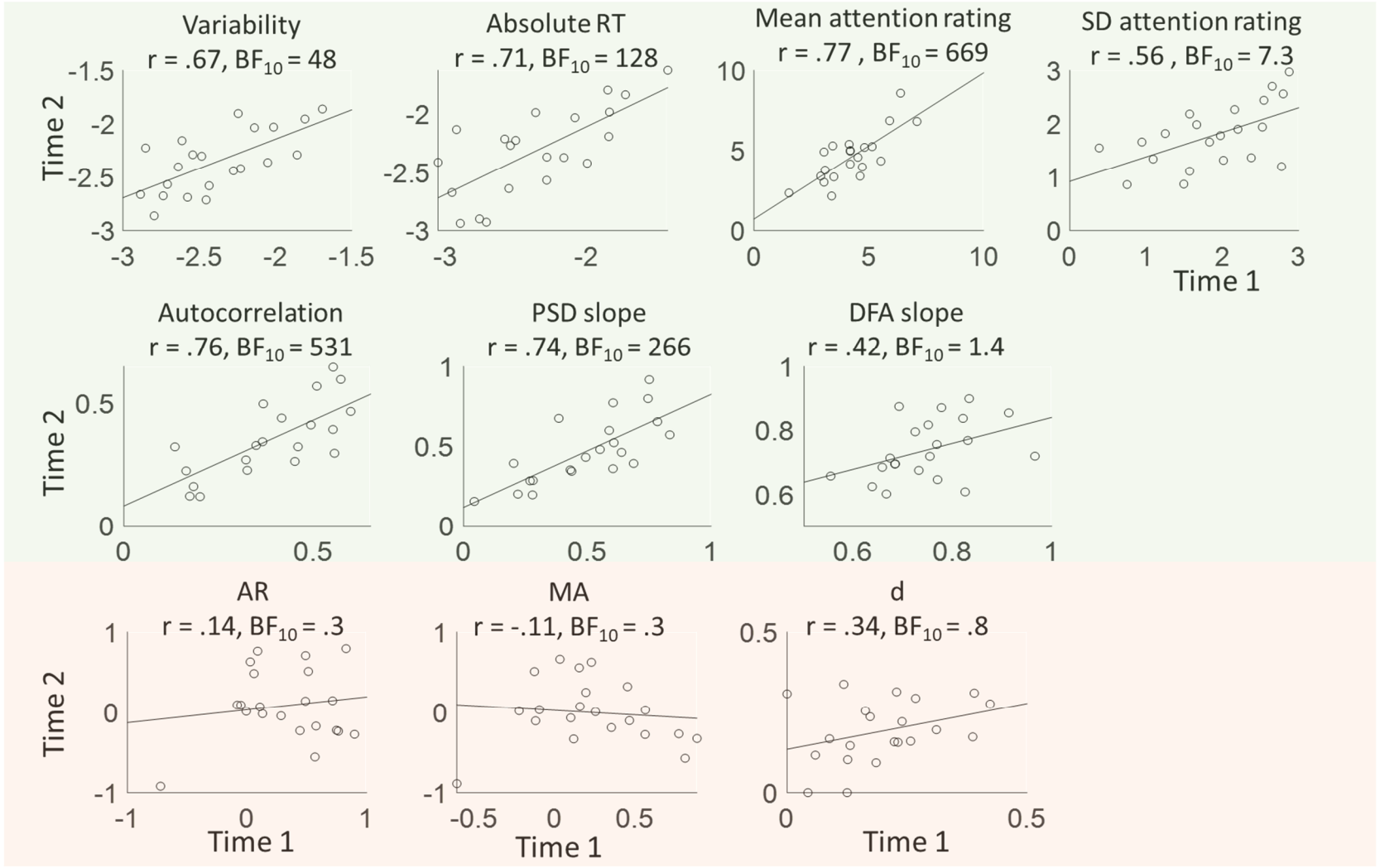

Variability
r = .67, BF10 = 48
Absolute RT
r = .71, BF10 = 128
Mean attention rating
r = .77 , BF10 = 669
SD attention rating
r = .56 , BF10 = 7.3
Autocorrelation
r = .76, BF10 = 531
PSD slope
r = .74, BF10 = 266
DFA slope
r = .42, BF10 = 1.4
AR
r = .14, BF10 = .3
MA
r = -.11, BF10 = .3
d
r = .34, BF10 = .8
Time 1
Time 2

Figure 3. Within-subject correlations between MRT session 1 and 2 for performance (variability and mean absolute RT), subjective attentional state ratings (mean and variability), and temporal dependency measures (autocorrelation at lag 1, Power Spectrum Density (PSD) slope, Detrended Fluctuation Analysis (DFA) slope, and the three ARFIMA(1,d,1) parameters – AR, MA, and d). Corresponding Bayes Factors above 1 are indicated by green shadings (indicating evidence in favour of that correlation), while Bayes Factors below 1 are indicated by red shadings (indicating evidence against that correlation). Overall, the ARFIMA parameters show poor reliability, while the other six measures show good reliability.

Between-subject correlates of temporal dependency

Next, we were interested in how these temporal dependency measures related to inter-individual differences in behaviour, subjective attentional state ratings, and self-assessed personality traits. Bayesian Pearson correlation analyses were conducted between the temporal dependency measures and: 1) RT variability (Figure 4, left column) and absolute mean RT (Figure 4, second column), 2) the mean and SD of the attentional state ratings (Figure 4, third and fourth column), and 3) the questionnaire scores (Figure 6). Overall, the autocorrelation at lag one and the PSD and DFA slopes correlated to performance – such that good performance was associated with lower temporal structures. However, we found evidence against correlations of temporal dependencies with both attentional state ratings and questionnaires. Below, the results are discussed in more detail.

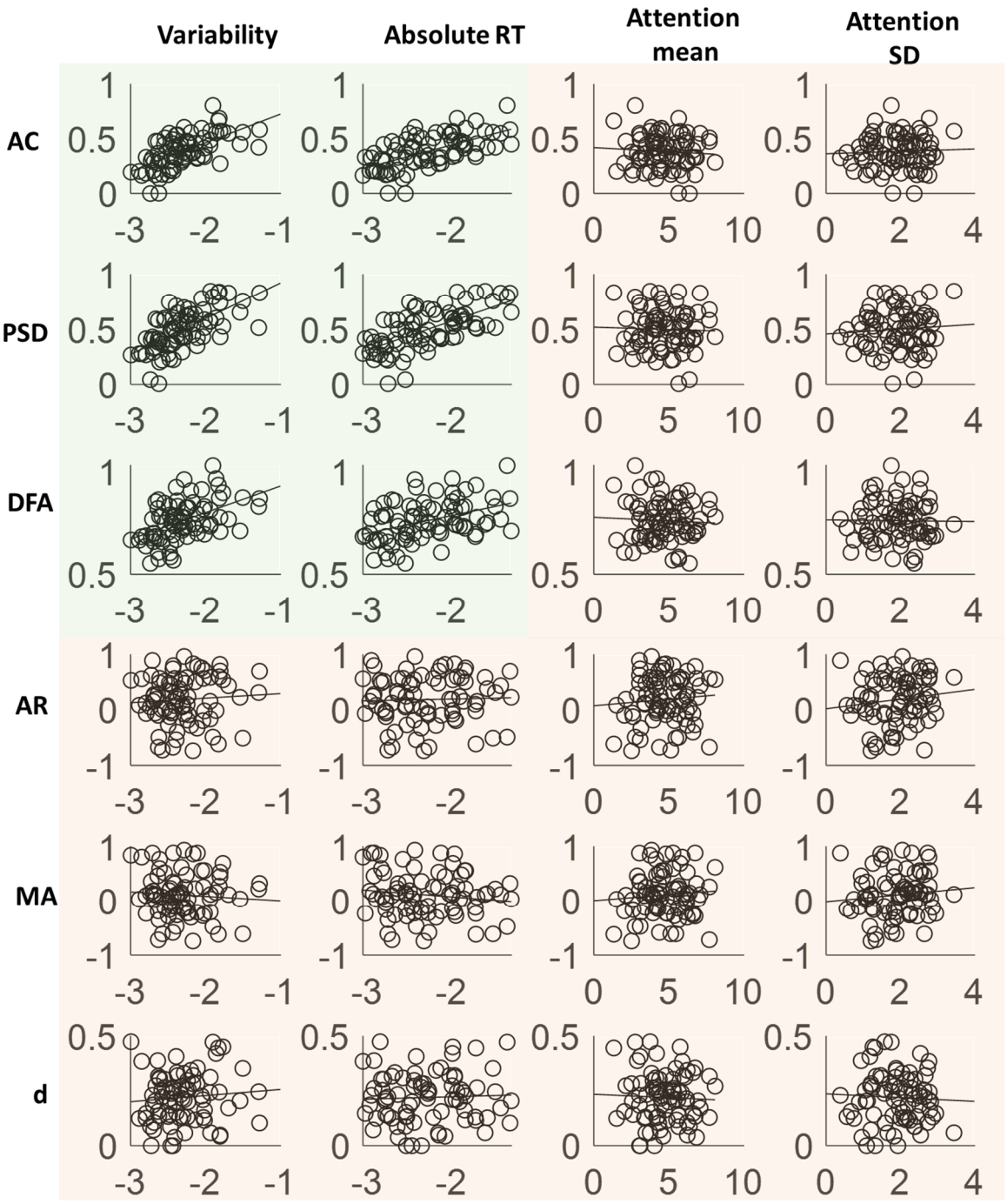


Figure 4. Between-subject correlations between performance and temporal dependency (left column for variability, and middle column for absolute RT), and subjective attentional state and temporal dependency (right column). Convention are the same as in Figure 3. Performance correlates moderately to strong with temporal dependency measures, with the exception of the ARFIMA(1,d,1) parameters. However, Bayes Factors show evidence against correlations between subjective attentional state ratings and temporal dependency.

Table 2. Pearson r-values between objective and subjective MRT measures and temporal dependency measures, with corresponding BF_{10} in brackets.

	Variability	Absolute RT	Mean attentional state	SD attentional state
AC	.62 (1.2e+7)	.62 (3.5e+7)	-.07 (.17)	.05 (.15)
PSD	.64 (6.5e+7)	.66 (3.5e+8)	-.04 (.15)	.08 (.18)
DFA	.48 (2617)	.41 (134)	-.05 (.16)	-.01 (.14)
AR	.07 (.17)	.03 (.15)	.08 (.18)	.13 (.28)
MA	-.07 (.17)	-.14 (.29)	.08 (.18)	.10 (.21)
d	.09 (.19)	.04 (.15)	-.04 (.15)	-.05 (.15)

Variability

Figure 4 shows the between-subject correlations of temporal dependency measures with both measures of performance, with corresponding r-values and BF_{10} in Table 2. Performance correlated with all the measures except the ARFIMA(1,d,1) parameters, for which there was evidence *against* correlations. For the other measures, we found that participants who performed well on the task displayed on average low temporal dependency. These correlations cannot depend on the variance of the time series, as we correct for this by subtracting the shuffled data series.

Figure 5 shows these dynamics in more detail over four different participants. Good performance, as indicated by low variability, was associated with a low autocorrelation coefficient at lag 1, that quickly decays over the increasing lags, as well as with relatively shallow PSD and DFA slopes (note that DFA slopes for white noise are .5). Poor performance on the other hand, as indicated by a high SD, was associated with high autocorrelation coefficient at lag 1, that slowly decayed over the next lags, as well as with relatively steep PSD and DFA slopes. Average performance (respectively showing SD around median and median values) showed intermediate temporal structures.

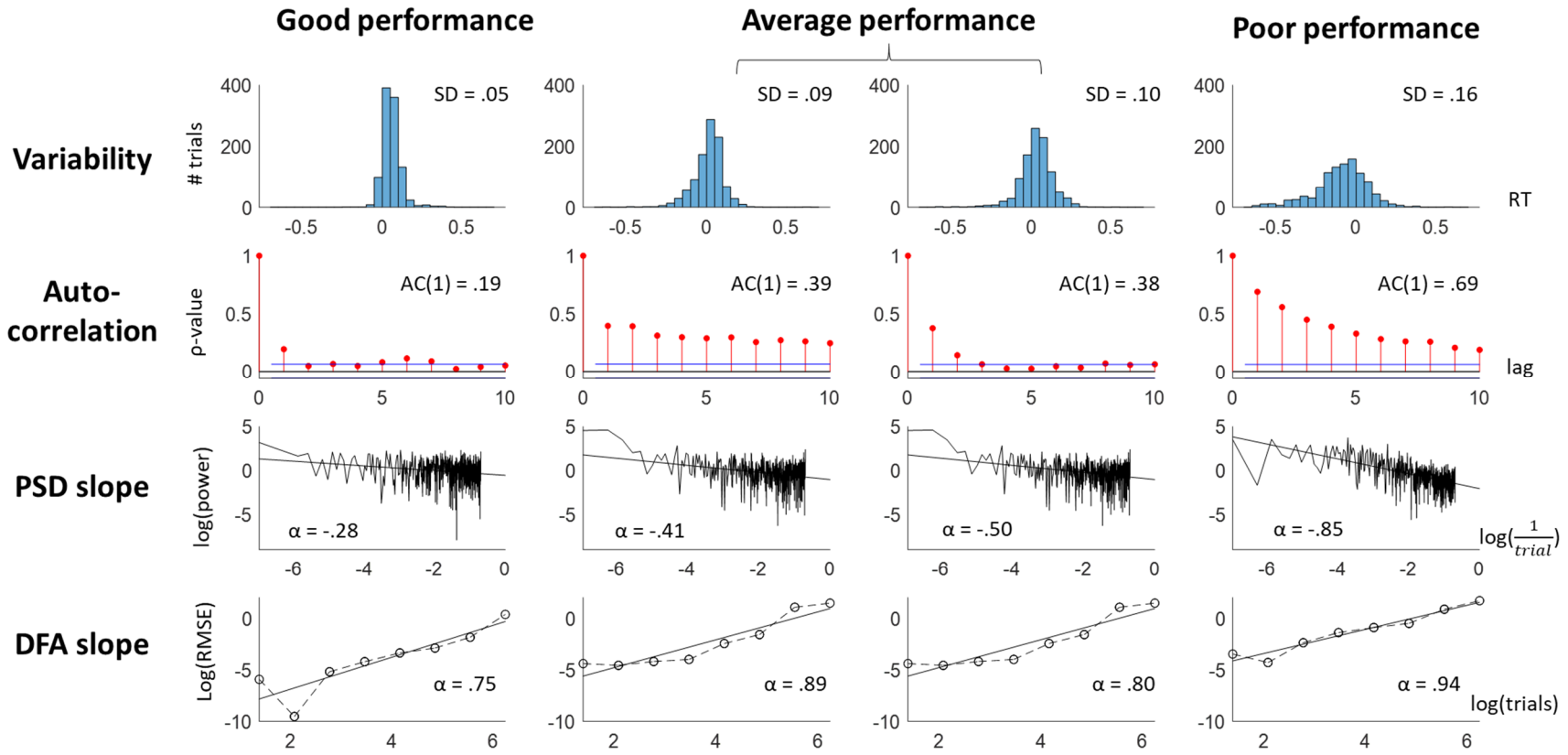


Figure 5. Examples from four participants over the temporal dependency measures, representing (left to right) good, close-to-median, close-to-mean, and poor performance. Good performance was associated with low temporal dependencies, as reflected in low autocorrelation and shallow PSD and DFA slopes, while poor performance was associated with high temporal dependencies, as reflected in high autocorrelation and steep PSD and DFA slopes. Note that the DFA plots have not been corrected for the shuffled RT series.

Subjective attentional state ratings

Figure 4 (right column) shows the between-subject correlations of temporal dependency measures with mean and variability of the subjective attentional state ratings, with corresponding r-values and BF_{10} in Table 2. There was moderate evidence against all the correlation pairs – indicating that participants' ratings of their attentional state throughout the task did not correlate with temporal structures in behaviour.

Personality traits

Figure 6 shows the between-subject correlations between temporal dependency measures with the questionnaire scores (ADHD tendencies, mind wandering tendencies, and impulsivity respectively), with corresponding r-values and in Table 3. None of the correlations were supported, and out of the 18 correlation pairs in total, 17 showed moderate evidence against a correlation. Furthermore, here was indeterminate (ADHD and mind wandering) to moderate (impulsivity) evidence against a correlation between the questionnaire scores and RT variability.

Table 3. Pearson r-values between the self-assessed personality traits and temporal dependency measures, with corresponding BF_{10} in brackets.

	ASRS	**DFS**	**UPPS-P**
Variability	-.08 (.18)	-.19 (.54)	-.08 (.17)
Autocorrelation	-.04 (.15)	-.04 (.15)	.001 (.14)
PSD slope	-.06 (.16)	-.09 (.20)	-.04 (.15)
DFA slope	-.03 (.16)	.04 (.15)	-.03 (.15)
AR	-.02 (.14)	.09 (.19)	-.17 (.40)
MA	-.01 (.14)	.10 (.20)	-.13 (.25)
d	.03 (.15)	-.01 (.14)	.20 (.59)

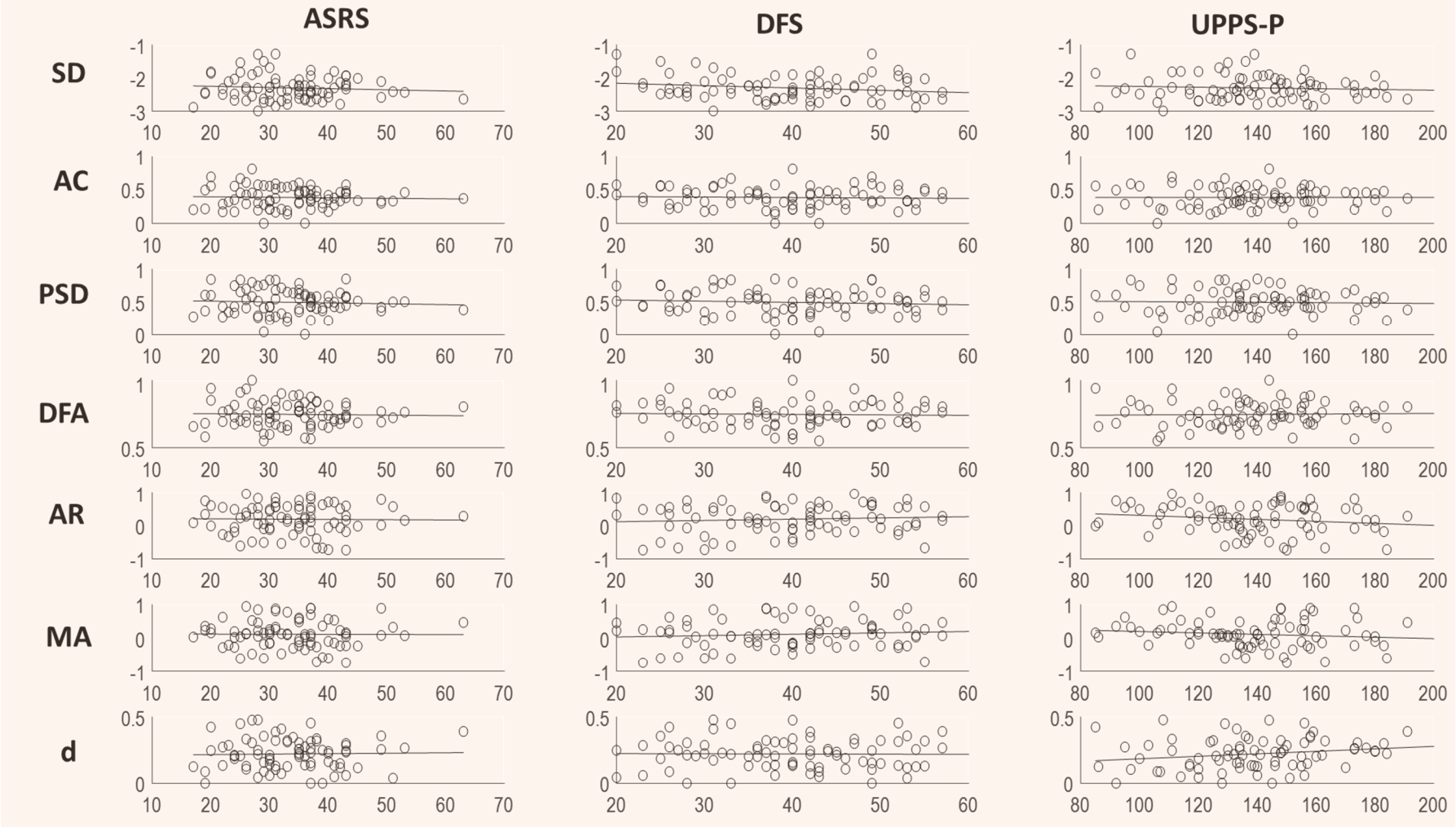


Figure 6. Between-subject correlations between the temporal dependency measures and self-assessed personality traits. Convention are the same as in Figure 3 and 4. None of the correlation pairs were supported.

Testing the presence of long-term dependency

As mentioned in the introduction, one benefit of the ARFIMA(1,d,1) model is its direct way to test the benefit of a long-term parameter over only short-term parameters. To test this, AIC was calculated both for the ARFIMA(1,d,1) model and for the ARMA(1,1) model (following the procedure of Wagenmakers et al., 2004). The difference between ARMA(1,d,1) and ARFIMA was calculated, and is shown in Figure 7 for each participant (black points on left panel). Values above 0 indicate a better AIC for the ARFIMA model, while values below 0 indicate a better AIC for the ARMA model. In practice, however, only values larger than 2 are taken as clear support for one model over the other.

For the 78 analysed participants, the long-term model was clearly favoured for 59 of them (~75%). When using the more conservative goodness-of-fit measure BIC instead (as recommended by Torre, Delignières & Lemoine, 2007), the long-term model was still clearly favoured for 47 participants (~60%; right panel). As a control analysis, we ran the same analysis on the same 100 shuffled RT series previously used to correct the PSD and DFA analyses. The AIC and BIC differences for the mean of these 100 RT series show no clear preference for either model.

As the d parameter showed large intra-individual differences, we were interested to see if the magnitude of the parameter would be informative for the fit of the long-term model. Between-subject correlations were conducted between the AIC/BIC differences and the d parameter (Figure 7). Indeed, we found a positive correlation for both fit measures – indicating *d* is more likely to add substantial fit to the long-term compared to the short-term model if its value is higher.

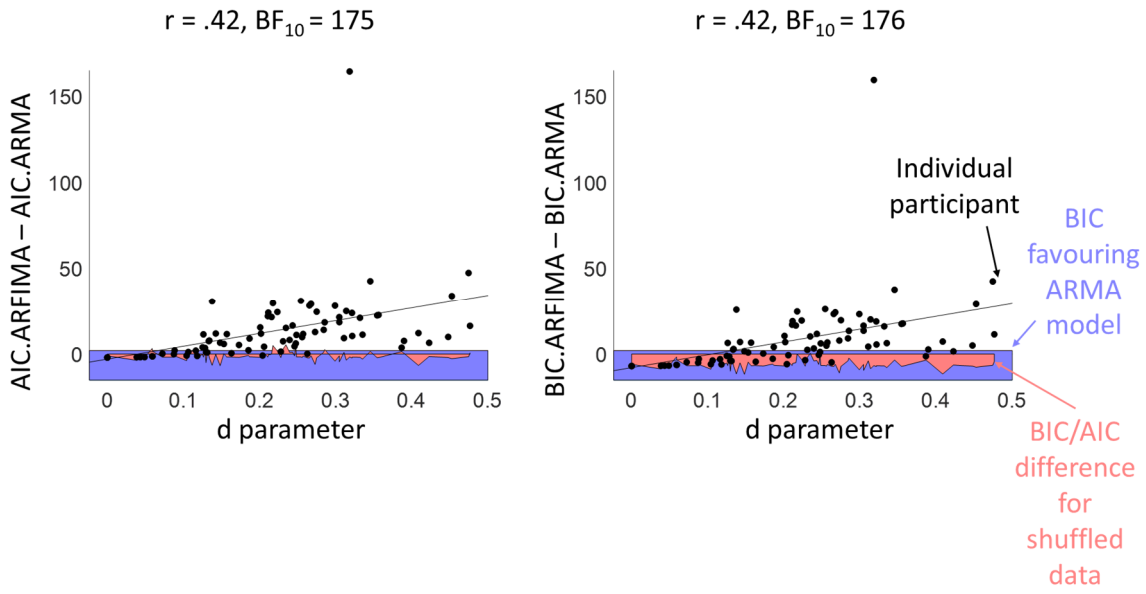


Figure 7. Shown are the difference in AIC (left) and BIC (right) between the ARMA(1,1) model and the ARFIMA(1,d,1) model, with each dot representing one individual subject. For points above the blue-shaded area (difference score of 2), the AIC/BIC clearly favours the ARFIMA(1,d,1) model – indicating that the d parameter adds substantial explanation to the model. This was true for most of the participants. The difference in AIC/BIC was related to the magnitude of the d parameter – indicating that ARFIMA(1,d,1) models with higher values of d were more likely to be favoured over the short-term model.

Inter-measure correlations

To examine the relationships between the different temporal dependency measures, Bayesian Pearson correlations were calculated on the data from the first session. Table 4 shows the correlation coefficients and corresponding BF_{10}. The results show a mixed picture. The measures that showed high within-individual repeatability (autocorrelation at lag 1, PSD slope, and DFA slope) also correlate highly with each other. These three measures also appear to correlate with ARFIMA's d parameter, though to a lesser extent.

On the other hand, the AR and MA parameters correlate highly to each other, but not to the other measures – with most Bayes Factor indicating evidence against a correlation. Remarkably, the autocorrelation does not correlate with AR, even

though their functions should be highly similar. Furthermore, while the fit measures (A/BIC difference, as plotted in Figure 7) correlate with the ARFIMA parameters, they do not correlate with the autocorrelation or the PSD, and only correlate weakly with DFA – suggesting that even if the traditional measures (autocorrelation, PDS, DFA) indicate a high temporal dependency, this cannot inform us whether a long-term dependency parameter is actually statistically beneficiary.

Table 3. Pearson r-values with corresponding BF_{10} between the different measures of temporal dependency.

	PSD	**DFA**	**d**	**AR**	**MA**	**A/BIC diff**[1]
Autocorr.	.94 (2.1e+35)	.82 (8.7e+16)	.30 (5.27)	.08 (.18)	-.09 (.19)	.16 (.36)
PSD slope	-	.71 (3.7e+10)	.14 (.31)	.16 (.37)	-.06 (.16)	.03 (.15)
DFA slope	-	-	.58 (8.4e+5)	-.11 (.23)	-.12 (.24)	.30 (4.56)
d	-	-	-	-.27 (2.53)	-.01 (.14)	.42 (175.46)
AR	-	-	-	-	.93 (5.6e+30)	-.43 (283.99)
MA	-	-	-	-	-	-.34 (15.47)
A/BIC diff	-	-	-	-	-	-

1. Note that the correlations between AIC and the other measures and those between BIC and the other measures were identical, and therefore, they've been combined into one column.

Discussion

In the current research, we investigated the intra- and inter-individual correlates of temporal structures in endogenous RT with a wide array of time series analysis

methods (autocorrelation at lag one, PSD, DFA, and ARFIMA(1,d,1) parameters). Results showed that on the group level, the data series indeed showed temporal dependency – as reflected in positive autocorrelations (AC(1) > 0) and in steep long-range slopes (PSD slope > 0; DFA slope > .05) – indicating that performance is dependent across trials. The dependencies are fairly high, as has been commonly found in tapping tasks (Delignières et al., 2004; Gilden et al., 1995; Lemoine et al., 2009).

Looking at intra-individual correlations of these measures, we found that temporal dependency showed high repeatability over time, though this was dependent on which measure was used: autocorrelation and PSD were highly reliable, but the ARFIMA parameters were not. Similarly, on the between-subject level, we found that the objective measures of performance and the subjective measures of attentional state were highly repeatable over time. However, there was Bayesian evidence against correlations of the temporal dependency measures with both subjective attentional state and self-assessed personality traits. The temporal dependency measures (with exception of the ARFIMA(1,d,1) parameters) did correlate with performance – such that good performance was associated with lower temporal structures.

Intra-individual repeatability of temporal dependencies

Although the existence of temporal dependency in RT data has been shown repeatedly (e.g., Gilden, 2001; Van Orden et al., 2003; Wagenmakers et al., 2004), its properties have been underacknowledged. It has been suggested that the structures show individual differences (Madison, 2004; Torre et al., 2011; Simola et al., 2017), and that they may be used to distinguish healthy individuals from those with attentional dysfunction (Gilden & Hancock, 2007). However, the reliability of these individual differences remains largely unknown.

Our conclusions are similar to Torre et al. (2011), who found intra-individual consistency in DFA slopes is moderate at best. Running a reliability analysis on our DFA slopes gives a Cronbach's α of .59, which is in the same range as Torre et al. (2011). Furthermore, our results indicate that the autocorrelation and PSD measures were more reliable (with consistency explaining respectively ~58 and

53% of the total variance) than the DFA measure (with ~18% explained variance). We also found intra-individual consistency in the performance measures themselves. Similar intra-individual reliabilities in reaction time variability have been found previously across time and across different tasks (Hultsch et al., 2002; Saville et al., 2011; Saville et al., 2012; but see Salthouse, 2012). Subjective attentional state was the most reliable measure (~61% explained variance). None of the ARFIMA(1,d,1) parameters showed intra-individual repeatability – indicating they are likely not suited for studying stable individual differences in temporal structures (e.g., personality traits, biomarkers).

Individual differences in performance and attentional state

The ACF, PSD, and DFA measures (but not the ARFIMA parameters) furthermore correlated with performance, with good performance being associated with reduced temporal dependencies – matching Irrmischer et al. (2018). While they speculate these positive correlations between performance and temporal dependency could be due to attentional fluctuations, we found evidence against between-participant correlations: 1) between temporal dependency and subjective attentional state ratings that were measured throughout the task, and 2) between temporal dependency and self-assessed mind wandering tendencies in daily life. However, thought probes on the MRT have mostly been used as a within-participant measure, capturing the fluctuations over time, rather than as an average attentional state. To get an idea whether the probes correlated with temporal structure within individuals, we calculated the autocorrelations at lag 1 of the five RTs before each thought probe, and correlated these with the attentional state ratings (mirroring typical analyses on MRT data) – giving us one correlation coefficient for each participant. On the group level, these were not different from zero (BF_{10} = .22), providing further evidence against a relation between temporal dependency and attentional state. However, the PSD, DFA, and ARFIMA measures cannot be estimated on such small data series. For the autocorrelation, it is recommended that the maximum lag size *k* should not be less than 25% of the total observations *N* (Box et al., 2016). In our case, this recommendation is met (*k* = 1, *N* = 5) – although the analysis is likely not very powerful.

At first sight, our results diverge from Simola et al. (2017), who found higher temporal dependencies in good compared to poor performance. They reason that temporal dependency indicates higher mental flexibility, which allows for better performance on the Go/No-Go task. Our results do not necessarily oppose this reasoning: It is possible that participants with low mental flexibility perform better on the Metronome Task, as they are better at sticking to the consistent action throughout. If this is true, and if lower temporal structures indeed indicate lower mental flexibility, then our results fit with the conclusions of Simola et al. (2017). However, this hypothesis cannot be tested with the current data and therefore remains highly speculative. Furthermore, even if true, this would not explain the differences in results between Simola et al. (2017) and Irrmischer et al. (2018).

Individual and clinical differences

To our knowledge, the current study is the first to directly relate temporal dependency to self-assessed personality traits. Bayesian analyses showed evidence against correlations with ADHD tendencies, mind wandering tendencies, and impulsivity. Gilden & Hancock (2007) have previously related ADHD symptoms to divergent temporal structures. They recruited fifteen undergraduates and fourteen members from the Alcoholics Anonymous (AA). All completed a mental rotation task and were subsequently ranked according to RT variability. The nine least variable (all undergraduates) and the nine most variable (eight AA-members) were divided into a ‘low’ and ‘high’ group respectively. The RT series from these two groups showed substantial differences in temporal structures. However, apart from methodological issues with the analysis method (see Farrell, Wagenmakers et al., 2006 for a critique), there are a number of other problems that make the findings difficult to interpret.

First of all, the researchers note that no one in the ‘low’ group reported ADHD symptoms on the questionnaire, while participants in the ‘high’ group did. However, the questionnaire appears self-constructed, and its precise content, number of questions, validity, and reliability are unreported. Furthermore, the questionnaire scores were not explicitly related to the temporal RT structures. Therefore, it is premature to conclude that clinical attention-deficits relate to differences in temporal

structures. Thirdly, the two groups may have differed on a large number of aspects, such as age, educational level, social-economic status, lifestyle, cognitive functioning, and familiarity with psychological testing. This again makes it hard to pinpoint the findings specifically to attention-deficits. Finally, the data of the 'medium-variability' group was not used at all. This group was a mixture of undergraduates and AA-members, but their variability did not show clear differences between the two – making it unclear how robust the findings are. Overall, these findings and conclusions should thus be taken with caution.

Previous clinical studies have also looked at the temporal dynamics of RT in ADHD (e.g., Castellanos et al., 2005; Geurts et al., 2008; Johnson et al., 2007; Karalunas et al., 2012; see Karalunas et al., 2012; 2014 for reviews; see Kofler et al., 2013 for a meta-analysis), but their methods and aims have been different than described in the current research. Rather, given that people with ADHD are more variable than neurotypicals, these studies have examined whether this increased variability is driven by rhythmic fluctuations – i.e., if the longer RTs observed in ADHD are temporally predictable. Typically, the RT series are transformed to the frequency domain by either a Fast-Fourier or Morlet-wavelet transform to obtain a power spectrum – to see if people with ADHD have higher power within certain bands (focusing on low frequencies, < 1.5 Hz). These performance rhythms may subsequently be associated with underlying neural/bodily rhythms (Adamo et al., 2015; Castellanos et al., 2005). However, this line of research mostly shown that increased variability is reflected in all (low) spectral bands, rather than in specific ones (see Karalunas et al., 2012 for a review; see Kofler et al., 2013 for a meta-analysis) – though the examined ranges appear to be highly limited. Future studies may investigate the structure of the entire series with a larger range of analysis methods.

ADHD in children has also been associated with *reduced* autocorrelations in RT compared to healthy samples (Aase & Sagvolden, 2005; Aase, Meyer & Sagvolden, 2006). However, both studies used reinforcement learning tasks – for which performance may exhibit both positive and negative autocorrelations, with the temporal structures reflecting the training process. Results from learning tasks cannot be easily generalised to tasks that are very simple and/or have already been highly trained, such as the Metronome Task.

Overall, it remains largely unknown what drives the individual differences in temporal structure. It is important to note that the current study used healthy participants, who do not typically report clinical levels of ADHD symptoms. Oversampling for high ADHD tendencies, testing clinical samples versus healthy controls, or testing the effects of medication on clinical samples could still uncover differences in temporal structure.

Different measures of temporal dependency

All of the analyses methods used in the current study are so-called 'fractal methods' and are mathematically derivable from each other (Stadnitski, 2012). Similarities in results may therefore be expected. Still, we found important differences over the methods, both in their properties (reliability and relationship to performance) as well as in the extent to which they correlate with each other.

Our choice of methods was dictated by those previously used on cognitive data. However, these methods: 1) are not exhaustive – other methods, such as Rescaled Range Analysis and Dispersion analysis, fall under the same subclass (see Delignières et al., 2006; Delignières, Torre & Lemoine, 2005 for overviews) – and 2) may come with a number of variants and refinements. Furthermore, the analyses methods in the current research are all for capturing linear trends in the data over different time windows. Non-linear methods may capture more nuanced temporal trends in the data, and have been used previously on RT data (see Kelly, Heathcote, Heath, & Longstaff, 2001). Again however, these methods have been hardly used on psychological data, and overall, non-linear trends in RT series are difficult to capture.

Particularly striking is the extremely high correlation between the autocorrelation at lag-1 and the PSD-slope, with almost 90% shared variance. This implies that when studying individual differences, fitting a slope over the power of the *entire* time series (in this case: a range of 1050 trials) give little additional information to simply correlating each trial to the next. It is clear that the PSD method is not more informative, despite being more computationally heavy, less intuitive in interpretation, and hence more difficult to implement in practical contexts (e.g., physicians working with patients). Comparisons between the goodness-of-fit of the

ARMA(1,1) to the ARFIMA(1,d,1) models (Torre et al., 2007; Wagenmakers et al., 2004) showed the ARFIMA models were favoured – indicative of the presence of long-term structure. The high correlation between autocorrelation at lag 1 and PSD therefore seems to reflect that the time series contain clear, stable long-term structures that are also reflected in the short-term structures.

However, it should be noted that, as the ARFIMA parameters were not repeatable within individuals, the model may be more difficult to interpret. One possibility for this lack of reliability is that the model has to estimate three parameters at once. To test this, we fitted each of the parameters separately (e.g., fitting an ARFIMA(0,d,0) to estimate *d*) over the two session. Indeed, these parameters did show high consistency (AR: r = .77, BF_{10} = 689; MA: r = .77, BF_{10} = 538; d: r = .77, BF_{10} = 620). Furthermore, the single AR parameters were the exact same values as the autocorrelation at lag 1 – as one would expect.

As such, the individual parameters get altered when estimated together to obtain a better numerical fit. While this is not necessarily surprising, it does raise questions about the biological plausibility of the model: As short-term dependencies in behaviour (and neural activity) are much easier to explain than long-term dependencies, modelling may instead take an approach in which the short-term parameters are fitted first, and the contribution of a long-term parameter is assessed afterwards. Future work should investigate how this could be achieved.

While the autocorrelation and PSD have the highest repeatability, the DFA may come with more flexibility. One can decide on how many time windows to take into account, and whether these should overlap or not. The fitted slope can be plotted over the windows (see Figure 5 for examples), which allows one to directly assess the fit. Based on this fit, the window size can be adjusted (see Kantelhardt, Koscielny-Bunde, Rego, Havlin & Bunde, 2001; Krzemiński, Kamiński, Marchewka & Bola, 2017 for examples). This ensures the obtained slope actually matches the data – something which is not clear in the PSD slopes (Wagenmakers et al., 2004; Torre et al., 2007). However, this flexibility also has its drawbacks: It can make it more difficult to compare and replicate findings across studies. For example, Irrmischer et al. (2018) used windows of 2-60 RTs on the go-trials (but note that go-trials only occurred every 4 to 10 trials, meaning their DFA slopes are not calculated

on adjacent trials) with 50% overlap between the windows, while Torre et al. (2011) used a maximum window of 256 trials (on a series of 512 trials) without overlap, and Simola et al. (2017) used windows of 30-300 seconds without overlap. While none of these analysis choices are necessarily wrong, it is clearly difficult to compare these findings, which stands in the way of replicability. Ideally, it should be reported how these choices were decided on, and how different choices may or may not alter the results.

Missing values in the time series

Regarding the extraction of the different measures, the issue of missing values (i.e., missed responses on trials) has been scarcely addressed. There appear three methods to deal with this issue. One can: 1) exclude the missing values entirely from the series (which appears the most common option in the literature), 2) replace the missing values by values that stay true to the distribution of non-missing values (for instance by using the median value, or a value obtained by statistical interpolation; see Adamo et al., 2015 for an example), or 3) replace the missing values by the most extreme value (e.g., the maximum response time).

The issue of missing values has been mentioned previously by both Kofler et al. (2013) and Karalunas et al. (2012; 2014). They rightly point out that the use of different methods across articles complicates comparison of results. We would like to take this one step further: As soon as the time series have a lot of missing values, interpretation becomes more difficult no matter which method is used. This is due to what missed responses possibly represent: extreme cases of poor task performance. By excluding the missing responses or by replacing them with average values, it appears that the participants are doing better than they actually are – by disregarding the moments in which they were doing the task so poorly that they did not respond at all. In other words, imputation of missing values only gives unbiased estimates when the missing values are ‘missing at random’, which is typically not the case in these experimental tasks – meaning there is no reliable way of estimating their values (see Donders, van der Heijden, Stijnen & Moons, 2006 for a review on data imputation). By replacing the missing values with the most extreme values, this

issue is solved, as the missing values are being represented by extremely poor performance on that trial. However, this method takes a toll on the RT distributions.

It should be emphasised that this problem is not trivial – particularly when studying clinical samples compared to healthy controls. It is a fair expectation that clinical populations show more missing responses – meaning that any method of dealing with the missing values will introduce systematic group differences unrelated to the temporal structures in the time series.

The current results are based on the time series with the missing values excluded. We reran the analyses with two different methods: 1) replacing the missing values within participants with their median RT, and 2) replacing with an RT of 650 ms (reflecting the maximum time a participant has to respond), Both of these methods gave mostly similar response patterns, but also a few substantial exceptions: 1) for the median replacement, the DFA slope was more reliable ($r = .60$, $BF_{10} = 11.9$), and 2) for the high replacement (650 ms), the d parameter showed good reliability ($r = .51$, $BF_{10} = 3.7$), and AR and MA parameters did correlate to variability ($r = .32$, $BF_{10} = 8.9$, and $r = .31$, $BF_{10} = 5.6$, respectively), though they remained unreliable over time. These changes occurred despite the amount of missed responses being low for most participants (group median < 1%), and after the participants with more than 10% omissions were excluded from analyses. One explanation for these increases in reliability may be that the number of omissions is itself a reliable trait ($r = .52$, $BF_{10} = 4.1$) – although this would not explain why the increase is not found in all the measures (e.g., for the median replacement, reliability of d instead went down, $r = .26$, $BF_{10} = .5$). As there is no straightforward way of dealing with these missing values, it may be recommended to also report alternative methods – particularly when the amount of missed responses is high and/or different across compared groups.

Conclusion

In the current study, we found good intra-individual reliability of temporal measures, though there were differences between the different measures. In particular, the

autocorrelation may be best suited as a potential biomarker, as its reliability is good, and the measure is relatively easy to implement in practical settings. While we found no correlations of the autocorrelation with self-assessed ADHD, mind wandering, or impulsivity, previous studies have hinted at differences when particularly studying clinical cases of ADHD versus healthy controls. The more reliable measures did correlate to performance, but the underlying mechanisms are still unknown. However, small changes in the analysis pipeline may lead to substantial changes, highlighting the importance of transparent reporting.

It should be noted that the less reliable measures (DFA and the d parameter) may come with other benefits, but remain very difficult to interpret. In particular, it is unclear how the measures behave under different conditions (e.g., different cognitive loads, different response strategies, different attentional constraints), which gets in the way of coming up with falsifiable hypotheses. This lack of clear predictions in the study of temporal measures has been raised previously by Wagenmakers et al. (2012). Some may argue that the temporal structures should manifest similarly under different conditions – reflecting their 'ubiquitous nature' – but this would make the measures mostly uninformative. Future research may therefore instead aim to directly manipulate the temporal structures with different experimental conditions, rather than just measure them – to get a better picture of their underlying neural-cognitive mechanisms.

Chapter 3

Examining the links between subjective attentional state ratings, behavioural variability, and underlying neural states

Abstract

Whatever task we are performing, our attention fluctuates between on-task and off-task focus – a phenomenon described as 'mind wandering'. Subjective reports of off-taskness have been positively associated with reaction time variability. These findings match common intuitions that there is a strong link between attentional state and performance. However, effect sizes are typically weak, and the exact link between attentional state and behavioural variability remains unclear. While some EEG studies have investigated neural states preceding mind wandering, these studies have: 1) largely scattered methodologies and findings, 2) not examined the direct link with processes of behaviour, and 3) not distinguished mind wandering from other forms of inattention, such as fatigue. The current research is the first MEG study to investigate the relationships between behavioural variability, meta-cognition, and underlying neural states. Twenty-one participants performed the Metronome Task, in which they press a button every three seconds. Subjective attentional states and performance were measured with quasi-randomly presented probes. Objective and subjective task-measures were correlated with preceding oscillatory activity across trials. We found that subjective ratings correlated with

behavioural variability, though effect sizes were low. Subjective attentional state ratings and behavioural variability – but not performance ratings – could be predicted from neural states at the group level, but with large intra-individual differences. Furthermore, we found overlap across all frequency band in the neural states underlying performance ratings and behavioural variability, as well as those underlying performance and attentional state ratings. For attentional state ratings and behavioural variability, there was only overlap in the β band. Overall, our results show that metacognitive ratings and performance are associated with each other, but overall are poor markers of each other.

Keywords: mind wandering; mind blanking; attention; intra-individual variability; noise; MEG; oscillations

Introduction

Whatever task we are performing, our attention on it is never stable over time. Regardless of whether the task is extremely boring (like doing one's administration) or extremely interesting (like reading a good book), people will experience periods in which their thoughts are fully on the task, and periods in which their thoughts are instead focused on what to have for dinner, or what to get their mother for her birthday, or on their crippling financial debts and accompanying anxieties, or on any other thought unrelated to the task. Thoughts like this may be described as 'internally-driven', as they seem to appear spontaneously from 'our own mind' without being invoked by an external stimulus. This is evident in psychological experiments, in which participants are tested in a room devoid from any external distractions, but still experience these attentional fluctuations.

In the last decade, this phenomenon of task disengagement has been increasingly studied as the phenomenon of 'mind wandering', which is often

described as a shift of cognitive and executive recourses from the task to task-unrelated thoughts (see for instance Seli et al., 2013; Smallwood & Schooler, 2015). Mind wandering has been associated with behavioural task-performance. This makes sense on an intuitive level: Just like attentional states, performance fluctuates over time in a partly-spontaneous manner. Indeed, when we notice that we are in a period of poor task performance, we often tend to verbalise this like: *"I am performing poorly right now because I'm not paying attention".* There has therefore been a practical interested to reduce both mind wandering and behavioural variability as much as possible within certain contexts – such as traffic and air control, where they might have severe detrimental effects. However, the exact theoretical link between mind wandering and performance, and their relation to related concepts – such as attention, fatigue, and boredom – are largely unspecified still.

A few studies have investigated neural states preceding subjective reports of off-taskness (reviewed below), but their methodologies have been very scattered, with overall inconclusive results. Different types of off-taskness (such as drowsiness or mind blanking) have not been taken into account; rather, any subjective off-taskness has been analysed as mind wandering. Furthermore, while most of these studies did look at the associations between mind wandering and behavioural performance, most of them have not looked at the neural states preceding behaviour. As such, the three-way link between neural states, subjective meta-cognition, and behaviour remains unclear. The current research will be the first magnetoencephalography (MEG) study investigating the neural processes of mind wandering. With this study, we examine whether performance and metacognitive reports can be predicted from preceding oscillatory power. Furthermore, we aim to investigate the relationships between objective measures of performance, subjective reports of perceived performance, subjective reports of attentional state, and preceding neural states. Overall, we aim to find out to what extent subjective attentional state and behavioural variability reflect truly similar processes.

Mind wandering and behavioural variability

Any repeatedly performed action will show substantial variability over time. Similarly to mind wandering, this is evident during psychological testing: Even on the simplest response tasks, participants show deviations from their own mean performance from trial to trial. While a small part of this variability may be caused by task-related features such as condition order and time on task (e.g., Bompas et al., 2015; Gilden, 2001), the largest proportion remains unexplained – and has therefore been referred to as 'spontaneous'.

This spontaneous variability has been linked to mind wandering. To study this, the 'probe-caught' method is commonly used: Participants are probed quasi-randomly throughout the task to enquire about their subjective attentional state just prior to the probe. Indeed, off-task reports are preceded by larger reaction time (RT) variability compared to on-task reports. More generally, the association between subjective attention ratings and performance has among others been found in detection tasks (MacDonald, Mathan & Yeung, 2011; Jin, Borst & van Vugt, 2019), vigilance tasks (Qin, Perdoni & He, 2011; Jin et al., 2019), reading comprehension (Schooler, Reichle & Halpern, 2004; Ward & Wegner, 2013), and driving (e.g., Baldwin et al., 2017). Although the link between off-taskness and variability appears to be found consistently, the effects tend to be quite small (for example, reports of being fully off-task showed a 3-7% increase in variability compared to on-task reports; Seli et al., 2013, Laflamme et al., 2018), and a large amount of RT variability remains unexplained.

Importantly, the intuitive association between subjective attentional state reports and performance seems so strong that they are often taken as highly similar – if not identical – processes. For example, in Qin et al. (2011), subjective thought probes and objective task errors are analysed respectively as 'subjectively reported mind wandering' and 'behaviourally indexed mind wandering' – as to reflect two different markers of the same underlying process. Even more so, Baldwin et al. (2017) speculate on the use of online detection of neural states underlying poor subjective attentional states during driving to subsequently prevent poor driving performance. However, for this approach to be fruitful, we need empirical evidence that mind wandering and performance are indeed strongly linked.

Mind wandering and neural activity

Electro- and magnetoencephalography (EEG/MEG) analyses fall broadly into two categories: 1) the activity evoked by a stimulus, and 2) ongoing oscillatory activity.

In the first category, subjective reports of off-taskness have been associated with reductions in parietal and fronto-central P3 (Baldwin et al., 2017; Kam et al., 2011; Smallwood et al., 2008) compared to on-taskness reports. Reductions in parieto-occipital P1 during visual (Baird et al., 2014; Jin et al., 2019; Kam et al., 2011; but see Smallwood et al., 2008) and central midline N1 during auditory (Kam et al., 2011) tasks have also been found. Furthermore, these off-task reports have been associated with a reduction in feedback error-related negativity (fERN) while monitoring feedback on correct (but not on incorrect) trials (Kam et al., 2012). The reduction in these early (visual and auditory) sensory components have been interpreted as a disengagement or 'decoupling' from the external environment, while the reduction in the later components have been interpreted as a reduction of task-focused involvement of cognitive and monitoring processes – the combination of which may lead to alterations in performance.

Fewer studies have focused on ongoing oscillatory activity preceding off-taskness (see Table 1 for an overview), with contradictory results. Macdonald et al. (2011) investigated the relationships between prestimulus α power at parieto-occipital areas and stimulus detection, subjective confidence ratings, and subjective attentional states on a difficult visual psychophysics task. On each trial, they asked participants about the presence or absence of a target, their confidence in their answer, and their attentional state during the trial in a 'four-grid response square' – with the top and bottom two grids respectively reflecting 'more focused' and 'less focused' (total scale ranging from 0-200), and the left and right grids receptively reflecting 'sure absent' and 'sure present' (also ranged 0-200). They found that prestimulus parieto-occipital α in the one second window preceding target onset was higher when participants reported poorer attentional states.

Similar results were found on a larger time scale in a study on driving behaviour on boring routes in a simulator, in which participants occasionally received thought probes with a binary response option (Baldwin et al., 2017). Parietal α power in a window of 10 seconds prior to the probes was compared

between on- and off-task reports, and was found to be increased for off-task reports. Frontal θ power was also examined, but no differences were found.

Table 1. Overview of the literature on attentional state ratings and preceding oscillatory power. Shown are the task, the manner in which off-taskness was assessed (including range of rating scale), the number of participants included in the final analyses (N), the analysed time window prior to the report or stimulus onset in seconds, and a summary of the findings on oscillatory power during off-task compared to on-task reports.

Article	**Task**	**Assessment (scale)**	**N**	**Window [sec]**	**Off- vs. On-task**
Macdonald et al., 2011	Visual detection	After every trial (0-200)	18	1	↑ parieto-occipital α
Baldwin et al., 2017	Driving simulation	Probe-caught (binary)	9	10	↑ parietal α
Qin et al., 2011	SART[1]	Probe-caught (binary)	14	4	No significant differences
Braboszcz & Delorme, 2011	Meditation + passive auditory oddball	Self-caught (*NA*)	12	10	↑ θ (full scalp) ↑ frontocentral δ ↓ fronto-lateral β ↓ occipital α
Jin et al., 2019	SART[1] & Visual Search	Probe-caught (6 options incl. on-task and mind wandering)	18	~ 30 to 36	↑ frontal + parietal α

1. Sustained Attention to Response Task.

However, Qin et al. (2011) also investigated differences in θ, α, β, and γ power between subjective attentional states on a sustained attention to response task (SART), but did not find any significant differences. Furthermore, opposite results were found in a study from Braboszcz and Delorme (2011), who investigated mind wandering during a breath counting meditation task. Rather than using probes, participants were asked to monitor their own attentional states, and press a button every time they noticed they had been mind wandering (known as the 'self-caught' method). Oscillatory power on a window of 10 seconds prior to the button press was compared to 10 seconds after the button press (representing respectively off- and on-taskness). Most prominent was a significant increase in θ power during off-taskness over the entire scalp (with the largest differences being in the occipital and parietocentral regions). Furthermore, they found increases of δ power (largest in the frontocentral regions), and decreases in fronto-lateral β and occipital α. The researchers interpret these findings as a decrease in task-related attention and vigilance (reflected in α and β) and an increase in slow frequencies that resemble the process of falling asleep (reflected in δ and θ).

It should be noted that the Braboszcz and Delorme (2011) study diverges from the other mind wandering studies in three important ways. First, participants performed this meditation task with their eyes closed, while the other studies were conducted with eyes open. Eyes-open versus eyes-closed has been known to strongly modulate occipital α power (Mo, Liu & Din, 2013) and may thus be an important confound. Secondly, participants had to perform a meditation-based counting task, meaning their task-related thoughts were 'internally driven', rather than dependent on an 'external' computer-driven task – and thus, may rely on very different neural mechanisms. Thirdly, this study was the only one to use a self-caught rather than a probe-caught method – although it is not clear to what extent this would modulate the *direction* rather than just the effect of the findings.

A recent study has combined the two analysis approaches (event-related and ongoing activity) to develop a machine learning classifier on EEG data, aiming to differentiate between the neural states underlying subjective on-taskness versus mind-wandering (Jin et al., 2019). Participants performed both a SART and visual search task, with thought probes presented during both. Included markers were parietal-occipital P1 and N1, parietal P3, and power and coherence for frontal,

parietal and occipital α and θ (separately for both hemispheres) – though it should be noted that the oscillatory power was measured from -400 to +1200 ms around stimulus onset, thus encompassing both pre-stimulus and event-related activity. Each of these measures were individually better than chance at predicting subjective on- versus off-taskness, but only frontal α and left occipital α reached the same performance level as the whole-head model (with the whole-head model still numerically superior). The whole-head model showed classification accuracy above 50%-chance level, though the mean accuracies were low (mean for SART = 64%, visual search = 69%) and showed large individual differences (ranged 50-85%). Interestingly classifiers were generalisable across the tasks (accuracy for classifier of SART on visual search = 60%, visual search on SART = 59%), suggesting neural states underlying mind wandering are partly task-independent.

Common issues in the mind wandering literature

All in all, findings on oscillatory power preceding poor attentional states have been scarce and contradictory, and have largely ignored the subsequent link to behaviour. Furthermore, most of them have not addressed what constitutes participants' subjective ratings – e.g., to what extent off-taskness ratings specifically represent mind wandering (rather than other forms of off-taskness), and to what extent the ratings can specifically capture the metacognitive experience of attentional state. Finally, effect sizes and individual differences are typically disregarded. Below, we address these issues in more detail.

Link to behaviour

Within the mind wandering literature, the association between attentional states and behavioural variability plays a key role. Indeed, as discussed above in the section *Mind wandering and behavioural variability*, the two phenomena are often seen as two markers of the same (or at least extremely similar) underlying processes. Still, this link appears largely based on intuition, as there is only weak empirical evidence to back this up. A typical paper will analyse the association of subjective ratings to performance and the association of subjective ratings with neural states, and will

subsequently assume there is a common link between the three (e.g., Qin et al., 2011; Baldwin et al., 2011).

Aside from the weak empirically observed link between subjective and objective measures, it is unclear why they would reflect the same process. MacDonald et al. (2011) found that subjective off-taskness was preceded by parieto-occipital α, but found no relationship between α and detection performance – though it should be mentioned that their stimuli were variable in contrast over trials, making the direct relationship between α and detection less straightforward. Even more so, Qin et al. (2011) found that off-task states were preceded by higher fronto-central γ power compared to objective errors – supposedly reflecting higher cognitive processes during mind wandering (unrelated to the task). If anything, these results appear to reveal differences rather than similarities between the subjective and objective 'markers'.

Different types of off-taskness

Throughout this introduction, both 'mind wandering' and 'off-taskness' have been used interchangeably. However, not all off-taskness consists of mind wandering; it is possible that one is not focused on the task nor on other thoughts. We can refer to this as 'mind blanking' – which is a separate phenomenological state (Ward & Wegner, 2013).

Research on neural states underlying off-taskness have typically not made this distinction – instead encompassing both 'mind wandering' and 'mind blanking' into a general 'off-taskness' (with exception of Jin et al. (2019), who explicitly only examined on-task versus mind wandering). However, it cannot be a priori assumed that mind wandering and blanking are driven by the same neural mechanisms – particularly as only the former has a clear cognitive element. Two recent studies highlight this in particular. First, van den Driessche et al. (2017) found that ADHD patients typically report more mind blanking, but not mind wandering, compared to healthy controls. Furthermore, medication specifically reduced mind blanking episodes to a neurotypical level. Secondly, in healthy samples, mind blanking has been associated with smaller pre-trial pupil diameter (Unsworth & Robison, 2018),

and with differentiated effects on performance behavioural performance compared to mind wandering (Unsworth & Robison, 2016; 2018; Ward & Wegner, 2013). In the current study, we therefore asked participants to classify the content of their off-task states, to examine the underlying neural states separately.

Different metacognitive experiences

Another reason why off-taskness ratings may be difficult to interpret is that, due to their subjective nature, they may reflect a mixture of metacognitive experiences. For instance, when participants feel that they are performing the task well, they may be more inclined to report being on task, but if they are making a lot of mistakes, they may be more inclined to report being off task. In this case, (part of) the correlation between subjective reports and variability may be driven by performance- rather than attention-monitoring. Indeed, Macdonald et al. (2011) found an association between subjective ratings of detection confidence and subjective ratings of attentional state ratings – although in this case, participants reported both simultaneously with one rating on a response grid. In our study, we extended the probe-caught method with a second question, on which participants were asked to rate their own performance – allowing for a closer examination of the reported experience of attentional state.

Individual differences

Last but not least, the literature on mind wandering often lacks calculated effect sizes and – related to that – has disregarded individual differences. However, if one is interested into what extent off-taskness states can be predicted *within individuals*, the group average is only interesting in so far as it is a good reflection of the individuals. Therefore, we examined both group and individual levels.

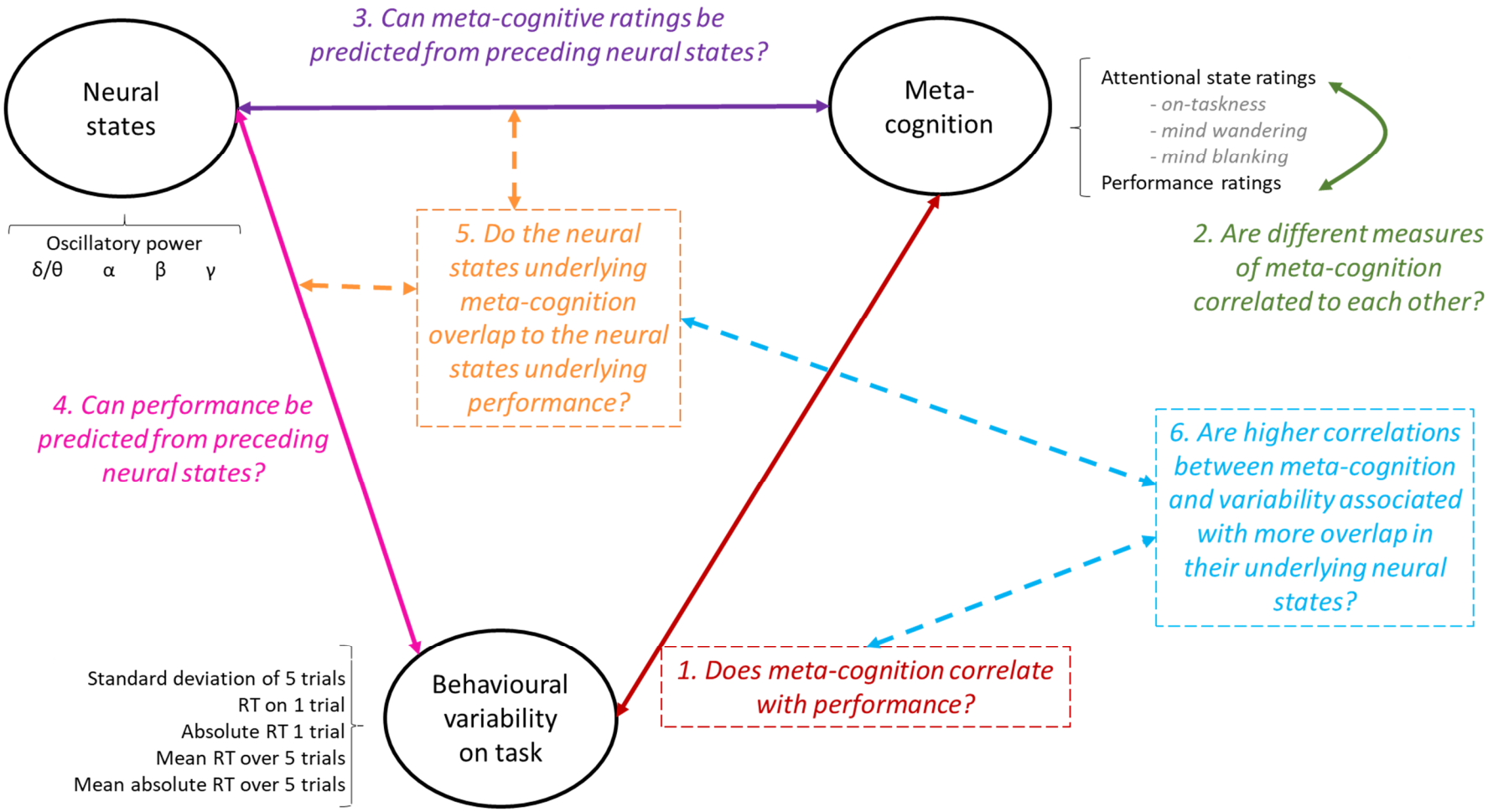


Figure 1. Graphical representation of the measures and research questions of interest in the current study, with each colour representing one question.

Current research

In the current study, we use MEG to investigate the underlying oscillatory processes, which, compared to EEG, comes with a higher signal-to-noise ratio, especially at higher frequencies, and with enhanced localisation (Baillet, 2017). We investigate six research questions (see Figure 1 for a graphical overview), described below in more detail. These sub-questions lead us to one overall aim – examining to what extent subjective attentional states and behavioural variability are markers of the same underlying processes.

Our participants performed an adjusted version of the Metronome Task (MRT) from Seli et al. (2013), in which they heard a tone every three seconds, and were instructed to press a button in synchrony with the tone. Due to the monotone and simple character of the task, it is thought to be well-suited for eliciting mind wandering (Cheyne, Carriere & Smilek, 2006; Giambra, 1995). Throughout the task, they pseudo-randomly received thought probes, in which they are enquired both about their attentional state and their subjective estimate of their performance just prior to the probe. We examined the relationship between different measures of metacognition and performance, and their links to neural states. Aside from typical group analyses, we were also particularly interested in individual effects.

Question 1. Does meta-cognition correlate with performance?

In the current study, we aim to replicate the positive association between subjective off-taskness and performance within participants. Note that previous studies have solely reported the association with SD (Laflamme et al., 2018; Seli et al., 2013). However, as we are interested in the extent to which subjective states and performance relate to similar processes, we want to find the best behavioural correlate of the subjective ratings. Therefore, we examine multiple markers of performance.

Question 2. Are different measures of meta-cognition correlated to each other?

While the MRT is simple to perform, participants do not have *exact* knowledge of their own performance – as opposed to SART or speeded detection tasks, in which participants are at least partly aware of the errors they make – allowing us to examine whether experiences of attentional state and performance are similar.

Furthermore, we can examine the performance correlates of the subjective performance ratings with the same analyses as Question 1, to compare the associations of different metacognitive reports.

Question 3. Can meta-cognitive ratings be predicted from preceding neural states?

Prior results on ongoing oscillatory activity preceding subjective off-taskness ratings remain highly inconclusive. Furthermore, they are often focused on a subset of frequency bands and on a limited number of electrodes. We used the advantages of the MEG by taking a whole-head approach over the different frequency bands. To investigate the link of subjective ratings with oscillatory power, data was source reconstructed using correlational synthetic aperture magnetometry (SAM-R; Bompas et al., 2015): Within participants, each of behavioural measures was correlated across trials to the ongoing oscillatory power preceding the response. This analysis was run separately for the attentional state ratings and the performance ratings. As we were interested in different types of off-taskness, we ran also ran these analyses separately for mind wandering and mind blanking.

Question 4. Can performance be predicted from preceding neural states?

The same SAM-R analyses were also run on objective performance. Anticipating on the behavioural analyses, we found that both subjective attentional state and performance correlated best to the SD of the last five RTs prior to the probe. Therefore, we focused our SAM-R analyses on this marker. This measure has also been consistently used in previous literature on this task, though it should be mentioned that it is not measure participants are instructed to focus on (but see Supplementary Materials, p. 145-146).

Question 5. Do the neural states underlying meta-cognition overlap with the neural states underlying performance?

Crucially, after finding the neural states underlying both subjective ratings and behavioural variability separately, we examine whether these neural states are actually overlapping. Therefore, we examine the associations between the neural states of subjective ratings and the neural states of behavioural variability. If the link between subjective off-taskness and poor performance is indeed as strong as intuitively assumed, there should be a high overlap in their neural states.

We performed these analyses for the overlap between: 1) attentional state ratings and behavioural variability, 2) performance ratings and behavioural variability, and 3) attentional state and performance ratings.

Question 6. Are higher correlations between meta-cognition and variability associated with more overlap in their underlying neural states?

While Question 5 refers to the amount of overlap at the group level, the individual effects are of interest here. Specifically, if there is a link between subjective off-taskness and poor performance, participants who show a high association between their subjective ratings and performance (Question 2) should show more overlap in the neural states of these two – reflecting that if subjective attentional state and performance are highly coupled, their underlying processes may show a strong overlap. On the other hand, participants who do not show this link between ratings and variability may also not show neural overlap.

Methods

Participants

Twenty-one participants (eighteen female, 21-40 old, M_{age} = 26.3) were tested in a two-session experiment. All of them had normal or corrected-to-normal vision and hearing, and none suffered from a psychiatric or neurological disease. Participants

were paid £10/hour (excluding potential reward). The study has been approved by the local ethics commission. Due to technical issues, one participant only completed two blocks instead of three on her second session, and for another participants, the behavioural responses of one session were not recorded.

Materials and apparatus

The experiment was generated outside the magnetic room using MATLAB version 8 (The Mathworks, Inc., Release 2015b) and Psychtoolbox-3 (Brainard, 1997; Kleiner et al., 2007; Pelli, 1997), with a HP Z230 Workstation PC, and was displayed to the participants on a projector screen inside the magnetic room. The background was set at light-grey with a white fixation dot in the centre. Participants were seated 1185 cm away from the screen, with their head on a chinrest to limit head movement. Eye movements were recorded binocularly at 500 Hz with an Eyelink 1000. Responses were recorded with a Nata Technologies 2-hand Button System.

Whole-head MEG activity was recorded at 600Hz with the CTF-Omega 275 channel radial gradiometer system (VSM MedTech). Each trial was manually screened to reject any artefacts causing excessive noise. Structural MRI scans were coregistered to the MEG data with use of three fiduciary markers at fixed positions near the left and right tragus and the eye centre. Photographs were made in order to verify the fiduciary positions afterwards.

Design

Reaction time (RT) was measured on each trial on the adapted version of the Metronome Task (Seli et al., 2013) relative to the tone. Subjective ratings of performance and attentional state were both measured with quasi-randomly presented probes on a scale from 1 to 9.

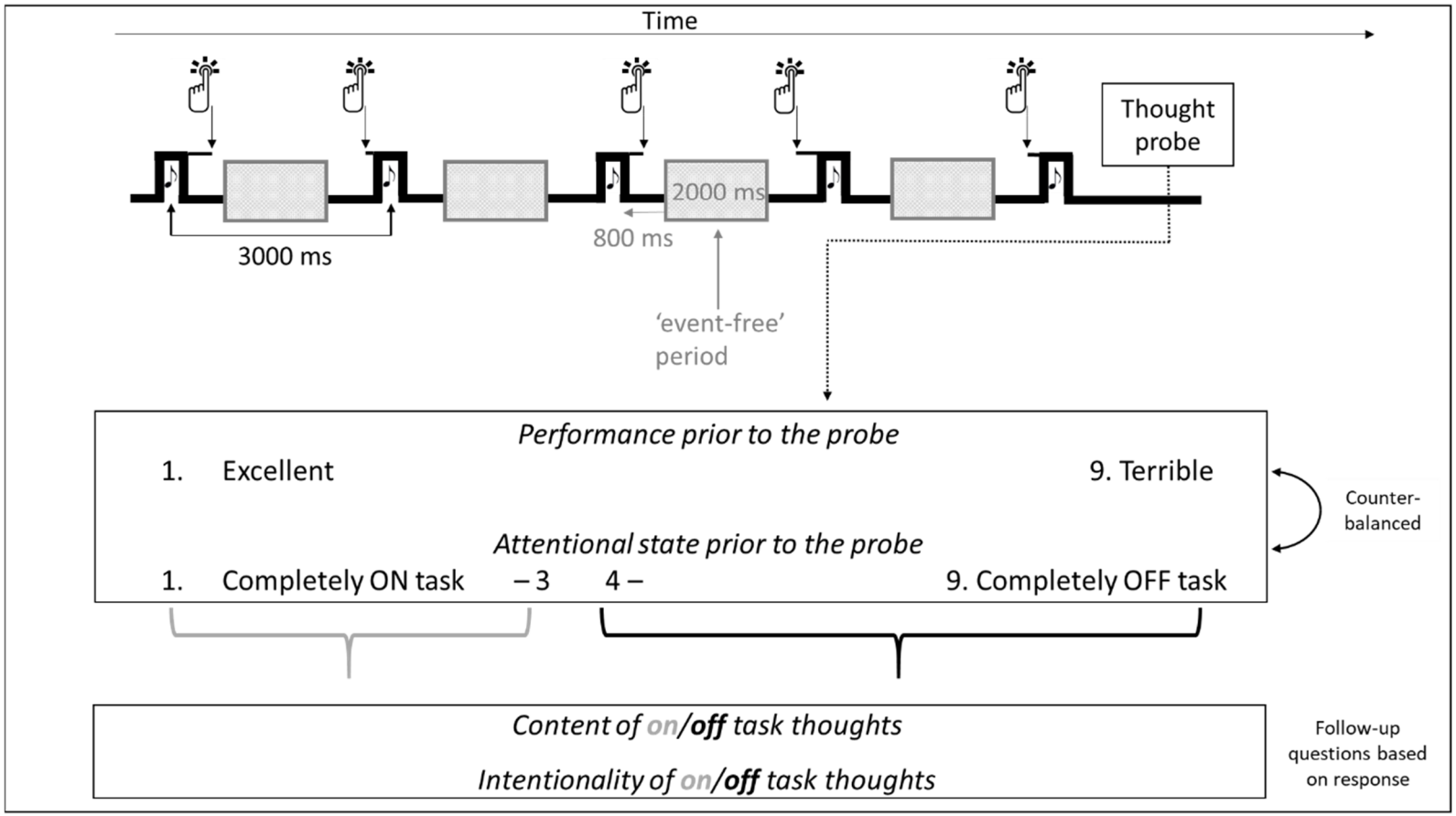


Figure 2. Overview of the Metronome task (figure adapted from Seli et al., 2013), and the four questions with which participants were probed quasi-randomly throughout the task. Participants were instructed to press in synchrony with a tone, occurring every three seconds, and their relative reaction times were recorded. The 'event-free periods' between the tones and presses were used to estimate ongoing neural activity.

Procedure

The experiment consisted of two experimental sessions, both lasting about 1.5 hour. Before and after the task, participants conducted a 4-minute resting state, in which they were instructed to fixate on the dot in the centre of the screen. Eye movements and pupil dilation were tracked throughout the experiment.

Figure 2 shows a snapshot of the experiment over time (adjusted from Seli et al,. 2013). Each trial lasted 3000 ms, and 1500 ms after trial onset, a short tone of 75 ms was presented. Participants were instructed to press in synchrony with the tone – meaning that perfect performance is indicated by a response time of 0 ms in relation to the tone. The relative time from response to tone constituted the objective RT and SD measures of performance.

For the subjective measures, participants were quasi-randomly presented with thought probes throughout the task. Each time, they were given four questions. The first two questions referred to their subjective attentional state (*"Please record a response from 1 to 9 which characterises how on task you were just before this screen appeared"*, with 1 as 'completely ON task' and 9 as 'completely OFF task') and performance (*"Please record a response from 1 to 9 which characterises how you rate your performance before this screen appeared"*, with 1 as 'excellent' and 9 as 'terrible'). The order of these two questions was counterbalanced over sessions and participants.

Next, the participants were presented with two follow-up questions on content and intentionality, depending on their attentional state rating. Any rating above 4 was considered 'off task', and was followed by the same question on content with different response options – 1 as 'bodily states' (i.e., thinking about being uncomfortable in the MEG), 2 as 'mind wandering' (i.e., off-task thoughts about anything else), 3 as 'mind blanking - alert' (i.e., thinking about neither the task nor anything else, but still in alert state), and 4 as 'mind blanking – drowsy (i.e., about neither the task nor anything else due to being too tired/falling asleep) – and a question on intentionality (*"Would you say your off-task thoughts were: 1. Intentional: Conscious off-task thoughts, or 2. Unintentional: Off-task thoughts despite your best efforts"*). Any attentional state rating between 1 and 3 was considered 'on task', and was followed by the questions: (*"Please indicate with the*

corresponding number what you were thinking about specifically?", with 1 as 'the tones', 2 as 'the key presses', 3 as 'the instructions of the task', and 4 as 'your performance on the task') and (*"Would you say your on-task thoughts were: 1. Intentional: Conscious on-task thoughts, or 2. Accidental: On-task thoughts without your best efforts"*). Note that the follow-up questions to on-task states were only introduced to equate the number of questions in all conditions, the answers were not analysed.

Both sessions had the same structure: The main experiment consisted of 24 blocks of 30 trials, resulting in (2 * 720 =) 1440 trials per participant. Each of these blocks contained one thought probe, resulting in (2 * 24 =) 48 thought probes per participant. The probe was presented on a random trial within the block, excluding the first and last five trials of the block – to ensure the probes were not administered too close together. The five trials following each probe were not included in the calculation of standard deviations measures and all related analyses. Prior to the main experiment, participants conducted a training of 60 trials, with a thought probe being presented after trial fifteen. After the training, the experimenter checked the performance to ensure the participant understood the task, and gave the participant feedback on their performance.

To give the participants motivation to perform as well as possible throughout the task, we used the random lottery reward system (Cubitt et al., 1998). At the end of a session, one random trial number was extracted. If the moving window standard deviation on this trial was lower than .075, the participant received £5 extra. This cut-off was based on pilot data, such that the chance of reward would be ~20%.

Results

Before conducting any analyses, we were interested in the amount to which participants reported being on- versus off-task, as well as in the content of their off-taskness. Figure 3 shows the percentage of on-task reports over the two sessions, as well as the breakdown into percentages of off-task reports, at the subject and group levels. Particularly striking is the large variability between participants, both in

amount of on- versus off-task reports and in content. At the group-level, participants reported to be more on- than off-task. When they reported being off-task, 'mind wandering' was the most common category. 'Drowsy mind blanking' and distractions due to 'bodily sensations' were rarer, but did occur. As such, we replicate prior findings showing that not all reported off-taskness represents mind wandering episodes (Unsworth & Robison, 2016; 2018; Ward & Wegner, 2013).

One participant never reported to be on-task, and another participant only reported to be off-task once throughout both sessions. These two participants were excluded from all analyses that involved the attentional state ratings.

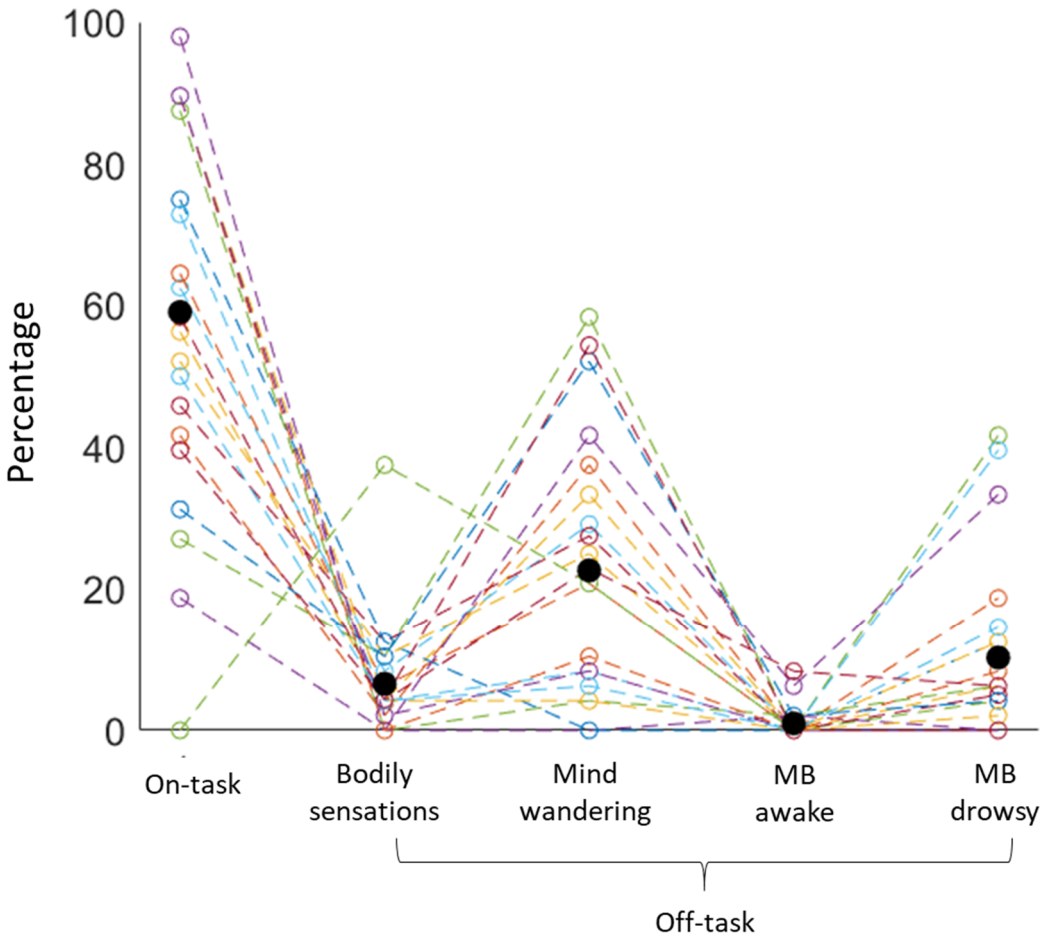


Figure 3. Percentages of on- versus off-task thoughts throughout the two sessions, with off-task thoughts broken down according to content as reported in the follow-up question. Shown are both the individual participants (with each dotted line showing one participant) and well as the means over the group (black).

Note that the results section is split up in the three sections *'Behavioural analyses', 'MEG analyses',* and *'Three-way link between subjective ratings, performance, and neural states'*, with the first concerning Question 1-2, the second concerning Question 3-4, and the last concerning Question 5-6. Statistical correction for multiple comparisons was conducted separately for each of these three sections.

Question 1 and 2: Behavioural analyses

We examined the relationship between subjective ratings collected throughout the task (both on attentional state and performance) and performance. Specifically, we were interested in which objective behavioural task measure correlated best with the subjective ratings. To assess this, we selected five different measures of performance: 1) SD on the RTs of the five trials prior to the thought probe (SD) – reflecting consistency, 2) mean RT on the last five trials (mean RT) – reflecting a bias to either predict the tone or respond to it, 3) mean of the absolute RTs on the last five trials (mean |RT|) – showing the departure from the ideal performance of zero, and representing the measure closest to participants' instructions, 4) RT on the last trial before the probe (RT), and 5) absolute RT on the last trial before the probe (|RT|) – with the last two quantifying performance closest to the subjective report.

For Question 1, five τ-values were calculated separately for each participant between the attentional state ratings and each of the objective markers. An example for one participant is shown in Figure 4. Next, we tested whether each of these five distributions of τ-values was significantly different from zero at the group level, using both a classical null-hypothesis significance and a Bayesian one-sample t-test.

For Question 2, we were interested in the extent to which the subjective ratings correlated to each other. For this question, a Kendall's τ correlation coefficient was calculated separately for each participant between the attentional state and performance ratings (see Figure 4 for an example). Furthermore, τ-values

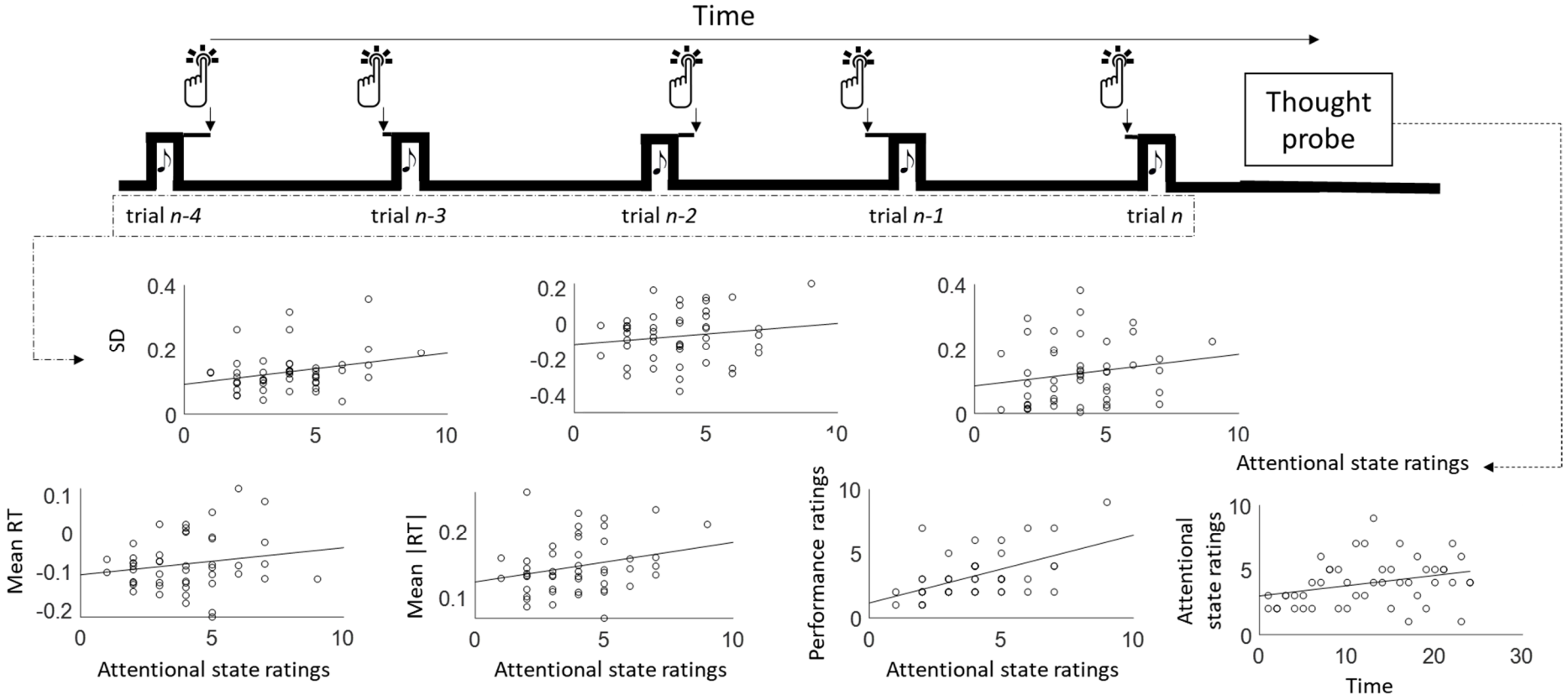


Figure 4. Examples of Kendall's τ correlational analyses for one participant, showing the relationship of attentional state ratings across trials with the five behavioural performance markers, as well as the relationship between the two metacognitive ratings, and the relationship between time and attentional state.

were calculated between the five objective markers with the subjective performance ratings.

P-values were corrected for multiple comparison with the False Discovery Rate (FDR; Benjamini & Hochberg, 1995) correction method, and tested for significance at $\alpha = .05$. Bayes Factors (BF_{10}) on the other hand are interpreted as continuous evidence – representing the ratio of the likelihood of the data under the alternative hypothesis (e.g., the distribution is different from zero) against the likelihood of the data under the null hypothesis (e.g., the distribution is not different from zero). For $BF_{10} > 1$, higher BF_{10} indicate more evidence for the alternative hypothesis, and for $BF_{10} < 1$, lower BF_{10} indicate more evidence for the null hypothesis. All Bayesian statistics throughout the current research were calculated in JASP (JASP Team, 2017) using equal prior probabilities for each model and 10000 Monte Carlo simulation iterations.

Note that to get a clearer insight into how participants perform the MRT, we also calculated the correlations between the different objective markers. Lastly, we aimed to quantify the linear trend over time within both subjective and objective measures. These results can be found in the Supplementary Materials (p. 145-146).

Question 1 – Metacognitive ratings correlate to behavioural variability

Out of the five objective behaviour measures (SD over five trials, RT on one trial, |RT| on one trial, mean RT over five trials, and mean |RT| over five trials), attentional state ratings correlated best to SD (see Table 2 for the descriptive and inferential statistics). The distribution of τ-values between subjective attention ratings and behavioural variability was statistically higher than zero – indicating that overall, participants were more variable when they reported to be more off-task (replicating Seli et al., 2013). There was a weaker correlation between the ratings with mean RT.

Question 2 – Different metacognitive ratings correlate to each other

The correlation between the two subjective ratings was by far the strongest association in the task-measures: Although the two ratings addressed separate metacognitive states, they shared a substantial amount of variance (~20% at group level; see Table 2 for overview).

Similarly for the subjective performance ratings, variability was the strongest behavioural correlate (see Table 2). Participants were more variable when they rated their performance as worse. Contrary to the attentional state ratings, performance ratings correlated to the absolute RT, both on the last trial and the last five trials – likely reflecting that precision was more important to participants than relative RT, which matches the instructions that they were given.

Although both attention and performance ratings correlated best with variability, the effects were weak – with the relationship between performance ratings and variability being stronger (explaining on average ~5.3% of the total variance) than between attention and variability (explaining on average ~3.2% of the total variance).

Table 2. Shown are the FDR-corrected p-value (p), Bayes Factor (BF_{10}), median and standard deviation of the distributions of Kendall's τ-values between subjective attentional state ratings (Attention), subjective performance (Performance) ratings, and five different objective behavioural performance measures (Behaviour), and time.

Attention to behaviour	**t**	**p**	**BF_{10}**	**Median**	**SD**
Attention – SD *(best measure)*	3.19	.012	9.21	.18	.14
Attention – RT	2.05	.093	6.51	.11	.17
Attention – |RT|	1.89	.113	.63	.05	.16
Attention – Mean RT	3.00	.020	1.32	.15	.15
Attention – Mean |RT|	1.51	.200	1.04	.08	.16

Subjective ratings	t	p	BF_{10}	Median	SD
Attention – Performance	7.21	<.001	17121.1	.44	.27
Performance to behaviour	**t**	**p**	**BF_{10}**	**Median**	**SD**
Performance – SD *(best measure)*	5.96	<.001	1846	.23	.16
Performance – RT	2.21	.093	.61	.08	.13
Performance – |RT|	3.84	.003	80.67	.11	.12
Performance – Mean RT	1.48	.201	1.46	.05	.14
Performance – Mean |RT|	4.32	.001	31.52	.11	.17

Question 3 and 4: MEG analyses

Next, we were interested in the extent to which the measures of performance and metacognition could be predicted from preceding oscillatory activity. To examine this, we extracted 'event-free' periods of two seconds for each trial, to investigate oscillatory power independent from activity related to the tones or button presses (see Figure 1). Because participants were inclined to respond before the tone, these event-free periods ranged from 2800-800 ms before the tone on trial *n*. These event-free periods were subsequently manually checked and excluded if they still included a response.

Using these event-free periods, correlational SAM analyses (SAM-R; Bompas et al., 2015) were conducted. Firstly, source activity was reconstructed separately for each session with synthetic aperturemagnetometry (SAM) analysis (Robinson & Vrba, 1999; van Veen, van Drongelen, Yuchtman & Suzuki, 1997; Vrba and Robinson, 2001), using a multiple local spheres forward model (Huang, Mosher & Leahy, 1999) – to estimate the global covariance matrices plus beamformer weight vectors for each voxel over a given frequency. The source construction was restricted to the brain volumes as identified by FSL's Brain Extraction Tool. The average source amplitude activity was calculated for each voxel and each trial separately. For the attentional state ratings, this analysis was done both on all the

ratings combined and separately for ratings split up between ‘mind wandering’-categorised off-taskness and ‘mind blanking-drowsy’-categorised off-taskness.

One non-trivial question is how many event-free periods prior to the probe should be included in the analysis. Prior literature does not give a clear guideline: windows from anything between .4 to 10 seconds have been used (see Table 1 for an overview). On the one hand, participants are asked to rate their attentional state *just prior* to the probe – meaning that smaller windows would give the best estimate of meta-cognition. On the other hand, the amount of thought probes is limited – meaning that too small windows will give a risk of being statistically underpowered. As a compromise, we correlated the ratings with the two prior event-free periods (see Figure 1). The analysis on behavioural variability comes with much higher statistical power (~1200 trials per participant), and therefore, it was correlated to only one preceding event-free period.

These analyses provided a Kendall’s τ correlation coefficient for each voxel per participants. The individual volumetric correlation images were subsequently averaged over all participants. Next, we tested for each voxel whether the correlation coefficients were significantly different from zero with permutation tests. This procedure was conducted for 2-7, 5-15, 15-30, and 30-100 Hz – providing us with one volumetric image for correlation coefficients and one for p-values separately for each frequency band. To correct for multiple comparisons across the four frequency bands, p-values were tested for significance at $\alpha = .0125$. A similar pipeline was conducted on the objective performance.

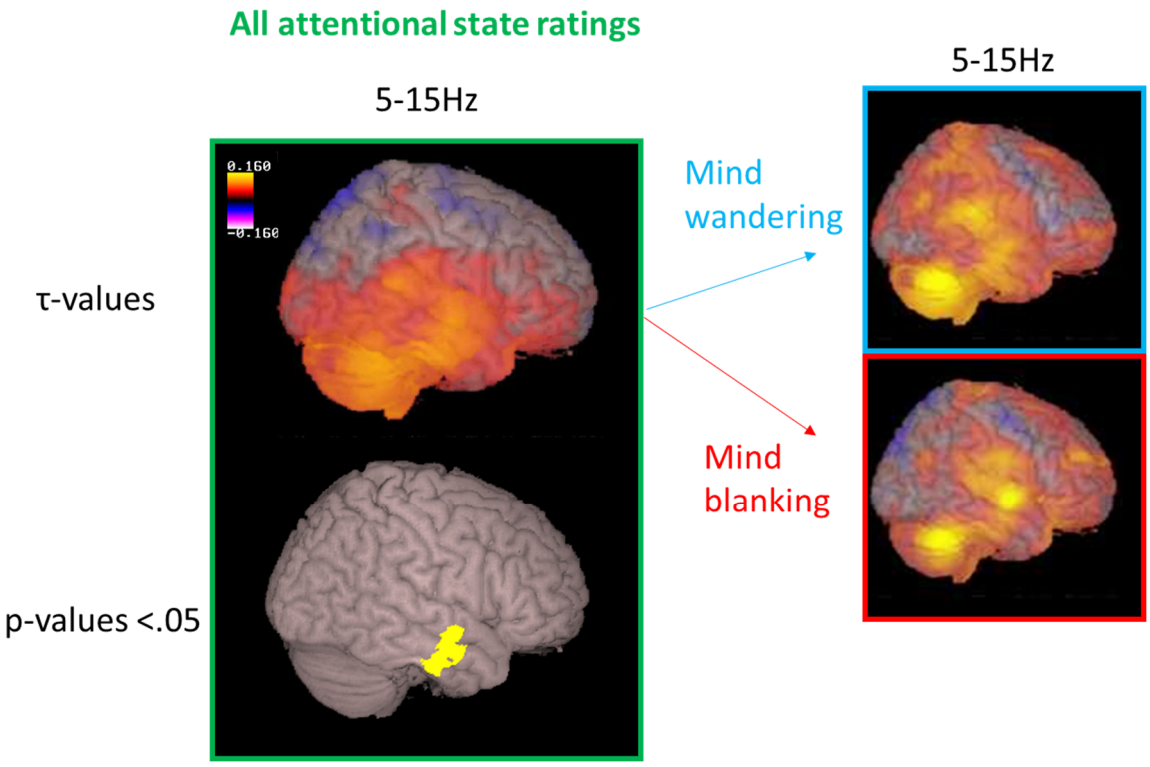


Figure 5. Group average of the volumetric correlational images between subjective attentional state ratings and oscillatory α – showing both the unthresholded correlation coefficients (top) and the corresponding significant p-values (bottom). Higher subjective off-taskness was correlated to increased right temporal α. These correlations were not found for the subjective performance ratings. Next, the attentional state ratings were split up into on-task+mind-wandering ratings only and on-task+mind-blanking ratings only. Exploratory analyses showed that the patterns were highly similar between the two different types of off-taskness.

Question 4 – Neural correlates of subjective ratings

As a first step, oscillatory power was correlated to all the attentional state ratings combined. Figure 5 shows the group averages of the correlation coefficients at 5-15 Hz, with the associated p-value map reflecting which correlations are significantly different from zero. We found that increased subjective off-taskness was preceded by increased right temporal α (reflecting higher baseline activity in the area associated with auditory processing; up to 2.6% explained variance). No significant correlations were found for the other three frequency bands. When using the

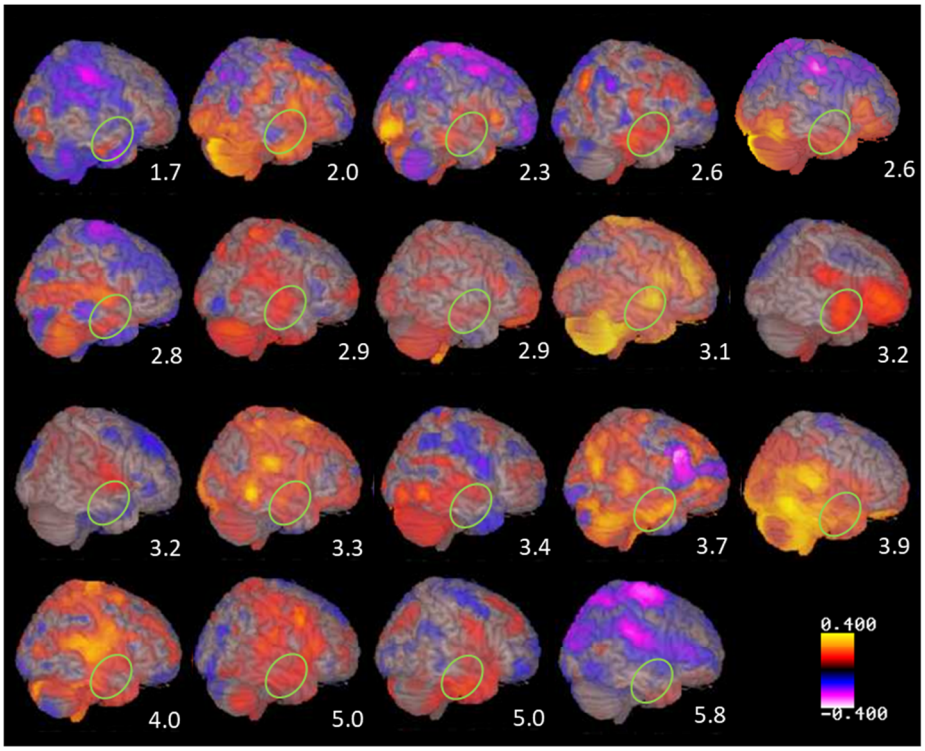
subjective performance ratings, we did not find any significant correlations in any of the four frequency bands.

Mind wandering versus mind blanking. Next, to examine different types of subjective off-taskness, we reran these analyses separately for 'mind wandering' – i.e., on all the on-task ratings (1-3), plus the off-taskness ratings (4-9) that were categorised by the participant as 'mind wandering' – and for 'mind blanking' – all the on-task ratings plus off-taskness ratings categorised as 'mind blanking – drowsy'. However, participants typically used one category more than the others (see Figure 2), causing an imbalance in trials between the categories. Therefore, participants were selected for this analysis only if they had at least two off-taskness ratings in *both* categories – leaving eleven participants for analysis. Due to the lower number of participants and the imbalance of trials, we did not calculate the p-value images – instead taking the results as strictly exploratory. As shown in Figure 5, the correlational images were highly similar across the two categories.

Individual differences. If one aims to predict people's upcoming attentional states (or upcoming performance), the group average is only useful in so far as it reflects the patterns of the individuals. Therefore, we were interested in the individual differences in activity. An FDR-corrected p-value map was calculated for each participant separately on the volumetric correlational image for attentional state. Figure 6 shows both the unthresholded correlation coefficients at 5-15 Hz (left) for each participant separately, as well as the coefficients that were significantly different from zero (right). These results reveal large inter-individual differences.

It is possible that the large individual differences arise due to differences in the way participants use the subjective ratings. We therefore aimed to investigate if differences in the subjective ratings itself were associated with differences in the correlational maps. Figure 6 shows the individuals' maps sorted by mean rating, from the participant with the lowest mean (1.7) on top-left to the highest mean (5.8) on the bottom-right. However, no clear patterns emerged – indicating that the inter-individual differences could not be explained by participants' mean rating. Similarly, participants' maps were sorted by: 1) SD of ratings, 2) range of ratings, 3) percentage mind wandering, and 4) percentage mind blanking. The individual differences could not be explained by any of these measures

All τ-values 5-15 Hz



τ-values 5-15 Hz with p < .05

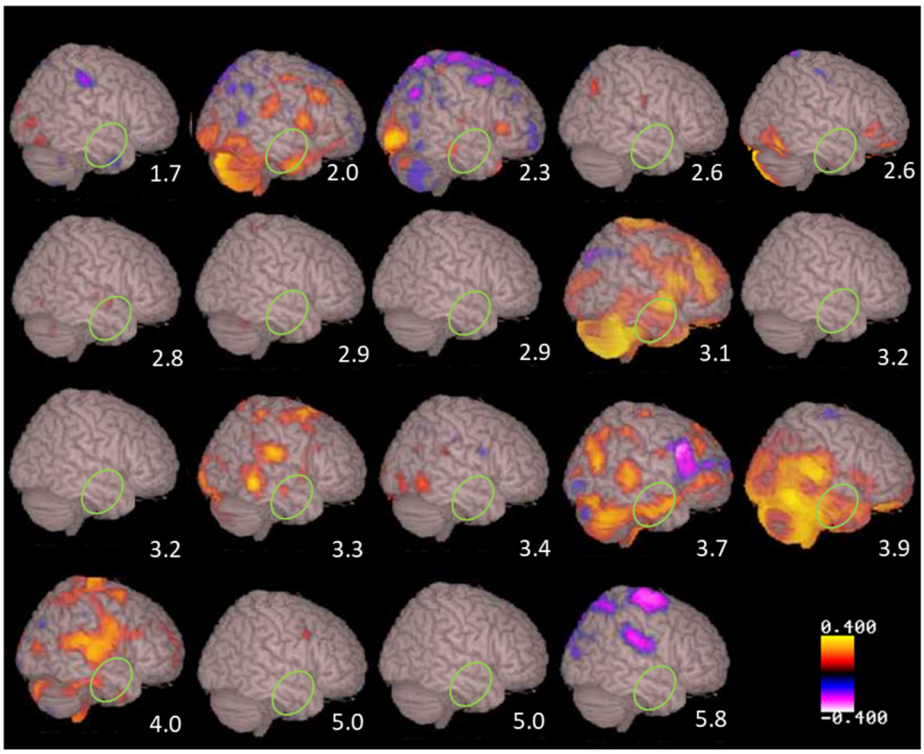


Figure 6. Volumetric correlation images between subjective attentional ratings and α power, separately for each participant – showing both the unthresholded τ-values (left) and the τ-values that are significantly different from zero (right) – with a green ellipse marking the approximate area that shows significant effects in the group analysis. Participants are sorted by their mean attentional state rating, with the participant who reported the least amount of off-taskness on the top-left (mean rating = 1.7), to the participant who reported the most amount of off-taskness on the bottom-right (mean rating = 5.8). No clear patterns were found in the individual differences.

Question 5 – Neural correlates of behavioural variability

Aside from the subjective ratings, we were interested whether the objective performance measures could be predicted from preceding neural states. On a behavioural level, the subjective ratings correlated best with RT variability. Therefore, we used this as the main performance measure. Please note that this measure is also the one that has been consistently used in previous literature using this task (Laflamme et al., 2018; Seli et al., 2013).

The left panel of Figure 7 shows the group averages of the correlation coefficients at 15-30 Hz, with the corresponding p-value maps below. Results showed that behavioural variability was positively correlated to occipital β (explained variance < 1%). No significant correlations were found for the other three frequency bands. Again, we found large inter-individual differences in the correlation maps. Similarly, the individual maps were sorted according to both mean and SD or the RT series – but no clear patterns emerged across participants (Figure 7, right panel).

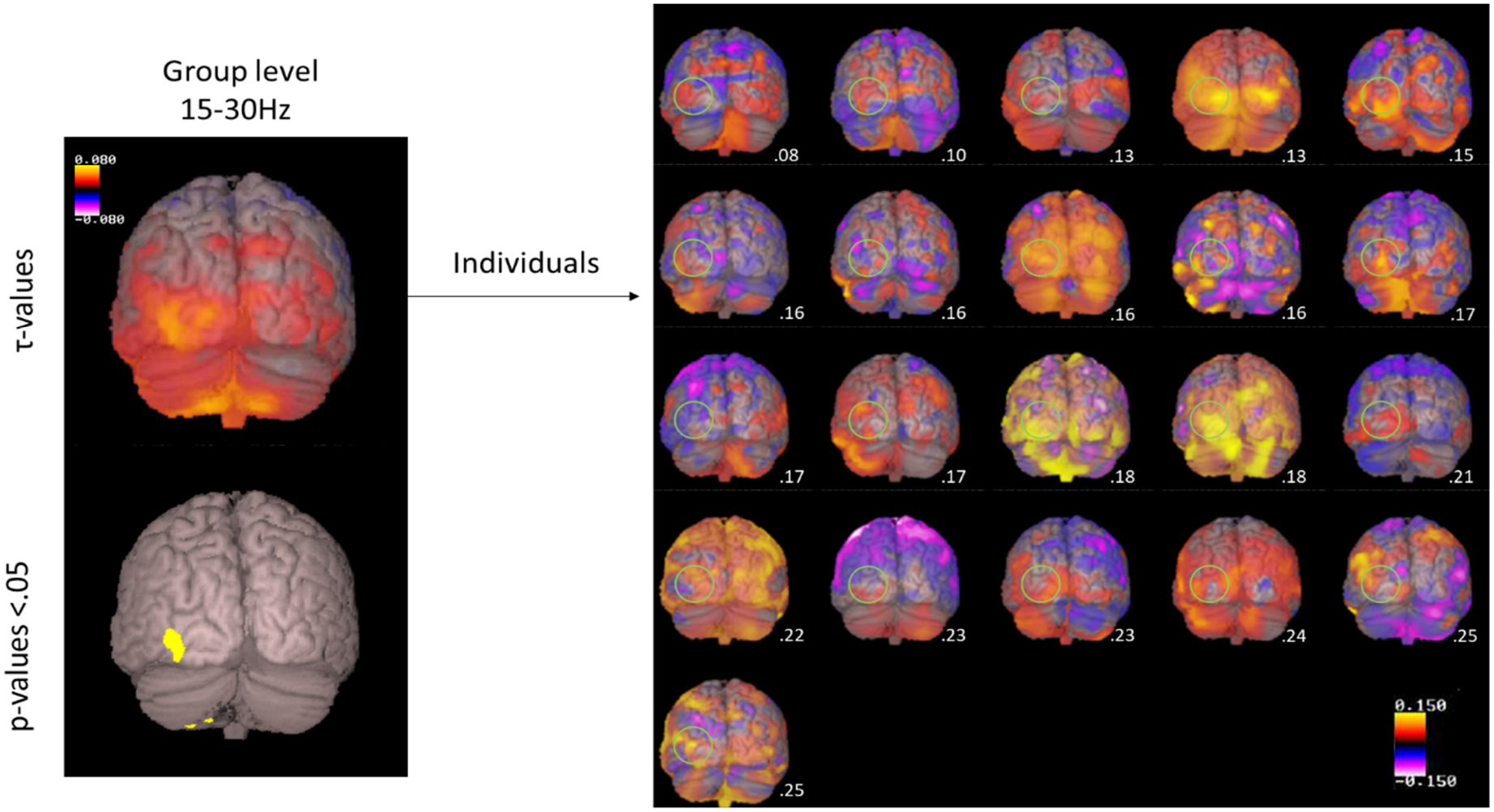

Group level
15-30Hz
τ-values
p-values <.05
0.080
-0.080
Individuals
.08
.10
.13
.13
.15
.16
.16
.16
.16
.17
.17
.17
.18
.18
.21
.22
.23
.23
.24
.25
.25
0.150
-0.150

Figure 7. Left. Group average of the volumetric correlational images between behavioural variability and oscillatory β (15-30 Hz) – reflecting the unthresholded correlation coefficients (top) and the corresponding significant p-values (bottom). Increased behavioural variability was correlated to increased left occipital β. No correlations were found for the remaining frequency bands. Right. Volumetric correlation images for the individuals – showing the unthresholded τ-values– with a green ellipse marking the approximate area that shows significant effects in the group analysis.. Participants are sorted by their overall RT variability, from the least variable participant on the top-left (SD = .08), to the most variable participant on the bottom-right (SD = .25). No clear patterns were found in the individual differences.

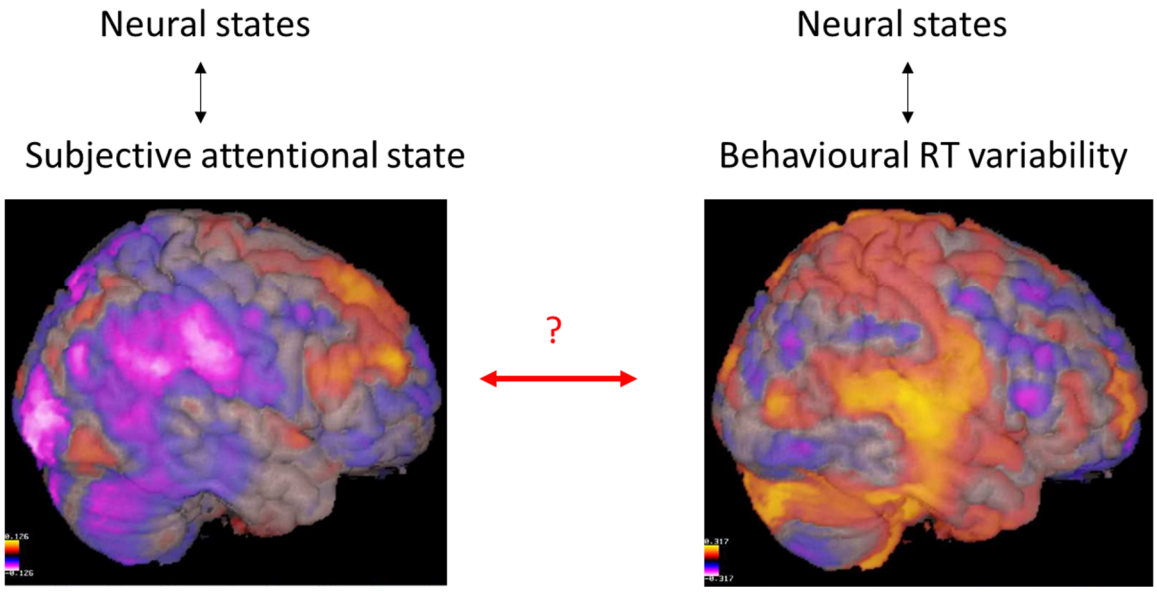


Figure 8. Example of the volumetric images of subjective attentional state and behavioural variability for one participant. To examine the extent to which the two measures have similar underlying neural states, a correlation coefficient was calculated between the two images across voxels.

Questions 5 and 6: Examining the three-way link between subjective ratings, performance, and neural states

So far, we examined a number of associations within task measures and neural states: 1) subjective attentional state and behavioural variability, 2) subjective attentional state and preceding neural states, and 3) behavioural variability and preceding neural states. However, it remains unknown to what extent the neural states underlying subjective attentional state and the neural states underlying behavioural variability overlap with one another (Question 5).

To examine this, we correlated the volumetric correlation image of attention ratings to the image of behavioural variability across voxels (see Figure 8 for an example) – resulting in one correlation coefficient per frequency band for each participant. One-sample t-tests were conducted to test if the distributions were statistically different from zero. Because this prediction specifically rests on positive associations between the measures, statistical tests were conducted one-sided, and tested at $\alpha = .025$. This analysis was also conducted for: 1) performance ratings and behavioural variability, and 2) subjective attention ratings and performance ratings. Figure 9 shows the distributions for each frequency band (statistically significant distributions in green), and associated descriptive and inferential statistics in Table 3.

Attentional state ratings and behavioural variability

We found statistical evidence for a correlation between underlying β power – indicating overlap in the neural activity of behavioural variability and of attentional ratings (4% shared variance). However, no correlations were found for the other frequency bands (with Bayesian evidence remaining indeterminate).

It is possible that the lack of an association on the other frequency bands is due to individual differences. For Question 1, we found that although there was a positive correlation between attentional state ratings and behavioural variability at the *group level*, there were large individual differences – with some individuals even showing a (weak) negative correlation between their ratings and variability. One could argue that the largest amount of overlap between neural states should be

found in those individuals that have the largest relationship between ratings and variability (Question 6). Therefore, between-subject correlation analyses were conducted between the distributions of the behavioural overlap (attentional state ratings ↔ behavioural variability) and the neural states overlap (neural states of attentional state ratings ↔ neural states of behavioural variability). Again, statistical tests were conducted one-sided.

Table 4 shows the correlation coefficient of these between-subject correlations, separately for each analysis, including inferential statistics. Figure 9 shows the associated correlational plots, with each dot representing one participant. If the largest overlap in neural states would be found individuals with a strong behavioural association, this would be represented in a positive correlation specifically in the blue quadrant (reflecting the participants whose behavioural and neural correlation coefficients were both numerically higher than zero) – and possibly extending towards the yellow quadrant (reflecting the participants whose correlations were both numerically below zero). However, we did not find this positive correlation, and Bayesian statistics showed clear evidence *against* an association between the distributions – meaning that participants who showed higher correlations between their subjective ratings and behavioural variability did not show higher overlap in the underlying neural states.

Performance ratings and behavioural variability

The same analyses were conducted between the neural states underlying subjective performance ratings and underlying behavioural variability. Correlations were significant at the group level on all four frequency bands (6.3-14.4% shared variance). Looking at the individual differences, we again find clear evidence against a correlation between the behavioural and neural correlations.

Attentional state and performance rating

Lastly, these analyses were conducted between the neural states underlying both subjective ratings. These distributions were statistically different from zero and

overall positive – showing that the neural states underlying attentional state ratings and the neural states underlying performance ratings do overlap with each other (amount of overlap ranging from 8-21% across frequency bands). Again, however, there was clear statistical evidence against a correlation with behavioural overlap.

Table 3. Within-subject correlations between the volumetric correlation images of subjective and objective markers (Question 5). Shown are the FDR-corrected p-value (p), Bayes Factor (BF_{10}), median and standard deviation of the distributions of Kendall's τ-values, separated for the correlations between neural states of attentional state ratings and of behavioural variability ($\tau_{attention\text{-}power,\ \tau_{SD\text{-}power}}$), neural states of performance ratings and of behavioural variability ($\tau_{performance\text{-}power,\ \tau_{SD\text{-}power}}$), and neural states of attentional ratings and of performance ratings ($\tau_{attention\text{-}power,\ \tau_{performance\text{-}power}}$). Note that the statistical tests were conducted one-sided.

$\tau_{attention\text{-}power,\ \tau_{SD\text{-}power}}$	p	BF_{10}	Median	SD
2 to 7 Hz	.137	.71	.09	.27
5 to 15 Hz	.078	1.10	.10	.28
15 to 30 Hz	.017	3.84	.20	.28
30 to 100 Hz	.041	1.85	.12	.28
$\tau_{performance\text{-}power,\ \tau_{SD\text{-}power}}$	p	BF_{10}	Median	SD
2 to 7 Hz	.009	6.41	.15	.25
5 to 15 Hz	.014	4.37	.22	.29
15 to 30 Hz	.015	4.20	.18	.32
30 to 100 Hz	.018	3.64	.16	.38
$\tau_{attention\text{-}power,\ \tau_{performance\text{-}power}}$	p	BF_{10}	Median	SD
2 to 7 Hz	<.001	188	.40	.33
5 to 15 Hz	<.001	259	.29	.33
15 to 30 Hz	<.001	128	.43	.36
30 to 100 Hz	<.001	58	.46	.39

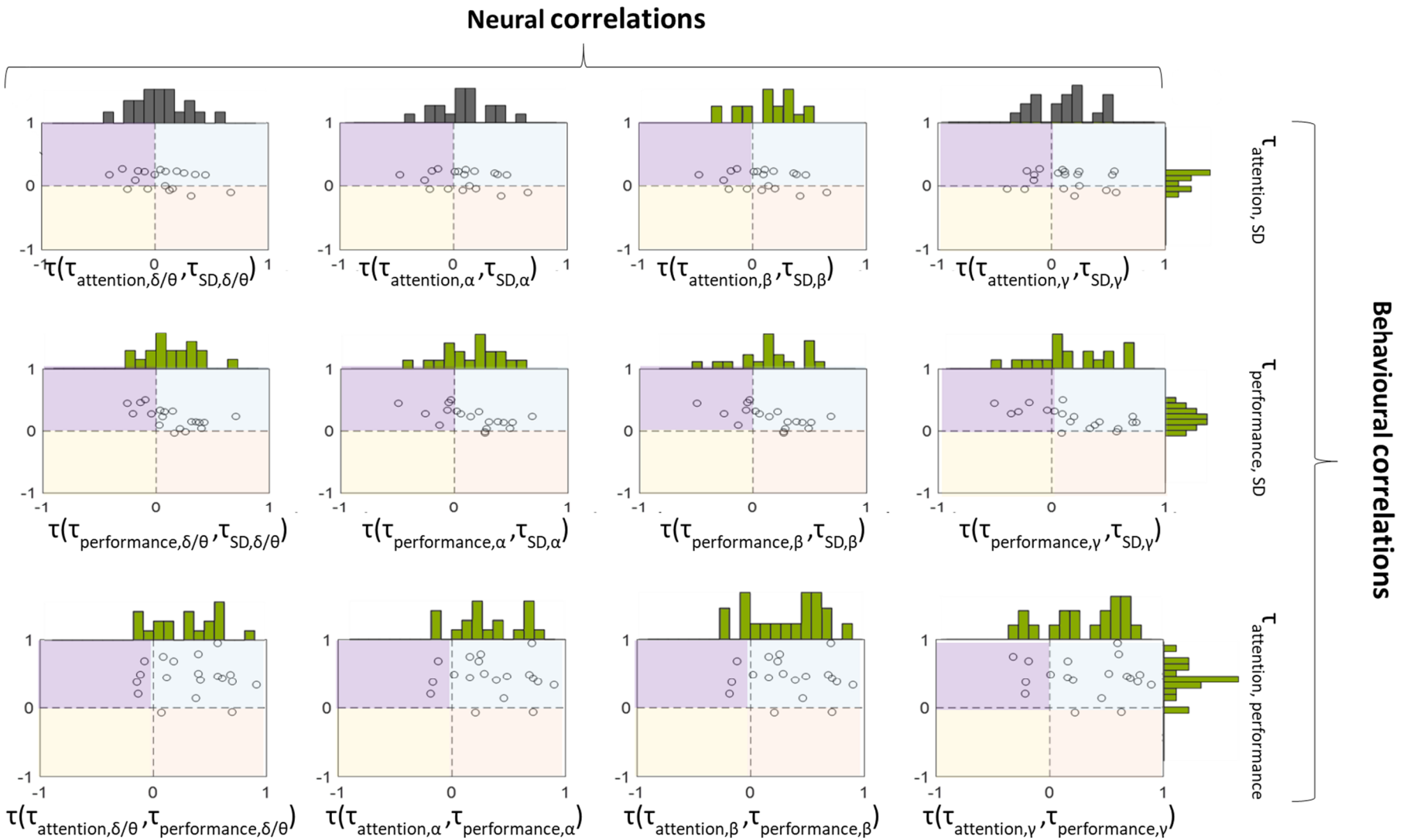

Neural correlations
Behavioural correlations
τ(τattention,δ/θ, τSD,δ/θ)
τ(τattention,α, τSD,α)
τ(τattention,β, τSD,β)
τ(τattention,γ, τSD,γ)
τ(τperformance,δ/θ, τSD,δ/θ)
τ(τperformance,α, τSD,α)
τ(τperformance,β, τSD,β)
τ(τperformance,γ, τSD,γ)
τ(τattention,δ/θ, τperformance,δ/θ)
τ(τattention,α, τperformance,α)
τ(τattention,β, τperformance,β)
τ(τattention,γ, τperformance,γ)
τattention, SD
τperformance, SD
τattention, performance

Figure 9. The twelve histograms on top of the correlation plots show the Kendall's τ-values between the MEG volumetric images of attentional state ratings and of behavioural variability (top row), between the images of performance ratings and of behavioural variability (middle row), and between images of attentional state ratings and of performance ratings (bottom row), separated for the four frequency bands – to estimate to which extent the task-measures are associated with the same underlying neural processes (Question 5). Distributions that are significantly different from zero are shown in green. There is a clear overlap in neural states between the two metacognitive ratings, as well as overlap between the performance ratings and behavioural variability, but evidence for an overlap between attentional state ratings and behavioural variability was mixed. To examine individual differences in this overlap, these correlations were associated to the behavioural correlations between ratings and variability (Question 1) – distributions shown on the vertical axis – to examine whether participants who have a stronger link between subjective states and behaviour also have more overlap between the underlying neural states (Question 6). These associations are shown on the correlational plots, with each dot representing one participant. We found clear statistical evidence against a positive association.

Table 4. Results of the between-subject correlations between the neural and the behavioural overlap (Question 6), showing the FDR-corrected p-value (p), Bayes Factor (BF_{01} – indicating evidence against a correlation), and correlation coefficient τ between the correlations in neural states and the correlations in behaviour. Note that the statistical tests were conducted one-sided.

$T_{neural\ overlap}$, $T_{behavioural\ overlap}$			
Attention, SD	**p**	**BF_{01}**	**τ**
2 to 7 Hz	> .999	7.43	-.22
5 to 15 Hz	> .999	6.72	-.18
15 to 30 Hz	> .999	6.25	-.16
30 to 100 Hz	> .999	4.71	-.08
Performance, SD	**p**	**BF_{01}**	**τ**
2 to 7 Hz	> .999	11.49	-.40
5 to 15 Hz	> .999	10.97	-.38
15 to 30 Hz	> .999	13.89	-.51
30 to 100 Hz	> .999	12.02	-.43
Attention, Performance	**p**	**BF_{01}**	**τ**
2 to 7 Hz	> .999	4.71	-.08
5 to 15 Hz	> .999	5.13	-.10
15 to 30 Hz	> .999	4.92	-.09
30 to 100 Hz	> .999	5.35	-.11

Discussion

In the current study, we were interested in the relationships between objective performance measures, metacognitive reports, and preceding neural states. Specifically, we examined six research questions: 1) the extent to which subjective ratings of attentional state and performance may be related to each other, 2) the associations between subjective ratings and objective task performance, 3) oscillatory power preceding subjective ratings, differentiating between different meta-cognitive experiences (attentional state and performance) and between different types of off-taskness (mind wandering and mind blanking), focusing both at group and individual level, 4) oscillatory power preceding behavioural variability, similarly focusing on both group and individual level, 5) the amount of overlap between those neural states underlying subjective ratings and those underlying behavioural variability, and 6) the extent to which this overlap would be higher in individuals with higher associations between subjective ratings and behavioural variability.

Our results replicate previous findings that that subjective off-taskness is associated with increased behavioural variability compared to on-taskness, with the effect size being in the expected range (Question 1; Laflamme et al., 2018; Seli et al., 2013). We found a similar positive relationship between subjective reports of perceived performance and behavioural variability, and while the effect size was slightly stronger, it remained in the weak range. Furthermore, we replicated that different subjective ratings correlate to each other (Question 2; MacDonald et al., 2011). Looking at the neural states preceding the task measures, we found significant effects at the group level – specifically, that increased off-taskness was positively associated with α power in the right temporal gyri (Question 3), while behavioural variability was positively associated with left occipital β (Question 4). For both however, the differences between participants were high, and the group level was not representative of individual patterns. We found clear statistical evidence for an overlap in neural states between performance ratings and behavioural variability as well as for an overlap between the two subjective ratings across all frequency bands. For attentional state ratings and behavioural variability,

overlap was only found for β power (Question 5). The neural overlap was not higher for individuals who had higher correlations between the subjective and objective measures on the behavioural level (Question 6).

Predicting subjective and objective markers from oscillatory power

Previous studies have found that subjective ratings of off-taskness were preceded by increased occipital and parietal α (Baldwin et al., 2017; MacDonald et al., 2011; Jin et al., 2019). Such increases in occipito-parietal α have also been found when people are performing 'internally-focused' mental imagination) compared to stimulus-based tasks (Cooper, Burgess, Croft & Gruzelier, 2006; Cooper, Croft, Dominey, Burgess & Gruzelier, 2003), and when people are anticipating an auditory stimulus while they are being presented with visual distractors (Foxe et al., 1998; Fu et al., 2001). On the other hand, redirection of focus (e.g., after errors) on the visually-external environment has related to a decrease in occipito-parietal α (Mazaheri, Nieuwenhuis, van Dijk & Jensen, 2009). On a behavioural level, decreases in α have been associated with better detection performance (van Dijk et al., 2008; Ergenoglu et al., 2004; Hanslmayr et al., 2007). As such, α has been linked to inhibition of the areas relevant for the task, and has been described as a reflection of current 'cortical excitability'. The increased α prior to self-reported off-taskness has been interpreted as 'mental withdrawal' or 'perceptual decoupling' – which is subsequently also reflected in reduced sensory ERPs (Baird et al., 2014; Kam et al., 2011; Jin et al., 2019).

Our current results are consistent with this interpretation: As we found increased temporal α prior to subjective off-taskness ratings during a task with auditory input, this instead may indicate inhibition in the areas relevant for auditory processing (with which the temporal gyrus has been associated; see Moerel, De Martino & Farmisano, 2014 for a review). However, it is important to emphasise that the current study is the first to specifically examine individual differences. We found that the group average was not a good representation of what was going on in the individuals – which makes it difficult to interpret group average activity. Future studies may include individual's activity to estimate the robustness of the effects and interpretations.

While the above-mentioned studies on detection performance use stimuli that are difficult to detect (e.g., low-contrast targets or targets within white noise), behavioural variability shows up even when stimuli are clearly detectable and entirely stable over time. Neural mechanisms of behavioural variability using high-contrast stimuli have been studied previously as well, these studies have been highly variable in used behavioural tasks, analyses methods, and findings, making it difficult to integrate findings across the literature (Bompas et al., 2015; Drewes & VanRullen, 2011; Everling, Krappmann, Spantekow & Flohr, 1997; Gonzalez Andino, Michel, Thut, Landis & Grave de Peralta, 2005; Hamm, Dyckman, Ethridge, McDowell & Clementz, 2010; Hamm, Sabatinelli & Clementz, 2012; Perfetti et al., 2011; Zhang, Wang, Bressler, Chen & Ding, 2008).

It should be mentioned that, with exception of Bompas et al. (2015), these studies have not differentiated between endogenous (i.e., arising from internal fluctuations within the participant) and exogenous (i.e., arising from external sources such as order in trials and experimental conditions) variability.

In the current study, we have used the MRT (Seli et al., 2013), which is particularly suited for measuring endogenous fluctuations in behaviour, as the task features no different conditions or different stimuli, and remains the same throughout the entire session. However, the task-measures may still be influenced by trial order – meaning that found associations between variables may be mediated by time and/or that true associations may be masked. Indeed, we found that particularly subjective attentional state ratings increase over time (see Supplementary Materials, p. 145-146). To test this, Kendall's τ partial correlations were calculated for the relationship between attentional state ratings and variability, performance ratings and variability, and attentional state and performance ratings, while controlling for time (as measured by block number) with:

$$\tau_{XY.Z} = \frac{\tau_{XY} - \tau_{XZ}\tau_{YZ}}{\sqrt{(1-\tau_{XZ}^2)(1-\tau_{YZ}^2)}}$$

These distributions were highly similar compared to the original distributions, only being about .01-.02 lower on the group level. As such, they found associations between subjective and objective measures are not driven by time. Still, this is only

true for the behavioural data – future steps may include determining the effect of time on the neural activity and its associations to the behavioural data.

The intuitive link between behaviour and subjective attentional state

Our different research questions emerged from one overlapping aim: To investigate the extent to which behavioural variability and subjective attentional state truly reflect the same processes. This is important because they have been taken as highly similar, likely due to their intuitive link. This is apparent even in the scientific literature – for instance, by assuming that subjective ratings and behavioural variability are markers of the same underlying mental process (Qin et al., 2013) – and affects the way questions within this field are tackled.

Imagine, for instance, workers at a control centre at an airport. In this context, we want the controllers to be as good and consistent at possible at their jobs, as delays and errors can have fatal consequences. Such situations result in an interest to reduce behavioural variability and off-taskness as much as possible (though both are known to have beneficial qualities in other contexts as well). This logic goes as following: 1) we experience spontaneous fluctuations both in our attentional state and in our behaviour, and specifically 2) we often experience that we are variable because of our fluctuations in attentional state – e.g., the air controller may have made an error because they were not paying attention, which means that 3) we must find the neural correlates of subjective off-taskness, because 4) we could then detect these in real life situations – e.g., by online EEG recordings of the air controllers when they are at work, so 5) we can interfere when the recording detects off-taskness – e.g., by issuing a warning signal to the air controller, which 6) results in a reduction of poor performance (and thus to less accidents). This line of thinking is highly dependent on a strong link between subjective off-taskness and behavioural variability. However, the current findings do not support a strong association.

Effect sizes

First, quantifications of effect are vital in the statistical analyses typically performed in cognitive neurosciences: While p-values can inform whether there is at all a significant effect, they cannot tell us the magnitude of said effect – just as a pregnancy test can tell whether one is pregnant, but does not give any estimate about how many weeks. As the positive correlation between behavioural variability and subjective ratings has consistently been found, one may rush to conclude that these two are strongly linked – as per our intuition.

However, our current effect sizes indicated that the relationship between subjective attentional state ratings and behavioural variability was very weak. Subjective performance ratings were correlated better to behavioural variability, (with follow-up two-sided paired t-test analysis confirming the correlations between performance ratings and variability were indeed statistically higher than correlations between attentional state ratings and variability, $t(18) = 2.62$, $p = .017$, Cohen's $d = .60$, BF10 = 3.32, but effect size were in a similar low range. As such, measures of metacognition seem largely uninformative for actual behaviour.

Unfortunately, the vast majority of articles on mind wandering do not report effect sizes, with MacDonald et al. (2011) as a noteworthy exception. The general absence of effect sizes hinders the interpretation of findings, as low effect sizes make it less likely that our inner states are directly (if at all) accessible to us.

Lack of neural overlap

Secondly, our MEG results suggest that behavioural variability and attentional state rely on different neural mechanisms – as the neural states underlying variability and the neural states underlying subjective ratings do not significantly overlap with each other. It should be noted that we did not find Bayesian evidence against an overlap, as the Bayes Factors remained in the indeterminate range. This should not reflect a flaw in our analysis approach, as the found overlap in the neural states underlying respectively subjective attentional state ratings and performance ratings shows its viability. Instead, the lack of clear Bayesian evidence is likely an issue of statistical power: when calculating a correlation coefficient, a lot of information (variance) gets

removed. Correlating two correlation coefficients may therefore be inherently noisier.

However, these findings imply that even if there is some overlap in the respective neural states, this overlap would be very small (current results showing 0.8- 4.0% shared variance). These results still go against the intuitive framework: If the relationship between attentional state and variability is indeed as strong as intuitively thought, then we would expect large and clear effects even in small samples. Furthermore, we did find clear Bayesian evidence against a positive correlation between behavioural and neural overlap: Participants whose subjective attentional states correlated better to their behavioural variability did not show more overlap in their respective underlying neural states. Importantly, if one wants to successfully predict one marker from the neural states underlying the other marker, these strong effects are needed. As such, our results imply this approach would not be fruitful.

Our results fit well with prior research, even if they have not been explicitly addressed as such. As discussed in the introduction, previous studies investigating oscillatory power during subjective off-taskness have either: 1) found a relationship of neural states to subjective ratings but not to performance (MacDonald et al., 2011), 2) found direct significant differences between activity underlying subjective ratings and underlying performance (Qin et al., 2011), or 3) have not investigated the neural processes of performance at all.

Similar evidence comes from fMRI research: Kucyi, Esterman, Riley, and Valera (2016) recorded fMRI data while 29 participants were performing a sustained attention task with pseudo-randomly presented thought probes. They found that Default Mode Network activity increased when participants reported to be off-task (which has been commonly found in fMRI research on mind wandering; see Fox, Spreng, Ellamil, Andrews-Hanna, & Christoff for a meta-analysis; see Andrews-Hanna, Smallwood & Spreng, 2014; Christoff, 2012; Gruberger, Simon, Levkovitz, Zangen & Hendler, 2011; Mittner et al., 2014; Smallwood, Brown, Baird, & Schooler, 2012 for reviews), but decreased when participants were highly variable on the task. These results likewise show that despite the intuitive link, the neural mechanisms

underlying the link between off-taskness and behavioural variability is not all straightforward.

Subjectivity

In general, veracity and informativeness of subjective reports remains a topic that is difficult to tackle. Even when we find seemingly meaningful associations (for instance with behaviour), this still cannot tell us what type of information participants are using to decide on their report, nor whether the found relationship is direct. While we found a clear association between the two metacognitive reports (similar to MacDonald et al., 2011), their sufficient amount of unshared variance suggests that participants do use (at least partially) different information to rate their different metacognitive experiences. There is some evidence that the two ratings show different time profiles, $t(18) = 2.45$, $p = .025$, Cohen's $d = .56$, $BF_{10} = 2.46$, and on average, participants reported a much larger range in their attentional state ratings than in their performance ratings. Our results did indicate that there was an overlap in underlying neural states, but unexpectedly, this overlap did not positively correlate to the overlap between the ratings – that is, participants whose subjective ratings of attentional state and performance were more similar to each other did not have more overlap in their respective underlying neural states.

Mind wandering versus mind blanking

The current study is the first that attempts to compare oscillatory processes underlying mind wandering versus mind blanking – prior studies have either studied them as one general process of 'off-taskness', or have excluded mind blanking from analysis (Jin et al., 2019). We did not find any differences in activity between the two, but these analyses remained without inferential statistics due to a lack of power. Furthermore, as we compared 'all on-task ratings + mind wandering' to 'all on-task ratings + mind blanking', there was a lot overlap in activity by default – as at the group level, on-taskness was the most common mental state. One potential solution for future research may be to record participants for longer sessions – potentially leading in an increase in off-taskness, and specifically mind blanking, states. Still,

this type of trial-imbalances remains a major practical limitation in mind wandering research in general; rather than being able to manipulate conditions for statistical comparison, we can only record participants and hope that they experience enough of each meta-cognitive state.

Intentionality of off-taskness

One issue that has been left unaddressed throughout this thesis is intentionality. In the current task, we asked participants to classify the intentionality of their on- or off-task thoughts (the framing of the question being dependent on their attentional state ratings). Some previous studies have used such classifications to differentiate between intentional and unintentional off-taskness – for example, by examining differences in structural and functional connectivity between the two (Golchert et al., 2017). However, looking at the current data, we found the reporting had an extreme bias: When participants reported to be on on-task, they classified this as intentional, and when they reported to be off-task, they classified this as unintentional. Intentional off-taskness was extremely rare (or even absent in most participants), and as such, we did not analyse these responses further.

One reason for not finding intentional off-taskness may be that we gave clear instructions to participants that their job throughout the experiment was to perform well on the task. Even more so, we aimed to motivate them with a reward system that encourages consistently good performance over all the trials. This may be important because mind wandering is typically described as a drift of focus from the *primary* task to task-unrelated thoughts – meaning it should be clear to participants that the experimental task is their main task to perform. Robison, Miller, and Unsworth (2019) recently reported a similar lack of intentional off-taskness. However, they found that the manipulation of instructions (instructions either to avoid mind wandering throughout the task versus, being told that mind wandering was fine, or attaching no specific value) did not influence either task-performance nor subjective ratings. Overall, the large discrepancy in subjectively reported intentionality between different research groups is remarkable, and different instructions and reward systems (or other motivating factors) may be investigated in more detail.

Conclusion

In the current study, we were interested in the extent to which the subjective experiences of attentional state and behavioural variability truly resemble or relate to the same processes – as is commonly assumed. We found that subjective ratings and behavioural variability were correlated to each other, and that there was overlap in their underlying neural states. However, this overlap was only found in the β band – indicating that the neural states underlying subjective ratings (in this case: α) is not necessarily the most informative for the three-way link between subjective ratings, neural states, and behaviour.

Aside from rating attentional state, participants were also instructed to rate their subjective estimate of performance on a separate question. We found these performance ratings were superior in multiple ways: 1) their correlation to behavioural variability was higher, 2) they showed neural overlap with behavioural variability over all the frequency bands, and 3) unlike attentional state ratings, they also correlated to absolute RT, which participants are instructed to focus on.

For both type of ratings however, effect sizes were weak. This suggests that the subjective and objective measures are poor markers of each other.

Supplementary Materials

In the current study, we used the Metronome Task (Seli et al., 2013), which is particularly suited for measuring endogenous fluctuations in behaviour, as the task features no different conditions or different stimuli, and remains the same throughout the entire session. However, it also comes with its downsides. First, participants may differ in the strategy they use. While this is likely true for most neurocognitive tasks, other, more established tasks may come with more straightforward ways to quantify these (e.g., modelling the speed-accuracy trade-offs in rapid action selection tasks). Second, the MRT's main measure does not quantify what participants are instructed to focus on – as we touched upon in the results of Question 1: while it is common to analyse the consistency of performance over multiple trials, participants are told to focus on precision on each trial separately.

To get a clearer picture of the different objective markers (SD on last five trials, RT on last trial, |RT| on last trial, mean RT on last five trials, and mean |RT| on the last five trials), these were correlated to each other within participants. The Supplementary Table shows the descriptive and inferential statistics of the distributions of τ-values. Note that the statistical correction for multiple comparisons was conducted together with the analyses of Question 1 and 2. Overall, the strongest correlation between the measures was found between the SD and the mean absolute RT on the last five trials. This indicates that although consistency is not the main goal, it still reflects the precision that participants are told to focus on.

We also tested how the objective and subjective measures evolved over the blocks. The Kendall's τ correlation coefficients were calculated for each participant between block number (from 1 to 24) and each task measure: 1) attentional state ratings (example shown in Figure 4), 2) performance ratings, 3) variability, 4) RT, 5) |RT|, 6) mean RT, and 7) mean |RT|. There was clear evidence for a correlation between time and subjective attentional state ratings (indicating that participants reported to be more off-task as time passed by) and evidence against a correlation between time and each of the four RT measures (indicating both relative and absolute RT remained stable over time). For the other measures, BF10 remained indeterminate, with non-significant p-values. Overall, time was most associated with

the attentional state ratings, although the effect was again weak (2.9% explained variance). To the extent that time may also be associated with behavioural variability and/or performance ratings, these effects appear to be even weaker.

Supplementary Table. Shown are the FDR-corrected p-value (p), Bayes Factor (BF_{10}), median and standard deviation of the distributions of Kendall's τ-values between the different objective behavioural performance measures (Behaviour), as well as the relationship of the subjective and objective measures with time.

Behaviour	**t**	**P**	**BF_{10}**	**Median**	**SD**
SD – RT	-2.38	.060	2.21	-.03	.10
SD – |RT|	5.08	<.001	347	.16	.14
SD – Mean RT	-2.07	.095	1.35	-.10	.23
SD – Mean |RT| *(best measure)*	7.49	<.001	27929	.45	.24
RT – Mean RT	16.76	<.001	3.5e+9	.33	.09
RT – Mean |RT|	-.70	.552	.30	-.10	.26
|RT| – Mean RT	-1.25	.281	.47	-.09	.23
|RT| – Mean |RT|	15.02	<.001	6.2e+8	.32	.11
Mean RT – Mean |RT|	-1.91	.116	1.06	-.19	.43
Effect of time on measures	**t**	**p**	**BF_{10}**	**Median**	**SD**
Time – Attention *(strongest effect)*	5.07	<.001	344	.17	.14
Time – Performance	1.74	.141	.84	.09	.18
Time – SD	2.44	.056	2.43	.02	.14
Time – RT	.23	.856	.24	-.04	.14
Time – |RT|	.99	.340	.36	.07	.12
Time – Mean RT	.22	.829	.24	.05	.14
Time – Mean |RT|	.37	.775	.25	.01	.12

Chapter 4

Inability to improve performance with control shows limited access to inner states

Abstract

Any repeatedly performed action is characterised by endogenous variability, affecting both speed and accuracy – for a large part presumably caused by fluctuations in underlying brain and body states. The current research questions were: 1) whether such states are accessible to us, and 2) whether we can act upon this information to reduce variability. For example, when playing a game of darts, there is an implicit assumption that people can wait to throw until they are in the 'right' perceptual-attentional state. If this is true, taking away the ability to self-pace the game should worsen performance. We first tested precisely this assumption asking participants to play darts in a self-paced and a fixed-paced condition. There was no benefit of self-pacing, showing that participants were unable to use such control to improve their performance and reduce their variability. Next, we replicated these findings in two computer-based tasks, in which participants performed a rapid action-selection and a visual detection task in one self-paced and three forced-paced conditions. Over four different empirical tests, we show that the self-paced condition did not lead to improved performance or reduced variability, nor to reduced temporal dependencies in the reaction time series. Overall, it seems that, if people

have any access to their fluctuating performance-relevant inner states, this access is limited and not relevant for upcoming performance.
Key words: Intra-individual variability; metacognition; attention; noise

Introduction

Variability is a prominent characteristic of all human behaviour. Any repeatedly performed action will show substantial variation both in how well the action is performed and in how much time is needed to perform it. This is not only true for behaviour in daily life, but can also be measured precisely during cognitive experiments. For instance, even on simple reaction time tasks featuring the same high contrast stimulus on every trial, response times (RT) show large fluctuations relative to their mean. Although variability can also have beneficial aspects (see our Discussion), it is often perceived as desirable to reduce variability as much as possible. In the lab, we seek to reduce measurement error and obtain cleaner data. In real life, we may strive to reduce variability anywhere from trivial situations, such as keeping up consistently good performance when playing darts or playing music in a band, to contexts where variability may lead to more serious consequences, such as traffic and air control.

In many situations in our everyday lives, we take it for granted that we can maximise performance by acting when we feel ready for it. For instance, in darts – and other shooting or throwing sports – players typically take a moment to concentrate and choose the ‘right’ moment to initiate an action. However, this intuition relies on two non-trivial assumptions. Let us accept that, if performance varies across time under unchanging circumstances, this has to be due to variations in some internal states. The intuition above then assumes that: 1) we can access aspects of these fluctuating internal states which are directly relevant to performance, 2) we can choose when to act accordingly in order to improve performance. The current article tests these assumptions. Specifically, we address

the effects of control upon variability and performance: Participants are given a means to only start each trial when they feel ready to continue. If it is possible to have access to these performance-related internal states and to act upon this information in a useful way, the control should be an effective measure for reducing response variability and errors.

Endogenous variability and its accessibility

In many lab-based tasks, behavioural variability can be attributed to factors inherent to the task (experimental conditions and their time order) or directly linked to the feedback (such as learning or post-error slowing, the latter referring to the phenomenon that an error on trial *n* is usually followed by a slow response on trial *n+1*; Rabbitt, 1966). However, in simple tasks, such as pressing a button in response to a visual onset, all these factors explain only a small proportion of the overall variability (Bompas et al., 2015; Gilden 2001). The residual variance, referred to as endogenous or spontaneous, has recently received growing interest, as its properties and causes still remain largely unknown.

First, not all of this endogenous variability is random noise. Indeed, in the lab, RT on a trial is partly correlated to that on subsequent trials, and such temporal dependencies unfold on short- but also on longer-term scales (Gilden, 2001; Kelly et al., 2001; Wagenmakers et al., 2004). Similar temporal dependencies have also been found in sports performance (van Beers, van der Meer & Veerman, 2013; Gilden & Wilson, 1995; Huber, Kuznetsov & Sternad, 2016; Smith, 2003; Stins, Yaari, Wijmer, Burger & Beek, 2018). It is tempting to attribute some of this endogenous variability to familiar concepts, such as fluctuations of motivation, attention, distractibility, fatigue, arousal, or mind wandering, which may also unfold at time scales larger than one trial. It remains unclear to what extent these constructs or meta-cognitive descriptors can contribute to *explain* variability (beyond providing a label for aspects of it), but if they indeed bear some relationship to relevant internal brain and bodily states, it would be intuitive to think that these can be used to reduce variability and improve performance.

Of these meta-cognitive constructs, the concept of mind wandering in particular has received growing interest over the last decade. Mind wandering refers to the subjective report of losing mental focus on a task, instead focusing on

thoughts that are not directly task-related (e.g. Cheyne, Solman, Carriere & Smilek, 2009; McVay & Kane, 2012). Studies designed to investigate this metacognitive construct often use the 'probe-caught' method (Weinstein, 2017), in which participants are interrupted during their task with a probe about their level of "on-taskness" or the amount of mind wandering they experience. Higher levels of mind wandering on these probes have been associated with higher RT variability just before the probe (Laflamme et al., 2018; Seli et al., 2013; Thomson et al., 2014). This may imply that: 1) people are able to report when their thoughts are on- or off-task, 2) this subjective report bears some relation to their recent performance, and as such, that 3) participants can access some aspects of their internal fluctuating states (but see Discussion). However, even if relevant information were available on these internal states, the extent to which people could use it to reduce their own variability or improve their upcoming performance is rarely addressed.

Mind wandering is tightly linked to the more traditional cognitive concept of attention, although the exact relation remains unclear. A possible distinction may be the level of awareness: While mind wandering requires some form of awareness (even if this awareness is 'post-hoc'), as it is primarily a subjective mental state, this may not necessarily be the case for episodes of low task-focus (also known as lapses of attention). Indeed, mind wandering is often divided into two categories: 'tuning out' (during which one is aware of the mind wandering episode as it occurs) and 'zoning out' (for which awareness only occurs after the episode has finished). These stages may also be seen as degrees of severity, with 'tuning out' being characterised by a flexible division of focus between on- and off-task thoughts (Cheyne et al., 2009; Smallwood, McSpadden & Schooler, 2007). Such severity is considered to come about sequentially, with mind wandering episodes starting off shallow and deepening over time (Cheyne et al., 2009; Mittner, Hawkins, Boekel & Forstmann, 2016).

Like mind wandering, attention has been linked to behavioural variability. It has been said that "attention quenches variability" (Masquelier, 2013, p.8), as more attention and higher predictability correlate with lower variability on both a neuronal and a behavioural level (Cohen & Maunsell, 2009; Ledberg, Montagnini, Coppola & Bressler, 2012; Mitchell, Sundberg & Reynolds, 2007). Lapses of attention are typically suspected when RTs are very slow, but also when they are extremely short (so called *'anticipations'*, Cheyne et al., 2009), the combination of which leads to

increased variance. Yet another link between attention and mind wandering is that patients with Attention-Deficit and/or Hyperactivity Disorder (ADHD) are typically thought to suffer from lapses of attention, and have been reported to show higher variability as well as higher spontaneous mind wandering in comparison to a non-clinical population (Seli et al., 2015; Shaw & Giambra, 1993; see Kofler et al., 2013 for a meta-analysis; see Tamm et al., 2012 for a review).

Although in the literature, there is a strong reliance on attention and mind wandering as causal factors for behavioural variability, it remains unclear what these concepts exactly refer to, how they relate to each other, and how they are exactly linked with variability. Still, it seems intuitive that variability in performance is caused by fluctuations in some underlying brain and body states. Our main question here is whether such states are accessible to us and whether we can act upon this information.

Reducing variability with control?

The potential use of control in reducing variability may seem intuitive when looking at sports. For instance, when thinking about playing darts, there is the implicit assumption that people have access to some internal states as well as means to act upon them – leading them to throw the darts when they 'feel ready for it'. When playing darts, people may feel that they have the ability to wait until they feel fully attentive to the board and to throw on this exact moment. Within this framework, taking away one's ability to self-pace their darts game should deteriorate performance. However, while the origins of variability in dart throwing have been of interest in sports and movement psychology (e.g., van Beers et al., 2013; Smeets, Frens & Brenner, 2002; Stins et al., 2018), this specific prediction seems not to have been empirically tested so far. For now, it remains unknown what constitutes this feeling of 'being ready', how it links to our internal states, and whether it actually influences performance.

Unlike in a game of darts, in a traditional experimental psychology paradigm, timing of actions is carefully planned and controlled: The time from each trial to the next ('inter-trial interval'; ITI) is determined externally, either by an absolute timing or by a jitter with a fixed range and mean. Being in an unfavourable internal state when the trial starts or when the target appears could lead to poor performance on

that trial. Thus, giving participants control over the timing of the task – by letting them start a new trial whenever they feel ready for it, thus creating a 'self-paced' task – may enable them to reduce their variability, by preventing extreme RT and errors.

To our knowledge, this is the first study that compares a self-paced condition (in which participants determine their own ITI) to 'forced-paced conditions' (in which the ITIs are calculated from the self-paced ITIs) as a means to reduce variability and improve performance. Kelly et al. (2001) investigated the effects of 'self-pacing' versus 'forced-pacing' on temporal structure of choice RT. However, in their study, the 'pacing' refers to the maximum response time allowed after stimulus onset – as a means to manipulate the difficulty levels of the conditions. While participants were thus given some form of control (they could allow themselves more or less time to respond, and this triggered the onset of the next trial), their design does not address the question of the current research – whether control to *start* a trial can help improve on-going performance and reduce variability. Specifically, Kelly et al. (2001) investigated differences in temporal structure of choice RT series on a four choice serial RT task, and found that RT series in the 'self-paced' condition (in which response time was unlimited) indeed showed less long-term dependency (i.e. being closer to white noise) compared to the 'forced-paced' conditions (a 'fast' version, in which the maximum response time was the mean of the self-paced condition, and a 'slow' version, in which the maximum response time was the mean plus two standard deviations of the self-paced condition). They also looked at performance (but not variability), and found that mean RTs were higher in the self-paced condition. However, because both of their forced-paced experiments consisted of a fixed ITI, while the self-paced condition was not fixed but rather differed from trial to trial, findings may therefore be attributed to differences in the variability of the ITIs.

Our aim is to test whether participants can access their fluctuating performance-related internal states and have the means and will to act upon these to improve their performance (referred to as *Hypothesis 1* or *H1* throughout the article). The alternative hypothesis (*Hypothesis 2* or *H2*) is thus that people either have no access to performance-related internal state, or no will to act accordingly or no means to improve their performance as a result. In most of the tests below, but not all, *H2* is equivalent to the null hypothesis. Because of this, we use Bayesian

statistics throughout the article in order to assess the evidence in favour of *H2* even when it is equivalent to a null finding.

In our first experiment, we test *H1* within its intuitive framework: With a darts-based task. Highly motivated participants played a game of darts both with and without control over when they could throw. If *H1* is true, participants should be able to use the control in the darts game to obtain higher and less variable scores compared to when they have no control. Under the alternative hypothesis (*H2*), no decrement in performance would be expected when control is taken away from participants.

The second experiment uses a computer-based design consisting of two different tasks (easy and hard visual detection tasks) – in order to converge two different literature fields (fast action selection and visual perception). In these two tasks, participants are given control or no control over the ITI. The goal of the second experiment is three-fold. First, to replicate and to generalise our findings from Experiment 1 over various forced-paced control conditions. Second, to test another two predictions of H1 related to the RT and ITIs (which were not available in Experiment 1), namely that 1) long ITIs should be associated with better performance, and 2) RT series in the self-paced condition should show fewer temporal dependencies. Third, Experiment 2 allows for closer examination of the self-paced ITIs themselves, to see how participants might use the control they are given.

Experiment 1 – Testing the use of control in a darts task

Rationale and Predictions

The first experiment involved participants throwing darts in self-paced and in forced-paced manners. There are multiple advantages to using darts: 1) there is a clear intuitive link between darts, control, and insight into perceptual-attentional states, as discussed in the Introduction, 2) similar to laboratory experiments, darts involves performing the same action over and over again, 3) unlike laboratory experiments, people typically can play darts for a good deal of time without getting bored, 4) the

darts board can be set up with a scoring system that allows for a measure of performance and, 5) participants can easily understand what constitutes ‘good’ and ‘bad’ performance (an explicit monetary reward was used to reinforce this) and 6) participants would be motivated to get the best performance, and thus, motivated to take advantage of the control when offered to (motivation was also independently assessed via a questionnaire).

The darts task consisted of two conditions: 1) the Self-paced condition, in which participants throw the dart whenever they feel ready, and 2) the Forced-paced condition, in which participants are instructed to throw in a forced-paced (but comfortable) manner according to a tone. To further increase motivation, social competition in pairs (Tauer & Harackiewicz, 2004) and a random lottery reward system (Cubitt et al., 1998) were used – both of which have been shown to be effective for increasing motivation in participants.

If participants can use the control in the Self-paced condition to throw at the ‘right’ moment (H1), this should result in higher average scores (darts closer to bull’s eye) and lower variability compared to the Forced-paced condition. However, if they cannot use the control (H2), performance and variability should be similar under both conditions.

It is important to note that the measure of variability does not stand on its own. All in all, we are looking for *consistently good* performance – meaning that the variability should be interpreted in light of the performance and not as a sole measure of performance (particularly since reducing variability was not part of the instructions – participants were instructed to perform well, but were not explicitly told to be consistent). For example, a lower mean score in combination with lower variability would indicate that participants are consistently worse, not better. Instead, consistently good performance would be reflected in the combination of higher scores and lower variability.

Because a self-paced darts task may be more familiar to participants compared to throwing darts to a tone, we analysed the scores over block, to examine if potential practice effects would be different between the conditions even after the initial training phase. An additional analysis was conducted on the scores of the last block only, as these blocks should be the least affected by practice effects.

Methods

Participants

In total, 38 participants (24 female, 19-39 years, M_{age} = 24.1 years) with normal or corrected-to-normal vision were tested. All participants were right-handed. They were paid £8 or received course credit as a base rate for participation (excluding reward).

The study was approved by the local ethics committee. At the end of the experiment, all the participants filled the Intrinsic Motivation Scale (IMI; McAuley, Wraith & Duncan, 1991). One participant was excluded from analyses because of a low motivation score (less than half of the possible highest score of 144, making her a statistical outlier) – leaving 37 participants for analyses, whose average IMI score was 109 (*SD* = 9.5, range 88-126).

As Bayesian statistics were used, it is possible to continue recruitment until the evidence reaches a set threshold (Rouder, 2014). First, we collected a sample of 22 participants. Afterwards, sample size was sequentially increased until the median Bayes Factor either reached 6 (indicating that the data is six times more likely under H1 than under H2) or 1/6 (indicating that the data is (1/6) = .16 times as likely under H1 than under H2; or in other words, that the data is 6 times as likely under H2 than under H1) – which has been proposed as a reasonable threshold for early research (Schönbrodt, Wagenmakers, Zehetleitner & Perugini, 2017).

Materials

The darts game was played using a 45.1cm by 45.1cm Winmau dartboard and twelve nylon shafted Winmau darts. The board was covered with printed target sheets with 20 black and white rings (see Figure 1B). The scores of the rings went up by one point per ring with the most outer ring being worth one point and the bull's eye (inner circle) being worth 20 points. For each participant, four target sheets were collected: one for each training condition, and one for each experimental condition.

The experiment was run using MATLAB 9 (The MathWorks, Inc., Release 2016a) and Psychtoolbox-3 (Brainard, 1997; Kleiner et al., 2007; Pelli, 1997). Tones were presented over Logitech s-150 USB digital speakers (Logitech, Lausanne,

Switzerland). During the experiment, participants' scores were recorded using a scoring sheet.

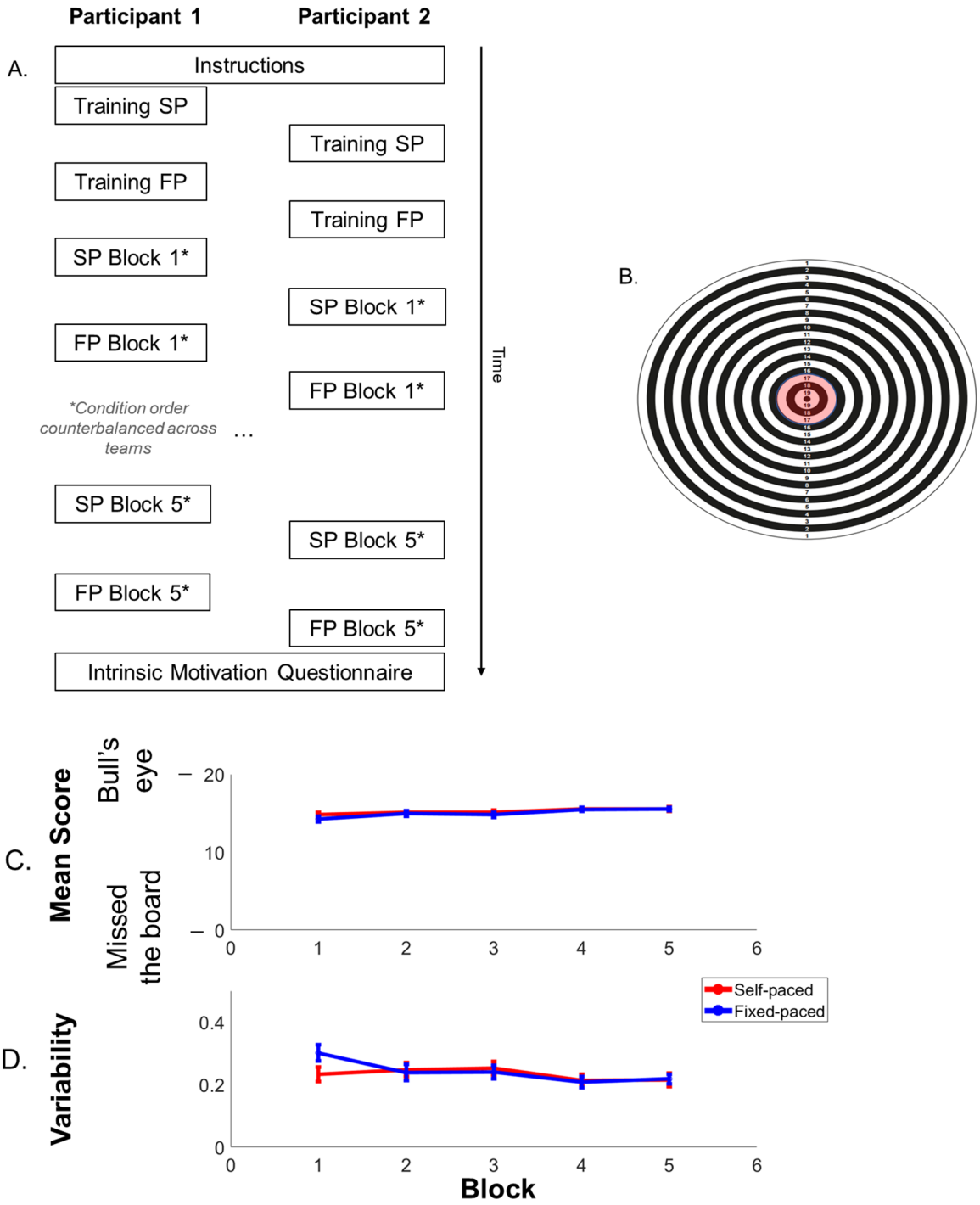


Figure 1. A. *Structure of Experiment 1. Participants played darts in pairs. In turns, they would first perform a training of the Self-paced (SP) condition, to get used to throwing the darts, and then a training of the Forced-paced (FP) condition, to learn*

the rhythm of the tones. Next, they would play five blocks of each condition, with the order of the conditions being counterbalanced across pairs. Each block consisted of twelve trials. At the end, participants filled in the Intrinsic Motivation Scale. ***B.*** *Target sheet (A2-sized) covering the dartboard, indicating the points for each ring, from the outer ring (1 point) to the bull's eye (20 points). Trials in which participant scored within the four most inner circles (in red) qualified for reward.* ***C.*** *Average score per dart on each block of twelve trials on the Self-paced (red) and Forced-paced (blue) condition.* ***D.*** *Average coefficient of variation (CV =* SD_{score} */* $\text{Mean}_{\text{score}}$*) on each block. Error bars show the within-subject standard error. There was no effect of condition (arguing against Hypothesis 1).*

Design

Both tasks had two conditions: Self-Paced and Forced-paced. In the Self-Paced condition, participants were instructed to throw the darts one by one in their own tempo – providing them with control over the timing of the action. In the Fixed condition, participants were instructed to throw in a fixed rhythm. On each trial, they heard three tones:

1) A low tone to indicate 'ready', on which participants were instructed to pick up a dart – followed by 1000ms of silence.
2) A low tone to indicated 'steady' – on which participants were instructed to get into a throwing position – followed by 1500ms of silence.
3) A low tone to indicate 'go' – on which participants were instructed to throw the dart – followed by 1000ms of silence before the next trial started.

The timing of the Forced-paced condition was based on pilot data, designed to ensure that the Forced-paced condition would be comfortable for participants and would have similar block durations as in the Self-paced blocks. In the main experiment, the Self-paced blocks turned out to have lower block durations on average than the Forced-paced blocks – and as such, any potential poor performance in the Forced-paced condition could not be due to the participants not having enough time to throw.

Procedure

Participants were tested in pairs, with the full session lasting about an hour. Figure 1A shows the complete timeline of a session. First, participants were given instructions on the structure and rules of the experiment. Then, they chose the order in which they played. In total, each participant completed two training blocks (one for Self-paced and one for Forced-paced) and ten experimental blocks (five for Self-paced and five for Forced-paced). Each block consisted of twelve trials. Participants would throw six darts, have a short break in which the experimenter would get the darts off the board, and then throw six more darts.

Participants alternated their game between each block: The first participant would play one block of one condition, next, the second participant would play the same block of the same condition, then the first player would play one block of the other condition, and finally the second participant would play one block of the other condition. Between each block, the experimenter switched the paper targets on the board. After both participants had finished a block, the total scores would be compared, and one participant was named the winner of that block. For the experimental blocks, half of the pairs started with the Self-paced condition and half of the pairs started with the Fixed-paced condition to counterbalance for an order effect. Training blocks were exempt from counterbalancing: To get used to throwing with the darts, all pairs started with the Self-paced training, followed by the Fixed-paced training.

The dartboard was hung at a height of 153 cm. Participants stood at 152cm from the board. A line of masking tape was put on the floor to indicate where they had to stand exactly. The six darts were laid out in a row on a table next to them. At the beginning of each run of six darts, the experimenter told the participant when to start and pressed a key on the keyboard to record the start time. At the end of the run, the experimenter again pressed a key, to obtain the total time of the run.

At the end of the game, the experimenter drew a random trial number and checked for both participants if they were eligible for the extra reward: If a participant had a score of seventeen or higher on that trial (four most inner circles), he/she would receive £5 extra, but if the score was sixteen or lower, he/she would only receive the base rate of £8. This cut-off was chosen to get a 20% chance of winning the reward (based on pilot data).

Results

Training trials were excluded from all analyses. Average scores and CV (coefficient of variation, equal to standard deviation of score divided by the mean score) were calculated over the five blocks of twelve trials. All statistics were Bayesian and conducted in JASP (JASP Team, 2017), using equal prior probabilities for each model, and 10000 iterations for the Monte Carlo simulation. Participants performed quite well on the task (group mean score across all blocks and conditions = 15.1, SD = 1.71, CV = .11; mean across blocks for Self-paced = 15.2, SD = 1.7, CV = .11, mean across blocks for Fixed-paced = 15.0, SD = 1.76, CV = .12 – see Figure 1 for a graphic break-down for both conditions over the blocks).

First, to assess the overall effect of condition, Bayesian 2x5 Repeated Measures (RM) ANOVAs were performed on the scores and CV, with condition and block as factors. Figure 1C and D show the means and CV of the Self- and Fixed-paced conditions over the five blocks. For both measures, the model with only block as factor performed the best. Table 1 shows the BF_{01} for each model – reflecting how much more likely the data is under the best model compared to each other model. For example, for mean score, the data is 68806 times more likely under the 'block only' model compared to the 'condition only' model. Furthermore, Table 1 shows the $BF_{inclusion}$ of each factor – which reflects the average of all models that include that factor. For both measures, only the $BF_{inclusion}$ for block is above 1. All in all, adding the factor 'condition' lowers the likelihood of the data compared to a 'block only' model. These results show there is a clear effect of practice but provide no evidence for an effect of condition.

Secondly, to directly assess H1 versus H2, Bayesian Paired t-tests were conducted on the last block of both conditions, looking both at mean and CV. As H1 specifically predicts an improvement in the self-paced condition, while H2 could predict either no difference or worsened performance, the t-tests were conducted one-sided. There was moderate evidence for H2 over H1 on the score (BF_{21} = 5.8) and CV (BF_{21} = 6.3). Note that data collection was only stopped until the median of these two tests reached 6.

Lastly, two-sided Bayesian Independent Samples t-tests were conducted on the scores and CV on each block for each condition, using order as grouping

variable. There was no evidence for such confound on either measure (BFs ranging between .30-.96).

Table 1. *Statistical outcomes of the Bayesian RM ANOVAs on the mean score and the coefficient of variation (CV) of score, using condition and block as independent factors.* BF_{01} *reflect the Bayes' Factors reflects how much more likely the data is under the best model ('block only') compared to each other model. The* $BF_{inclusion}$ *reflects the average of a factor over each model in which it is included.*

	BF_{01}		BF_{inc}	
Model	**Mean**	**CV**	**Mean**	**CV**
Block *(best model)*	1.00	1.00	24557	4.04
Condition	68806	27.53	.38	.16
Condition * Block	*NA*	*NA*	.01	.47
Condition + Block	2.61	6.30	*NA*	*NA*
Condition + Block + Condition*Block	57.30	13.33	*NA*	*NA*
Null model	23913	4.41	*NA*	*NA*

Position variability

Note that the interpretation of CV may not be straightforward, as it only reflects the variability of the raw scores, and not the variability of the position on the board. For example, imagine that a participant throws one dart on ring 14, one on 15, and one on 16. These darts could be close together or scattered over the board, but the measured variability would be the same in either case. To control for this, the cartesian coordinates of the darts were also extracted from the A2 sheets as distance from the centre. Bayesian one-sided paired t-tests were conducted on the combined variance of the horizontal (*x*) and vertical (*y*) coordinates, as calculated by:

$$\sigma_{x+y}^2 = \sigma_x^2 + \sigma_y^2 + 2r_{x,y}\sigma_x\sigma_y$$

There was no evidence for either *H1* or *H2* on this combined variance (BF_{12} = 1.4), nor in the standard deviation of just the x- (BF_{12} = .8) or y-coordinates (BF_{12} = 1.5).

Interim discussion 1

Overall, we found no evidence for a benefit of control in Experiment 1; throwing darts in a self-paced manner did not lead to higher performance or reduced variability compared to throwing on a fixed rhythm. When looking only at the scores of the last block, in which participants are most familiar with both of the conditions, we found moderate evidence against an effect of control – suggesting that if there was any initial benefit of throwing in a self-paced manner, it was due to unfamiliarity with the paced protocol.

There was also no evidence for reduced variability with control when looking at the landing position of the darts. However, this measure of variability has its drawbacks. Most importantly, participants were instructed to maximise the scores, and not to reduce variability in landing position, therefore the scores are easier to interpret. Furthermore, the same target papers were used across blocks and the darts positions were only extracted afterwards, so the temporal information was lost. While our score-based analyses show that the factor 'block' explains the most variance, it cannot be included in the position-based analyses. It is therefore likely that the latter include more unexplained variance – which increases the chance of statistical errors.

One limitation of Experiment 1 is that the two conditions are quite different from each other in terms of timing – similar to the design of the Kelly et al. (2001) study. Rather than using a standard fixed pace for each participant, using the participants' own self-paced timings may provide an improved control condition. However, this is difficult to achieve in the darts experiment, as we did not possess a way to easily measure RTs. Therefore, we aimed to replicate our findings in a computer-based experiment, to have more flexibility over the timing of the forced-paced condition. This experiment will also allow us to measure temporal dependencies in RT series. Another limitation of Experiment 1 is the relatively low number of trials per participant (60), while the complex manual action is sensitive to trial-to-trial motor noise – the combination of which may lead to decreased statistical

power. Due to its traditional set-up, Experiment 2 has a larger amount of trials and is therefore more sensitive to capturing the mean score and intrinsic variance.

Because the self-paced ITIs are recorded in Experiment 2, it allows us to examine these in more details, to see what potential strategies participants may use while handling the control. The analyses for Experiment 2 are therefore split into two parts, with the first part being focused on the effects of control, and the second on characteristics of the ITI.

Experiment 2 – Testing the use of control in two computer-based tasks

Rationale

The second experiment involved two different tasks: An easy, action-oriented task (a rapid action selection task) and a difficult, perception-oriented task. The action-oriented task is easy to perform, and therefore participants will immediately notice their own errors. The perception-oriented task involves near-threshold stimuli tailored to produce 25% errors on average. These two tasks aim to cover two different literatures: The mind wandering literature, in which it is common to use simple tasks that are highly familiar and repetitive in nature (see for example: Cheyne et al., 2009; Seli et al., 2013; Thomson et al., 2014) – as these types of simple tasks are well-suited for inducing mind wandering (Cheyne et al, 2006; Giambra, 1995) – and the literature on perception and noisy decision making (see for example: Ergenoglu et al., 2004; de Graaf et al., 2015; Romei et al., 2008; 2010), in which it is common to use visually-challenging detection tasks.

In both tasks, a target appeared either on the left or right side of the screen on each trial, and participants were asked to indicate on which side the target appeared. Both tasks consisted of four conditions: 1) Self-paced, in which participants manually start each trial themselves, 2) Fixed, in which the ITI is the same for each trial, 3) Replay, in which the ITIs of the self-paced condition are replayed in the exact same order, and 4) Shuffled replay, in which the ITIs are replayed in a shuffled order. The conditions were inspired from Marom & Wallach (2011), although their research question was different from ours. Importantly,

because the self-paced ITIs differ from traditional ITIs on multiple aspects, the three forced-paced conditions were chosen such that each of them allows for comparison with the self-paced condition over a different aspect (see Table 2 for an overview). This means that to ascribe any found difference to an effect of control, the result has to be consistent over all three forced-paced conditions.

Table 2. *Summary of the four different conditions and the main characteristics of the ITIs. The three forced-paced conditions (shaded in grey) allow for comparison to the Self-paced condition on these different characteristics.*

		Condition		
Aspect	**Self-paced**	**Fixed**	**Replay**	**Shuffled replay**
Control over ITI	Yes	No	No	No
Predictability of trial onset	Yes	Yes	No	No
Variability of ITI	Yes	No	Yes	Yes
Time structure in ITI	Yes	NA	Yes	No

The Fixed condition is most similar to both the forced-paced condition of Experiment 1 and to traditional experimental designs. Due to the repetitive nature of the Fixed condition, we can examine the effects of self- versus forced- pacing when target onset is always predictable. The Replay condition is an exact replica of the self-paced ITIs and thus has the same variability. In terms of timing – but not of predictability – the Replay condition is thus most similar to Self-paced. However, the self-paced ITIs will likely contain temporal dependencies[4] – similar to typical RT series. As traditional experimental designs do not include such temporal dependencies in their ITIs, their potential effects are unclear. Therefore, we also included the Shuffled Replay condition, in which these dependencies are removed.

Table 3 gives an overview of the two hypotheses and their corresponding empirical predictions and findings over Experiment 1 and 2. To contrast our hypotheses, we investigate the effects of task-control and spell out four empirical tests, the predicted outcomes of which differ across hypotheses. We compare

[4] Note that in the section “Characterising the self-paced ITIs in the computer-based tasks” and in Supplementary Table 1 (p. 212), we confirm that the self-paced ITIs indeed contain temporal dependencies, preserved in the Replay condition.

performance (RT and accuracy), intra-individual variability (CV of RT), and serial dependencies in the self-paced condition with the forced-paced conditions. With Test 1 and 2, we aim to replicate the findings from Experiment 1 – that participants cannot use the control to improve their performance and reduce their variability. Test 3 and 4 offer additional tests of *H1,* examining the impact of long ITIs on performance and contrasting temporal structures in RT series across conditions.

Table 3. *Summary of the two alternative hypotheses and their respective predictions over the two experiments. Green shading indicates those predictions that were supported by the data in the present article. Evidence favoured H2 (people have no access to performance-relevant inner states or no will/means to act upon it) over Hypothesis 1 (people have access to performance-relevant inner states and will plus means to act upon it).*

Empirical predictions Experiment 1	**Hypothesis 1**	**Hypothesis 2**
Improved performance and reduced variability in Self-paced	Yes	No
Empirical predictions Experiment 2		
1. Improved performance and reduced variability in Self-paced	Yes	No
2. Reduced extreme RTs	Yes	No
3. Performance following long self-paced ITIs is:	Better	Worse
4. Reduced temporal dependencies in RTs in Self-paced	Yes	No

Unlike Experiment 1, Experiment 2 comes with the possibility of recording self-paced ITIs, and therefore using these in designing the forced-paced conditions. In order to record and replay the self-paced ITIs, the Self-paced condition will have to come first. Although we showed in Experiment 1 (which allowed for counterbalancing of conditions) that order did not matter, a concern might be that participants could continue to show training effects in the Self-paced condition – which could mask differences between the conditions. To anticipate, we found no evidence for such training effects, making it unlikely that the results are explained by condition order.

Test 1. The effect of control on performance and variability

First, we aimed to replicate the results from Experiment 1, i.e. that having control does not lead to improved performance or reduced variability. For both tasks, we calculated for each condition: 1) mean reaction time, 2) percentages of errors, and 3) coefficient of variation of RT (CVRT). Again, if we do not have access to our own internal performance-related rhythms or have no means to act upon them, having control over the timing should *not* lead to lower RT means, lower error rate, or lower CVRTs.

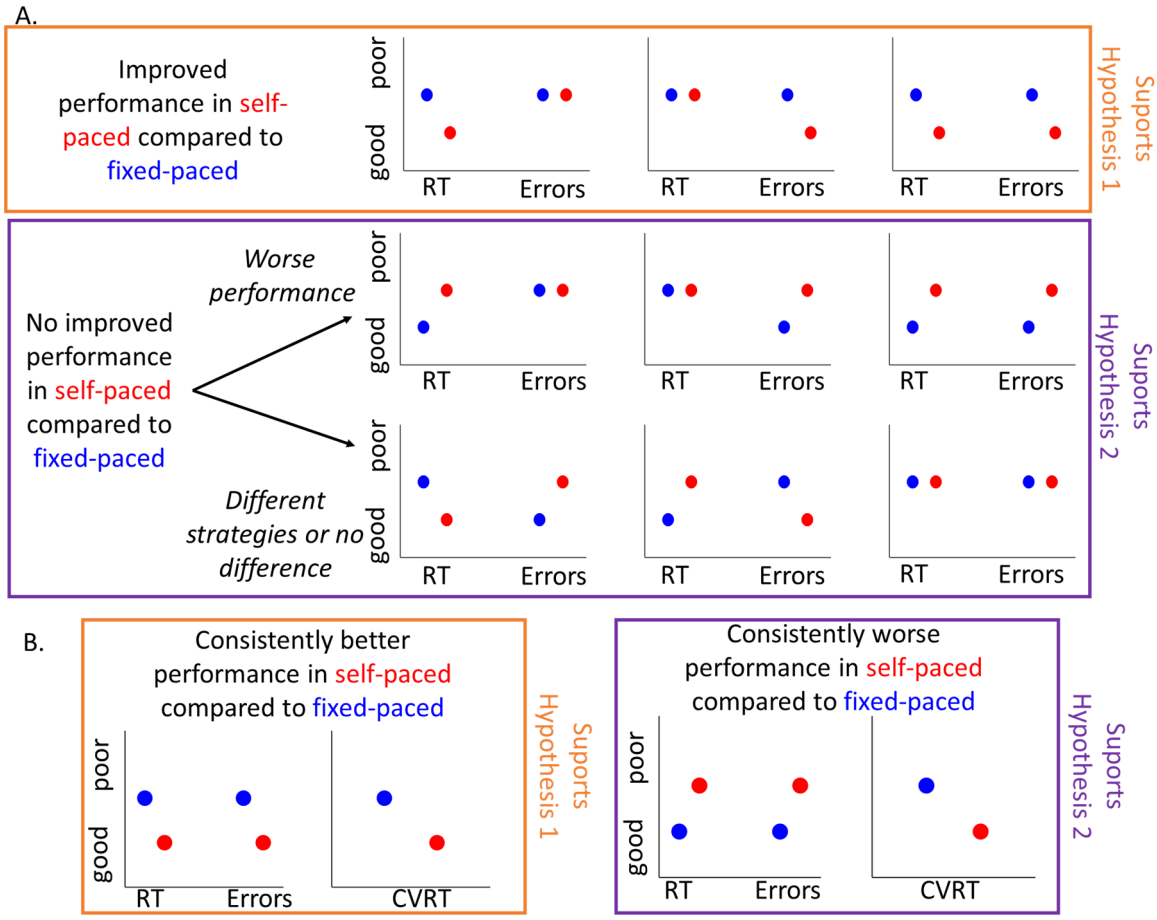


Figure 2. Overview of the empirical predictions and their respective conclusions in graph space, showing that it is important to interpret the measures concurrently rather than separately. This means that H1 is only supported when A) performance is improved in the Self-paced condition on RT, error percentage, or both. In contrast, H1 would not be supported when one measure is improved but the other is deteriorated the Self-paced condition – instead implying a difference in speed-accuracy strategy only. B) Similarly, decreased variability in the Self-paced condition

would only support H1 if it is combined with improved RT/error measures – showing consistently better performance.

Just as in Experiment 1, it is important to not interpret the measures individually. When investigating RT and accuracy, good (or poor) performance is not just indicated by each of them separately, but by the combination of the two. For example, a reduced mean RT with an increased error rate is not indicative of improved performance as such, as it could also reflect an adjustment of speed-accuracy trade-off – see Figure 2A for an overview of different patterns and their respective interpretation Here, we use the EZ-diffusion model to investigate this in more detail (see Test 1b). Furthermore, we are again looking for *consistently good* performance – meaning that the variability can only be interpreted in combination with performance (see Figure 2B for examples).

Test 1b. The effect of control on performance as EZ-diffusion model parameters

The EZ-diffusion model was used to disentangle strategy adjustments from true performance improvements. The EZ-diffusion model is based on the drift-diffusion model (DDM; Ratcliff, 1978), which is a computational model for two-alternative forced choice tasks – in which participants have to make a choice between two options (in this case, 'left' or 'right'). The model assumes that evidence accumulates between two boundaries, each representing one response option, until one of them is reached, which initiates the corresponding response.

The EZ-diffusion model is a simplified version of the DDM (Wagenmakers, Van der Maas & Grasman, 2007), which uses calculations rather than a fitting procedure. It provides three parameters: 1) drift rate (v), which reflects the rate with which evidence is gathered (or in other words, how quickly information is processed), 2) boundary separation (α), which reflects a response criterion (or in other words, reflects how much evidence is needed before an action can be initiated), and 3) non-decision time (T_{er}), which reflects the time spent on any processes but decision making (such as sensory and motor execution). Improved performance may be reflected in higher drift rates and/or in lower non-decision

times, while differences in speed-accuracy trade-offs may be reflected in the boundary separation.

Test 2. Reduced extreme RTs

It is possible that participants are not able to reduce the constantly ongoing ('subtler') variability in their performance and hence do not improve their mean performance, but can still use the control to avoid extreme RTs – which are considered the hallmark of severe mind wandering and lapses of attention. If severity of off-taskness indeed comes about sequentially (Cheyne et al., 2009; Mittner et al., 2016; but see Discussion), participants should be able to detect, at least sometimes, when they are in the shallow stages of mind wandering, and use the control to avoid reaching the more extreme off-task states. To test for this, the number of very long *and* very short reaction times (likely anticipations) was calculated for each condition and each participant. Under the intuitive framework of H1, in which people can wait for the 'right' moment to perform, participants in the Self-paced condition should be able to delay the start of the next trial to 'refocus' on the task, leading to a reduced amount of extreme reaction times. But under H2, there should be no difference between conditions.

Test 3. The effect of longer self-paced ITIs

To get more insight into potential ways participants may have used the ITIs, we tested whether longer ITIs reflected moments when participants waited for a more optimal moment to initiate the trial. ITIs were divided 'regular' and 'long', and the mean reaction time, coefficient of variance, and accuracy were calculated on the 'regular ITI'-trials and on the 'long ITI'-trials. If participants can effectively make use of the control – i.e. if they can use these longer breaks to wait until they feel ready to continue (*H1*) – their performance should increase and their variability decrease on the trials with long self-paced ITIs compared to trials with regular self-paced ITIs.

Alternatively, participants may simply show fluctuating good and poor modes of responding throughout the experiment over which they have no control, similarly affecting both RT and ITIs. If this were the case, these long ITIs may be indicators

of being stuck in an overall poor mode of responding, leading to poorer performance on these trials compared to trials triggered following regular ITIs.

Test 4. Time structure in the reaction time data

Kelly et al. (2001) found reduced temporal dependencies in a self-paced task compared to fixed-paced conditions. In our attempt to replicate this finding, both autocorrelations and power spectra were considered, following Wagenmakers et al. (2004). Autocorrelations measure the degree of dependency in a (reaction time) series with itself over time, by calculating the correlation between trial *n* and trial *n* + *k*, with *k* indicating the lag. Power spectra also measure temporal structures, but express this in frequencies – which allows for classification into different types of noise. Series with no temporal structures are called 'white noise', and are characterised by flat null autocorrelation functions as well as flat power spectra. It has been proposed that empirical data contains 'pink noise' or *1/f noise*, a mixture of strong short-term dependencies and slowly reducing long-term dependencies (Gilden, 2001; but see Farrell et al., 2006; Wagenmakers et al., 2004), and is characterised by exponentially decreasing autocorrelation functions and power spectra with a slope around -1. Note that the power spectra can be mathematically derived from the autocorrelations.

It has been suggested that long-term correlations in performance may reflect 'spontaneous fluctuations in attentional state' (Irrmisscher et al., 2018) – one example of the internal states our participants may aim to counteract with the control. Successful mitigation against such temporally-correlated internal states would result in reduced temporal dependencies in their RTs (i.e. closer to white noise) – reflected in reduced autocorrelations and flatter power spectra of the RTs in the self-paced condition. The temporal dependencies may instead be transferred to the self-paced ITIs (analysed in Part 2).

Methods

Participants

In total, 39 participants (32 female, 18-36 years, M_{age} = 24.5 years) with normal or corrected-to-normal vision were tested. Of them, 36 participated in the action-oriented task, and 39 participated in the perception-oriented task. Participants were paid £10/hour or received course credits for participation. Two participants in the action-oriented task and four in the perception-oriented task were excluded from analyses due to poor performance (see *Data preparation and analysis*). The study was approved by the local ethics committee.

As we are considering multiple tests in parallel (some of which are dependent on each other numerically and/or in terms of interpretation), it would have been very difficult to ensure that all of them reach a pre-determined Bayes Factor. Therefore, we again sequentially sampled until the *median* value across all tests reached either 6 or 1/6 (see Interim discussion 2). As a first sample, 24 participants were recruited. Afterwards, we sampled until the threshold was reached.

Materials

The stimuli were generated using MATLAB 8 (The MathWorks, Inc., Release 2016a) and Psychtoolbox-3 (Brainard, 1997; Kleiner et al., 2007; Pelli, 1997), using a Bits# Stimulus Processor video-graphic card (Cambridge Research Systems, Cambridge, UK) and a Viglen VIG80S PC (Viglen, Hertfordshire, UK), and were displayed on an hp p1230 monitor (Palo Alto, US) with a resolution of 1280 by 1024 and a refresh rate of 85Hz. Responses were recorded with a CB6 Push Button Response Box (Cambridge Research Systems, Cambridge, UK), which was connected to the Bits#. Participants were positioned in a chin- and head-rest, 92 cm away from the screen.

The experiment was shown on a grey background (55.8 cd/m^2), featuring a fixation dot (112.1 cd/m^2, .18°) or a fixation cross (112.1 cd/m^2, .42°). Both tasks featured a vertically oriented Gabor patch as target (spatial frequency = 1.81 c/°, sigma = .26°). In the action-oriented task, the contrast of the target was always set at the maximum of 1. The perception-oriented task featured a low-contrast (difficult

to detect) target that was adjusted to individual detection-thresholds of 75% accuracy and ranged between .021-.070 (*M* = .039 *SD* = .011).

Design

Both tasks had four conditions: Self-Paced, Fixed, Replay and Shuffled replay. In the Self-Paced condition, participants started each new trial manually whenever they felt ready –they were given control over the ITI. In the Fixed condition, the median of the ITIs in the Self-Paced condition was used as ITI-length. The ITI was thus kept fixed throughout the trials while keeping the pace as similar as possible to the self-paced trials. In the Replay condition, the recorded ITIs from the Self-Paced condition were replayed in the exact same order – thus controlling for the different ITI lengths without giving control to the participants – and in the Shuffled replay condition, the ITIs were replayed in a different order – to allow for the different ITI lengths while removing any possible time structure between the ITIs.

Procedure

The experiment consisted of four testing days of about an hour – two for each of the tasks (Figure 2). The first day of both tasks started with a training of 300 trials, followed by the Self-paced condition, and then one of the three control conditions (Fixed, Replay, or Shuffled replay). The remaining two conditions were administered on the next day. On each day, the testing session was preceded by three minutes of rest with eyes open, to provide a common baseline to all participants before starting the task.

Main Experiment. Figure 3 illustrates the time course of each trial. Every trial started with a light grey screen with a fixation dot in the centre. Each condition consisted of 300 trials, with the first 30 being training trials. In the Self-paced condition, participants were instructed to press with the left and right index fingers at the same time whenever they felt ready for a new trial. They were told that they could wait as long as they wanted before continuing, but were discouraged from taking very long breaks. The time between fixation dot onset and double key press

was recorded and subsequently used as ITI in the other conditions. Participants were unaware that their own self-paced ITIs would be used. After the button press, the dot was replaced by a fixation cross. In the three forced-paced conditions, the participant's recorded self-paced ITIs (Replay, Shuffled replay) or median (Fixed) were used to determine the time between fixation dot and fixation cross. Next, 500ms after the cross onset, a target appeared either on the left or right of the cross. Participants were instructed to indicate with a button press which side the target appeared, using their left or right index fingers. After 200ms, the target disappeared, and after another 100ms, the fixation cross disappeared. Participants were then shown a blank screen until they responded.

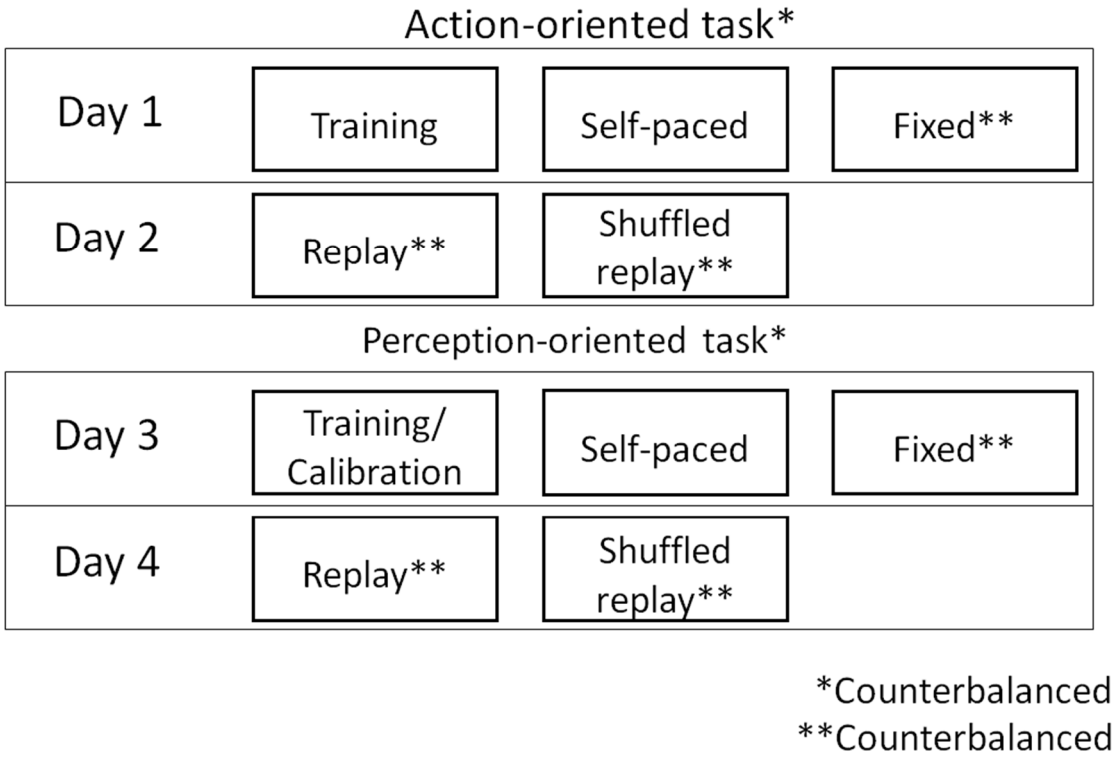


Figure 2. *Structure of Experiment 2. Each task (action- and perception-oriented) took place over two days, with the order of the tasks being counterbalanced over participants. For both tasks, participants started with a training of 300 trials followed by the Self-paced condition, and finally one of the three control conditions (Fixed ITI, Replay, or Shuffled Replay, the order being counterbalanced over participants). During the next session, they would perform the other two conditions.*

Training. Before the main experiment, each participant underwent a training using a fixed ITI of 1000 ms. After every 30 trials, participants were given feedback on their mean reaction time and accuracy. In the action-oriented task, participants were asked to be as fast as possible while avoiding errors, and in the perception-oriented task, they were asked to be as accurate as possible while avoiding producing too long RT. Again, the focus of the instructions was on good performance, and not on consistency. These instructions were repeated in the main experiment before each new condition. In the perception task, these trials were also used to determine the target contrast for each individual for the remainder of the task. The Psi method (Kontsevich & Tyler, 1999) was used to find the 75%-correct contrast detection threshold for each participant. Performance on training trials were excluded from all analyses.

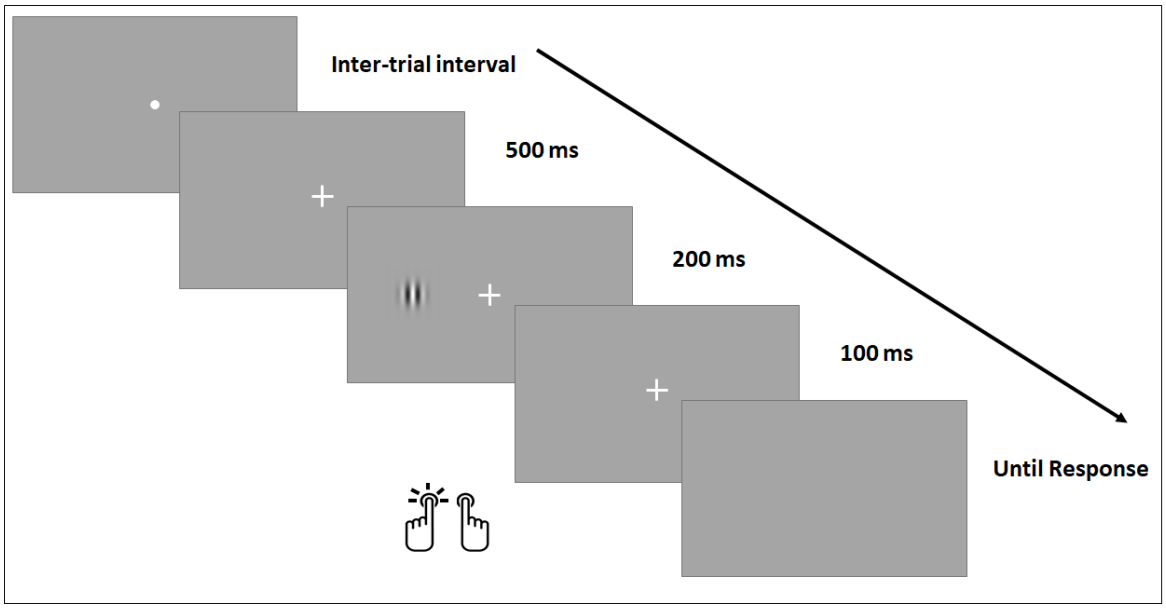


Figure 3. *Example of one trial over time in Experiment 2. The length of the inter-trial interval was manipulated over conditions. After the ITI, the fixation dot was replaced with a fixation cross. After 500ms, the stimulus (Gabor patch) appeared either on the left or the right side of the screen for 200ms. The fixation cross disappeared 100ms later, and the screen remained empty until the participants responded either with their left or right index finger.*

Results

Test 1. Participants do not perform consistently better with control

Average RT across conditions and across participants ranged from 204 to 932 ms in the action-oriented task and from 271 to 2143 in the perception-oriented task. However, participants' data were highly skewed, which had a large effect on the calculations of the mean (and variability) of the RT. Moreover, group distributions of mean RT and CVRT violated assumptions for normality. Therefore, RTs were log transformed. Because our hypotheses rely on the assumption that participants are motivated and able to perform the task, we first examined performance for each participant. One participant was excluded for both tasks due to below chance level performance on the training trials, and one participant was excluded from the action-oriented task for having more than 25% incorrect responses. Three participants in the perception-oriented task were excluded from the analysis as more than 15% of their correct RTs were outliers in at least one of the conditions (outliers included log(RT) higher than 3 standard deviations above the mean log(RT) and extreme RT - below 100 or above 1000 ms in action-oriented, and below 150 or above 1500 ms in perception-oriented task). As these participants performed poorly in all conditions, this did not bias either hypothesis.

Examining the unfiltered data of the remaining participants, average RT across conditions ranged from 204 to 592 ms in the action-oriented-task and from 271 to 1551 ms in the perception-oriented task. Mean accuracy scores were calculated for each participant and for each condition. Mean reaction times and standard deviations were calculated on the logged values of the correct trials.

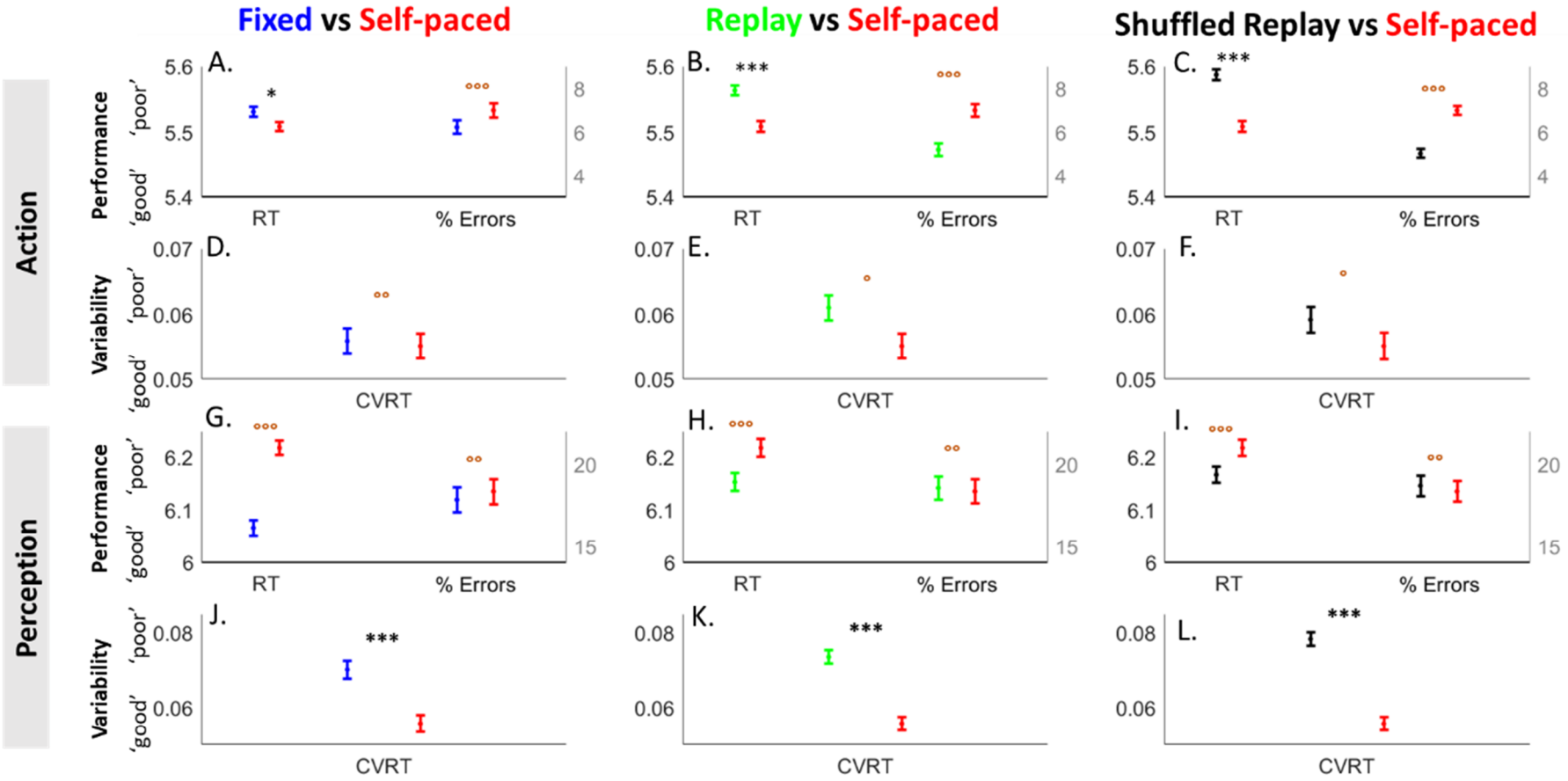
For both tasks, Test 1 involved paired Bayesian t-tests conducted on: 1) reaction time, 2) percentage of errors, and 3) CVRT, to test if the self-paced condition differed from any of the three control conditions. Because Hypothesis 1 is specifically based on *better performance* in the self-paced compared to the other three conditions, the t-tests were conducted one-sided.

Figure 4 compares the Self-paced condition to each of the forced-paced conditions on individual measures of performance (RT and percentage of errors) and intra-individual variability (CVRT). Table 4 shows the corresponding Bayes' Factors. Altogether, we did not find any consistent benefit of the Self-paced

condition over the forced-paced conditions, and evidence overall favoured *H2* over *H1*. Below the results are described in more detail.

Performance. Altogether, none of the comparisons in both tasks revealed any clear benefit of the control on performance. In the action-oriented task, the comparisons with the Fixed condition (Figure 4A) showed clear evidence against an improvement in accuracy, while the comparison on RT were more mixed. The comparison with the Replay and Shuffled replay conditions showed that participants were on average faster in the self-paced condition (providing strong evidence for *H1*), but also made more errors (providing strong evidence against *H1*, Figure 4B-C). This pattern is actually suggestive of an adjustment in speed-accuracy strategy, probably in response to the target onset being predictable (versus unpredictable in the Replay and Shuffled Replay conditions), rather than the improvement in performance expected under *H1*. This interpretation is supported by modelling using the EZ-Diffusion Model (see Test 1b). In the perception-oriented task (Figure 4G-I), all six comparisons favoured *H2* (BF21 ranging from 4.4-51.1).

Variability. In the action-oriented task (Figure 4D-F), two comparisons were in the indeterminate range and one showed moderate evidence against lower CVRT. In the perception-oriented task, all the comparisons showed strong evidence for *H1*, i.e. lower CVRT in the Self-paced condition compared to the forced-paced conditions. It is noteworthy that such reduced intra-individual variability was not accompanied by a reduction in mean RT (in fact, mean reaction time was highest in the Self-paced condition). One interpretation could be that participants made less anticipatory responses in the Self-paced condition, possibly due to the additional button press to initiate the trial. In this case, this reduction in CVRT would not be interpreted as an improvement in performance, but rather as an indication that participants are behaving differently in the self-paced condition. It should also be noted that this decrease in anticipatory responses did not lead to an increased accuracy, even though anticipations are characterised by accuracy scores at chance level (suggesting that a reduction in them would increase overall accuracy). Test 1b and Test 2 below address this in more detail.



Stats Legend: Evidence for an improvement in self-paced (BF12: *between 1-3, **between 3-10, ***above 10)
Evidence against an improvement in self-paced (BF21: °between 1-3, °°between 3-10, °°° above 10)

Figure 4. *Mean log RT, percentage of errors, and CVRT in the Self-paced condition compared to each of the three forced-paced conditions: A) Fixed ITI (blue), B) Replay (green) and C) Shuffled Replay (black). Black stars indicate evidence for an improvement in the Self-paced condition (consistent with H1), while orange circles indicate evidence against an improvement in the Self-paced condition (against H1). Top and Bottom panels show results of the action- and perception-oriented task. Error bars show the within-subject standard error across conditions. Increasing scores on the Y-axes show decreasing performance in a single measure (lower speed, accuracy or consistency), but the measures should be interpreted in relation to each other.*

Table 4. *Statistical outcomes for Test 1 – Does control improve performance and reduce variability? Shown are the Bayes' Factors for H1 over H2 for each comparison on RT, % errors and CVRT on both the action-oriented and the perception-oriented task. T-tests were conducted one-sided by contrasting the Self-Paced (SP) to each of the three forced-paced conditions, Fixed-paced (F), Replay (R) and Shuffled Replay (SR).*

BF_{12}	Action-oriented			Perception-oriented		
Comparison	RT	% Errors	CVRT	RT	% Errors	CVRT
SP < F	1.13	.09	.22	.02	.14	20.83
SP < R	51.78	.05	.99	.07	.20	1517
SP < SR	1150	.04	.49	.07	.23	48867

Test 1b. EZ-model suggests strategy-adjustments, not performance improvement

Drift rate, boundary separation, and non-decision time parameters were calculated for both tasks on each condition. Because the estimations are sensitive to outliers, extreme high RT (1000 ms for the action-oriented and 1500 ms for the perception-oriented task) were excluded before calculating the parameters. Next, Bayesian Paired t-tests were performed on i) drift rate (specifically testing one-sided for *increased* drift rate in the self-paced condition compared to the other three

conditions, which may reflect improved performance), ii) non-decision times (specifically testing one-sided for *decreased* non-decision times in the self-paced condition compared to the other three conditions), and iii) boundary separation (specifically testing two-sided for any difference between the conditions, reflecting changes in response strategies).

Figure 5 shows the means of drift rate (v), non-decision times (T_{er}), and boundary separation (α) as calculated by the EZ-Diffusion model in the Self-paced condition compared to each of the forced-paced conditions, with corresponding Bayes Factors shown in Table 5. The first two parameters may reflect differences in performance (with good performance being indicated by *higher* drift rate and *lower* non-decision times), while boundary separation indicated differences in speed-accuracy trade-off (with higher values indicating a more cautious strategy).

Altogether, in the action-oriented task, the comparisons suggest that the differences between conditions are caused by adjustments in speed-accuracy trade-off. These adjustments seem dependent on predictability of target onset rather than on control. This supports the conclusion that there is no benefit of control on performance – supporting *H2* over *H1*. For the perception-oriented task, differences are best explained by an increase in non-decision times, and thus, a decrease in performance. Again, this supports *H2* rather than *H1*. Below the results are described in more detail.

Performance. In the action-oriented task, there was no consistent improvement in the Self-paced condition (Figure 5A-C). Out of the six comparisons, none of the comparisons favoured *H1* (reduced non-decision times in Self-paced compared to Shuffled Replay), and four showed moderate evidence for *H2*. In the perception-oriented task, there was strong evidence against a decrease in non-decision times in the Self-paced condition compared to each of the forced-paced conditions (Figure 5G-I). In fact, non-decision times were higher in the Self-paced condition, with no evidence for increases in drift rate. This clearly suggests that processing of information did not improve in the Self-paced condition compared to the forced-paced conditions, but rather, that sensory or motor processes took longer (see Test 2 for complementary evidence).

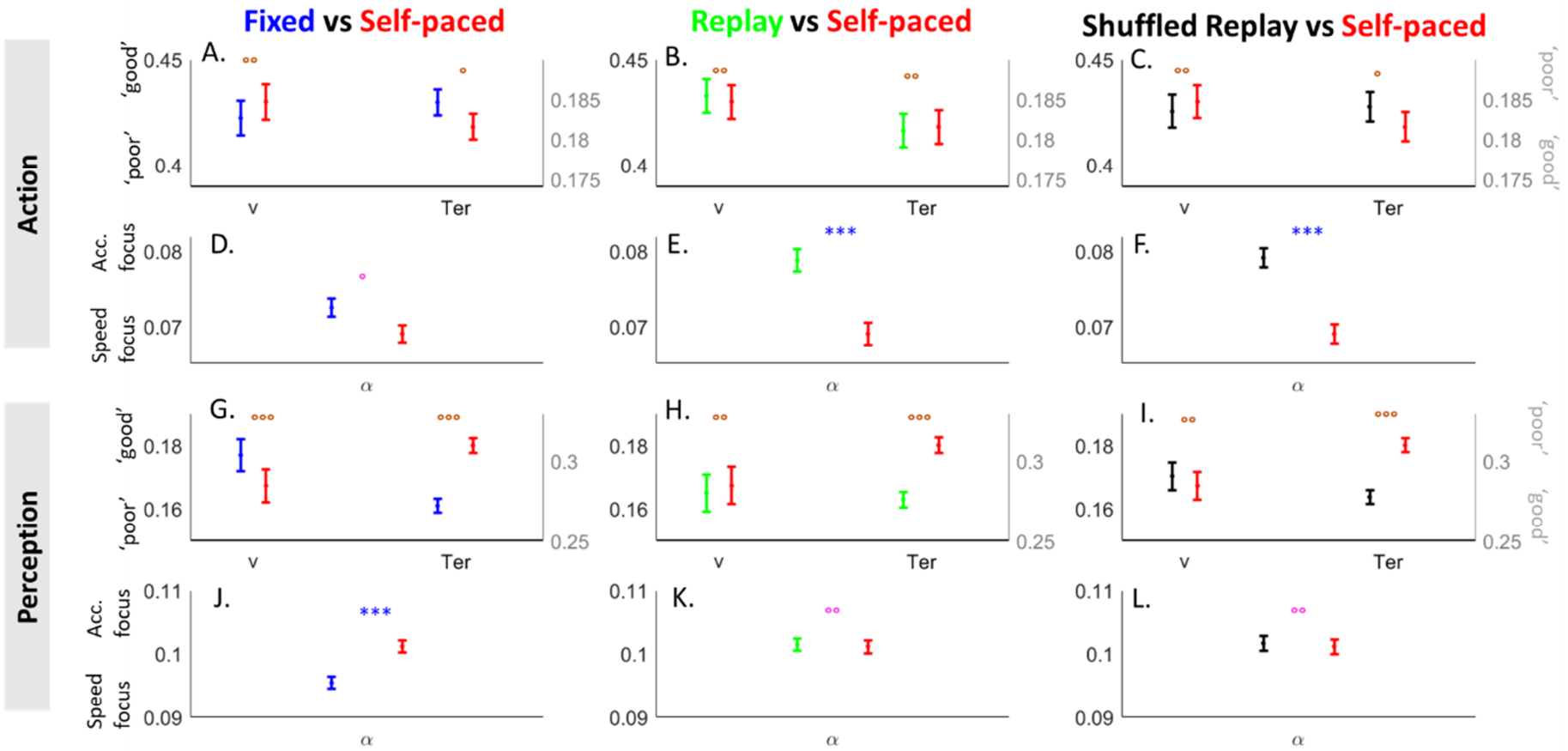


Figure 5. *Averages of the EZ-diffusion parameters on the Self-paced (SP), Fixed (F), Replay (R), and Shuffled Replay (SR) conditions. Error bars show the within-subject standard error.*

Speed-accuracy strategies. In the action-oriented task, indeed, boundary separation in the Self-paced was lower compared to the Replay and Shuffled replay condition (Figure 5B-C). High boundary separation values indicate that a lot of information needs to be gathered before one option can win (thus taking longer on the decision process, but with fewer chances of errors), while low values indicate that less information needs to be gathered before one option can win (leading to shorter RT, but reduced accuracy). Further testing showed that boundary separation was also lower in the Fixed condition compared to Replay and Shuffled replay (BF of 6.6 and 17.2 respectively), confirming that participants were less cautious when the target onset was predictable. In the perception-oriented task, this pattern was reversed: There was strong evidence for a change in caution in Self-paced compared to Fixed, with participants being more cautious overall in Self-paced. There was strong evidence against a change in boundary separation compared to Replay and Shuffled Replay (figure 5J-L). Further testing showed participants were also less cautious in Fixed compared to Replay and Shuffled Replay (BF of 14.5 and 20.6 respectively). It is possible that participants again had lower boundary separations in the predictable condition, but that this was not found in the Self-paced task due to the longer non-decision processes.

Table 5*. Bayes' Factors contrasting the EZ-diffusion parameters between the Self-paced and each Forced-paced condition (v: drift rate; T_{er}: non-decision time; α: boundary separation). Same conventions as Table 4.*

	Action-oriented			Perception-oriented		
Comparison	v*	T_{er}**	α***	v*	T_{er}**	α***
SP - F	.27	.45	.51	.10	.04	10.81
SP - R	.16	.17	12.24	.21	.05	.18
SP - SR	.23	.34	114.62	.14	.05	.19

*****Tested for higher drift rates in the self-paced condition than the other conditions.
**Tested for lower non-decision times in the self-paced condition than the other conditions.
***Tested for no difference between the conditions.

Test 2. Differences in extreme RTs

The amount of extreme reaction times (including trials defined as outliers above) in each condition was calculated for each of the participants. As a lower-bound cut-off, trials were counted if the RT was below 150 ms or below 200 ms for the action-oriented task and the perception-oriented task respectively. Trials below these cut-offs showed chance performance (i.e. an average accuracy of 50%) and as such reflect anticipations. For the upper-bound cut-off, trials were counted if the RT was above 500 ms or 1000 ms for the action-oriented and perception-oriented task respectively. Bayesian Paired one-sided t-tests were conducted, testing if the number of extreme reaction times was lower in the Self-paced condition compared to each of the fixed-paced conditions.

In the action-oriented task, anticipations were not less frequent is the Self-paced condition compared to each of the three fixed-paced conditions (see Figure 6A-C for means and Bayes' Factors), providing moderate to strong evidence for *H2* over *H1*. The number of high reaction times were more mixed: While there was moderate evidence against *H1* in the comparison with the Fixed-paced condition, the BF for Replay was close to 1, and the BF for Shuffled Replay showed a BF of 3.2 in favour of *H1.* These patterns may partially reflect the different speed-accuracy trade-offs of the different conditions.

In the perception-oriented task (see Figure 6D-F), support for *H1* was found only for anticipations, while the high RTs favoured *H2* overall. This reduction of the very short reaction times in the Self-paced condition was consistent with the overall higher mean RT compared to all the forced-paced conditions – bringing support to the interpretation from Test 1. One possibility is that this is due to the interference of the additional button press in the Self-paced condition. This interpretation is consistent with our modelling using the EZ-Diffusion model, which suggested that only non-decision times were higher in this condition.

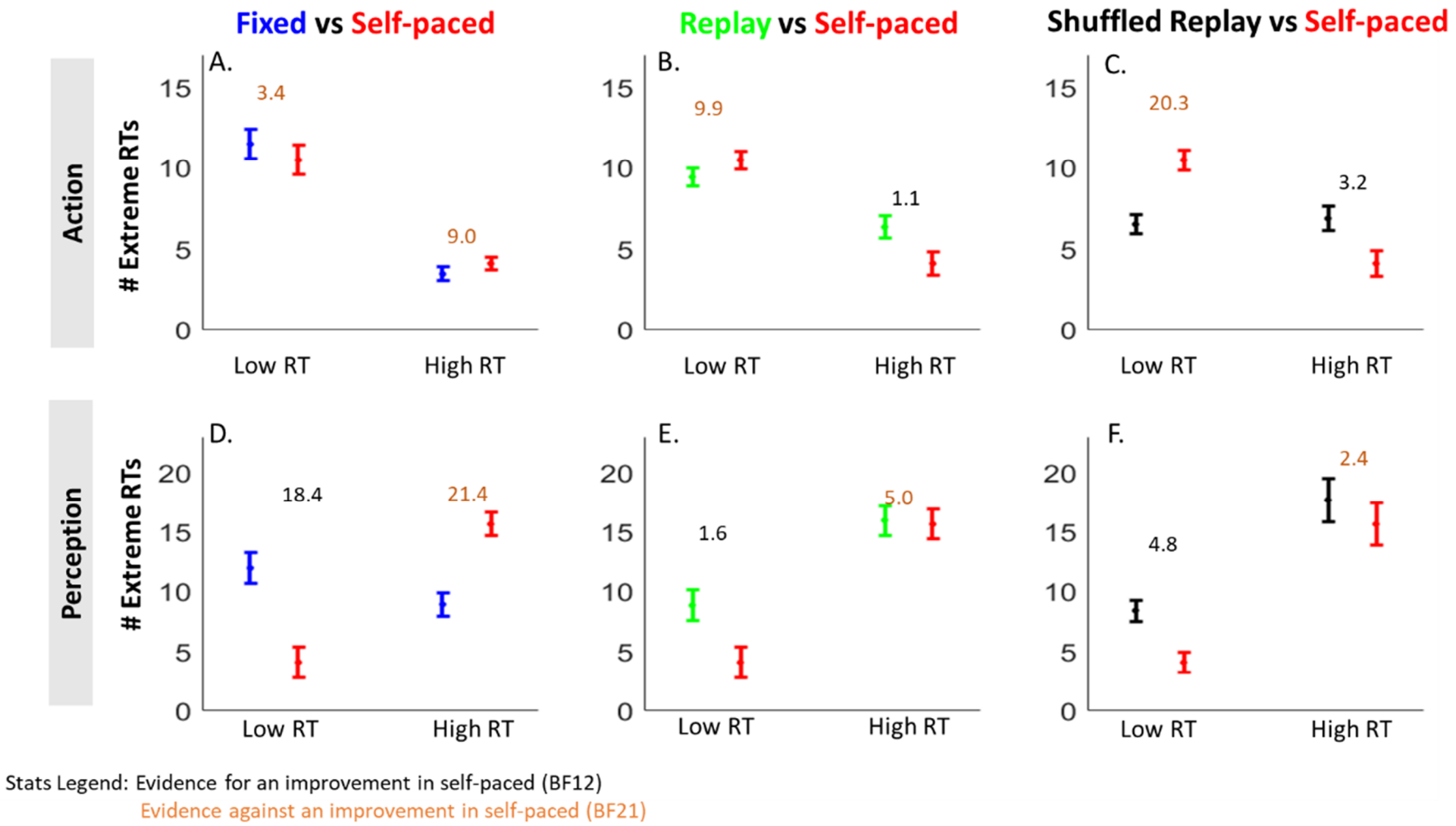


Figure 6. *Number of extreme reaction times averaged across participants. Same conventions as Figure 4.*

Test 3. Longer ITIs lead to poorer performance, not better

ITI-distributions were calculated by taking the time between the response on trial *n-1* and the self-paced ITI-press on trial *n*. To define typical and longer ITIs, the RT distribution from the Fixed condition was used as a reference: For both tasks, the 95th percentile of the RT-distribution was calculated for each participant as cut-off (see Figure 6A for an example). Self-paced ITIs below this cut-off were classified as 'regular', and may reflect as fast as possible responses to the fixation dot indicating one can start a new trial – thus resembling regular RT. ITIs above the cut-off were classified as 'long', and may reflect times in which participant felt they needed to wait longer before feeling ready to continue.

Mean error scores, mean reaction times, and standard deviations were calculated for trials following on from regular ITIs as well as for trials following long ITIs. Because there was a lot of variation in the number of regular and long trials between participants, ten trials were randomly selected 10000 times and the mean accuracy, reaction time, and CVRT over these 10000 iterations were calculated. Subsequently, Bayesian paired one-sided t-tests were conducted on these means to see if performance improved following long ITIs. Three participants in the action-oriented task and four participants in the perception-oriented task were excluded from analysis because they had less than ten trials with regular self-paced ITIs

For both tasks, evidence was found against an improvement in RT, variability or accuracy –providing moderate to strong evidence against *H1* (Figure 7B). When testing in the opposite direction (long ITIs lead to worse performance), it was found that RT and variability (but not accuracy) were clearly worse following long ITIs than those following regular ITIs (BFs of 325.0 and 740.3 for the action-oriented task, and 2841.8 and 1215.3 for the perception-oriented task respectively).

In conclusion, long self-paced ITIs did not lead to an improvement in performance or a reduction in variability. Instead, these breaks were associated with subsequent lower performance and higher variability. The co-occurring long ITIs and longer reaction times suggest the same fluctuating internal states affect both measures. To confirm this, correlation coefficients between ITI and RT on each trial were also performed. For both tasks, correlation coefficients were positive overall on the group (BF for one sample t-tests 703.9 and 436.5) – suggesting that short

ITIs are typically followed by short RTs, and long ITIs by long RTs. This could reflect similar temporal dependencies as in typical RT series on consecutive trials.

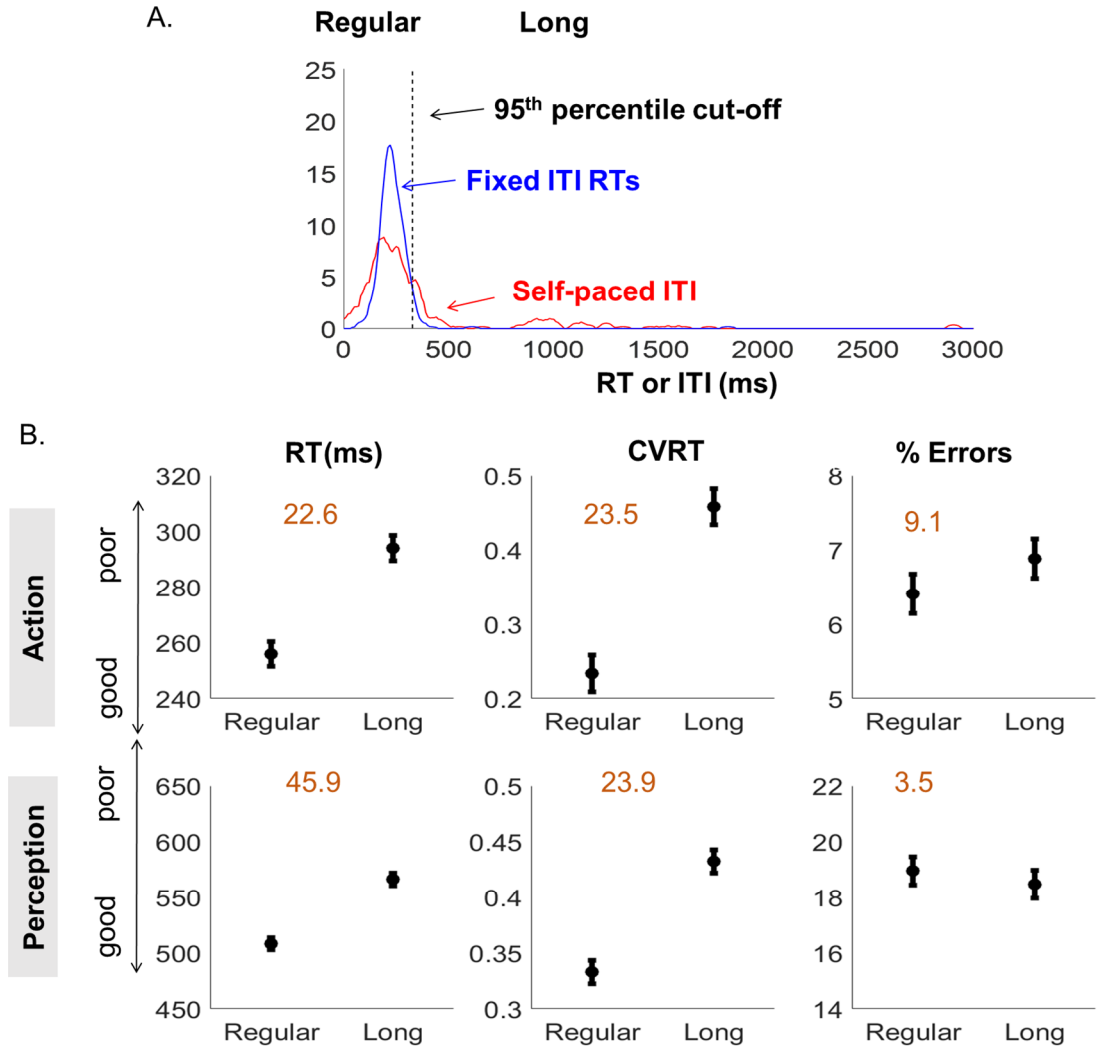


Figure 7. *Detrimental effect of long self-paced ITIs on performance and variability.* ***A)*** *Example from one participant of regular and long ITI-trials. Shown are the smoothed distribution of the self-paced ITIs (in red) and the distribution of the RT of the Fixed ITI condition (in blue). For each participant, the 95th percentile of the Fixed ITI RT distribution was calculated as a cut-off (black dotted line). Self-paced ITIs above the cut-off were deemed 'long', while ITIs below the cut-off were deemed 'regular'.* ***B)*** *Mean RT, CVRT, and % errors were calculated in the Self-paced*

condition for trials following regular and long ITIs. Orange Bayes' Factors indicate the strength of evidence against H1. None of the comparisons were in favour of H1. Error bars show the within-subject standard error.

Test 4. Control does not reduce temporal dependencies in RT series

The autocorrelations in the reaction time data were calculated separately for each participant and condition. Furthermore, the power spectrum was calculated over each reaction time series in R (R Core Team, 2013), following Wagenmakers et al. (2004). Although Wagenmakers et al. (2004) showed that the power spectrum for white noise of variance 1 is flat and null, this is not the case for white noise with the same variance as our experimental data, nor for series obtained from randomly shuffling our data. Instead, the spectrum of randomly shuffled RT series was positively correlated with the variance of that series – meaning that without correcting for this variance, potential differences between conditions could be due to variance rather than to actual temporal structures. Therefore, to correct for the power spectrum expected in our time series irrespective of any temporal dependency (our null hypothesis), the power spectrum was calculated 100 times on the randomly shuffled reaction time data, and the mean of these 100 spectra was subtracted from the unshuffled power spectrum. As such, the difference of these spectra reflects the time structure in the reaction time data. These difference-spectra were calculated separately for each participant and each condition.

Next, a linear regression line was fitted on the log of each power spectrum (still following Wagenmakers et al., 2004). Paired Bayesian t-tests were then conducted on the autocorrelations at the first lag and on the spectral slopes – to test if the self-paced condition differed from any of the three forced-paced conditions. Again, because *H1* is based specifically on a decrease in temporal dependency (and thus a flatter slope), t-tests were conducted one-sided.

First, we checked that our RT and ITI series actually showed clear temporal structure. As there was evidence for dependencies across the two measures (See Supplementary Table 1, p. 212), we carried on with contrasting these temporal dependencies across conditions.

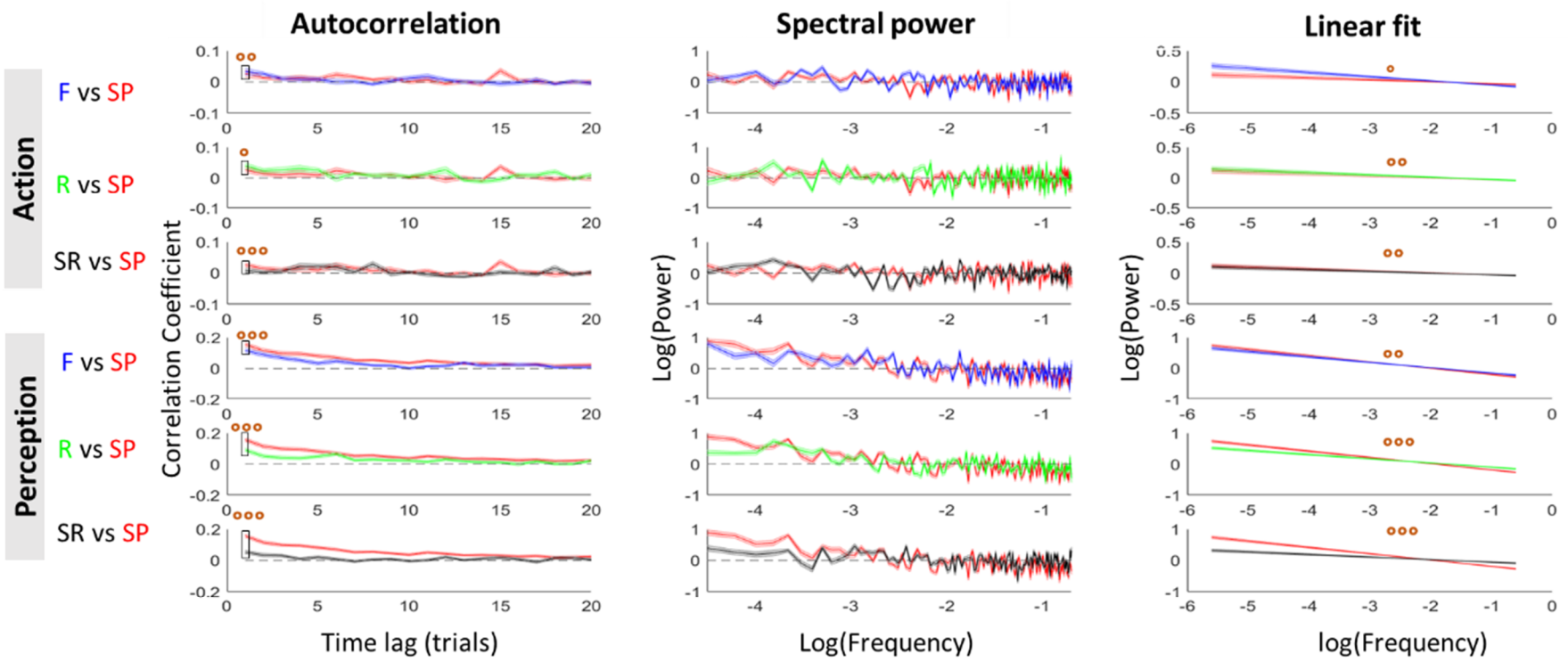


Figure 8. *Autocorrelation and spectral power and corresponding linear fit over the spectral power averaged across participants (same conventions as in Figures 4 and 5) for the RT, comparing the Self-paced with each of the forced-paced conditions.*

Table 6. *Bayes' Factors for Test 4, comparing temporal dependencies in the Self-paced versus each forced-paced condition, as reflected in the first point of the autocorrelation (AC) and the fitted slopes on the spectral power. Same conventions as Table 4 and 5.*

	Action-oriented		Perception-oriented	
Comparison	AC	slope	AC	slope
SP < F	.25	.62	.08	.11
SP < R	.37	.22	.05	.07
SP < SR	.08	.16	.04	.05

Figure 8 shows the mean autocorrelation functions and power spectra. Table 6 shows the Bayes' Factors associated with comparing each forced-paced condition to the Self-paced condition. Across both tasks, all comparisons provided evidence for *H2* over *H1* (showing no decrease in temporal dependencies in the Self-paced condition), though two were in the indeterminate range. Overall, our results suggest that control over trial initiation does not affect temporal dependencies.

No training effects in the Self-paced condition

Because the three fixed-paced conditions depended upon participants' own self-paced ITIs, the Self-paced condition always had to come first – making full counterbalancing impossible. While the potential effects of this are not straightforward, we conducted an extra analysis to test if participants were still learning the task in the Self-paced condition even after the training block. For each condition, the mean RT and accuracy were calculated for each participant on: 1) the first 30 trials (excluding the first trial), and 2) the rest of the trials. A Bayesian paired t-test was conducted to test if participants performed worse on the first set of trials than on the rest of the experiment (reflecting training effects).

No differences were found in either RT or accuracy in either task between the first 30 trials and the remaining trials in the Self-paced condition, (BF_{01} = 14.4, 23.3, 2.0 and 20.8 for RT in action-oriented task, accuracy in action-oriented task, RT in perception-oriented task and accuracy in perception-oriented task

respectively). It is thus unlikely that any outcome of the analyses from Experiment 2 could be ascribed to condition orders.

Testing for an effect of sleep

Relatedly, both tasks took place over two different sessions – that is, two conditions (one self-paced, one fixed-paced) on the first session, and the other two on the second session. As such, our set-up resembles that of a 'sleep experiment', in which participants are tested on a task of interest to assess the effect of sleep on learning and memory consolidation. Potential training effects may therefore particularly be found between the first and the second session. Additional analyses were conducted to check for this.

Scores for each measure (RT, % errors, CVRT) were collapsed within both sessions. Bayesian one-sided paired t-tests were conducted on each measure to examine if the score on the second session had improved compared to the first session – see Table 7 for the BF_{10}. On the perception-oriented task, there was evidence against an improvement. On the action-oriented task, there was only clear (moderate) evidence for an decrease in error % from session 1 (M = 6.5%, SD = 4.4) to session 2 (M = 5.3%, SD = 4.1). However, it appears this was not caused by an increase in drift rate ($BF_{S1<S2}$ = .07), nor by a decrease in non-decision times ($BF_{S1>S2}$ = .34). Evidence on whether the decrease was actually caused by an increase in caution was indeterminate ($BF_{S1<S2}$ = 1.25). Therefore, it remains unclear to what extent the decreased error rate reflects an actual performance improvement.

It should be noted that typical sleep studies are better controlled to pinpoint the effect of sleep as much as possible – e.g., by carefully selecting participants who have consistent and healthy sleep patterns, recording the sleep quality and the number of hours slept, and having a fixed delay across participants between different sessions. For the current study, the between-session measures contain much more noise, and any absences of effects are thus difficult to interpret.

Importantly though, these results do not affect our current conclusions. Even though the Self-paced condition always had to be completed first, the three fixed-paced conditions were counterbalanced across participants, meaning that none of

them could systematically affected by sleep. Any potential effect of sleep or order cannot explain the current result patterns (Figure 4-5).

Table 7. Bayes Factors between the mean scores for session 1 and session 2 – testing if performance was improved (i.e., reduced RT, CVRT, % errors) on the second session compared to the first.

Measure	Action BF_{10}	Perception BF_{10}
↓ **RT**	.12	.21
↓ **% errors**	6.47	.41
↓ **CVRT**	1.91	.04

Interim discussion 2

After reaching the same conclusions separately for Experiment 1 and 2, Bayes Factors of each statistical comparison between *H1* and *H2* in both experiments were summarised in Figure 9 with violin plots – distribution plots that show the entire range of Bayes Factors (y-axis) with horizontal thickness indicating density. Note that the Bayes Factors are logged for graphical purposes. The most-left (dark-blue) violin represents Experiment 1 plus Tests 1-4 of Experiment 2, excluding the EZ-parameter comparisons from Test 1b – showing an overall bias towards *H2*. While there are some BF that highly favour *H1*, these relate to comparisons that likely represent differences in speed-accuracy trade-offs, and do not reflect actual improvements in performance.

In the most-right violin (purple), the comparisons on RT, CVRT, and percentage correct have therefore been replaced by the comparisons between the parameters of the EZ-Diffusion model that relate to performance (drift rate and non-decision times from Test 1b, also seen separately in the middle violin). The comparisons on boundary separation are not included because they do not favour either hypothesis by default. Again, the overall results favour *H2*, showing evidence *against* a benefit of control.

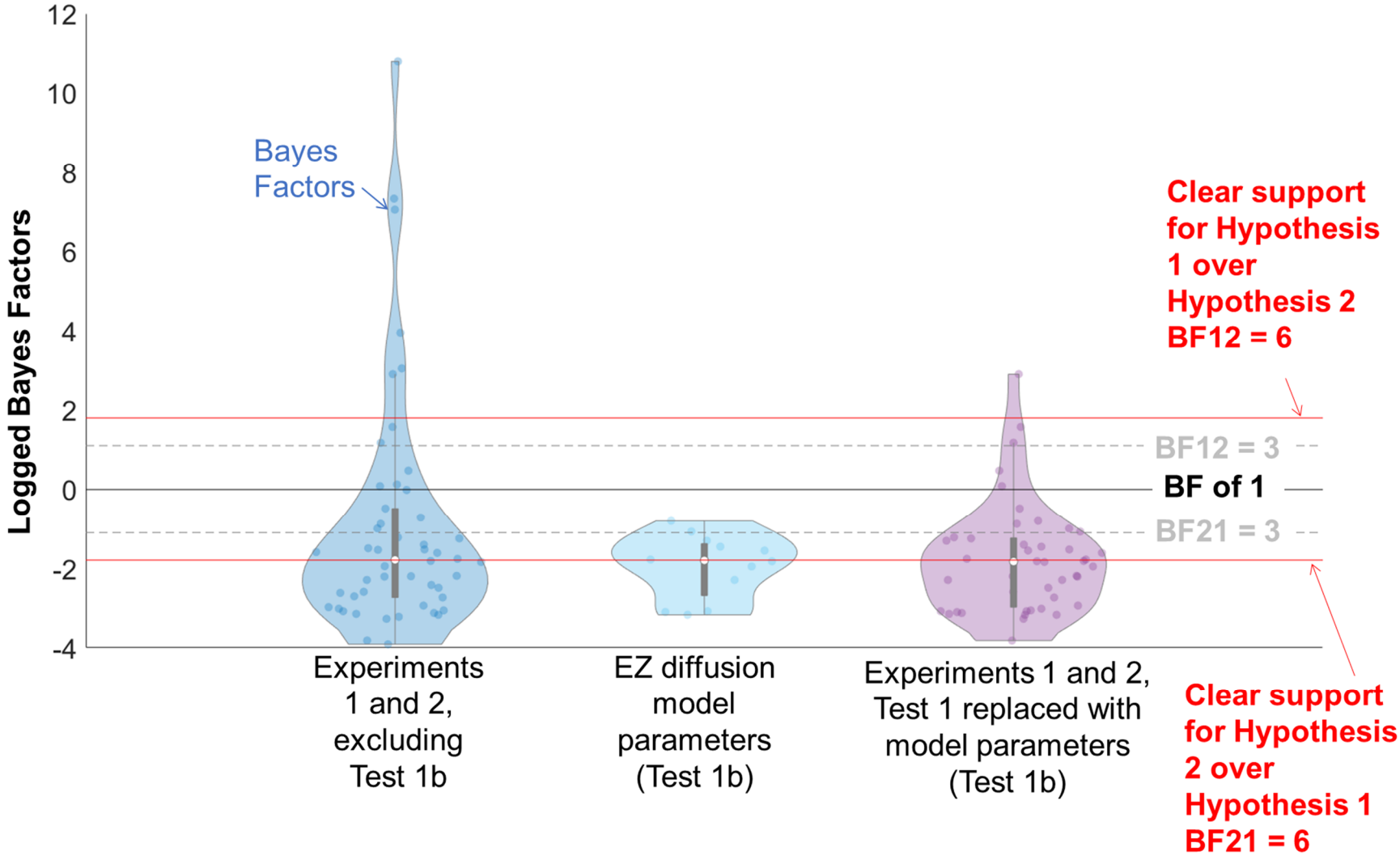

Logged Bayes Factors
12
10
8
6
4
2
0
-2
-4
Bayes Factors
Clear support for Hypothesis 1 over Hypothesis 2 BF12 = 6
BF12 = 3
BF of 1
BF21 = 3
Clear support for Hypothesis 2 over Hypothesis 1 BF21 = 6
Experiments 1 and 2, excluding Test 1b
EZ diffusion model parameters (Test 1b)
Experiments 1 and 2, Test 1 replaced with model parameters (Test 1b)

Figure 9. *Distribution of logged Bayes Factors from the statistical tests that compared Hypothesis 1 to Hypothesis 2, with each coloured dot representing one Bayes Factor, and each white dot representing the median of that distribution. Dots above the black line reflect higher support for Hypothesis 1, while dots below the black line reflect higher support for Hypothesis 2. The most left distribution (dark blue) encompasses the Bayes Factors from Experiments 1 and 2 (Test 1-4, excluding the EZ-model comparisons from Test 1b). In the right distribution (purple), the comparisons of RT, CVRT, and percentage correct (Test 1) have been replaced by the comparisons of the modelling on drift rate and non-decision times (Test 1b – shown separately in the middle graph). Overall, the distributions show our results favour Hypothesis 2 over Hypothesis 1.*

Note that for both experiments separately, data collection was continued until the median value of Bayes Factors that directly assessed Hypothesis 1 against Hypothesis 2 reached either 6 or 1/6. This approach was taken as both experiments featured multiple analyses, that cannot be interpreted independently from each other (such as mean performance and variability, or drift rates and non-decision times). For Experiment 2, the median value of the most-right distribution was used as a criterion for stopping recruitment (excluding the values from Experiment 1), with the final median BF_{21} being 6.6.

Characterising the self-paced ITIs in the computer-based tasks

Rationale

The results from Experiment 1 and Experiment 2 show that performance did not improve when having control – implying that participants cannot access their internal states, or alternatively, that they have some form of access but no means or will to act upon it. While we cannot fully rule out either possibility, we can have a closer look at *how* participants behaved when given control. Because Experiment 2 allows for the recording of the self-paced ITIs, it provides an opportunity to examine these ITIs in more detail – to see what potential strategies participants may have used in

handling the control they were given. Although the control did not benefit participants, their ITIs may still show characteristics that diverge from regular RT characteristics. To get more insight into these strategies, we examined three different measures in the self-paced ITIs: Variability, temporal dependencies, and post-error slowing.

Variability in the ITI

If participants use the control in the self-paced condition and do not continue to the next trial when they do not feel ready, one would expect the distributions of the self-paced ITIs to be different from typical responses to a stimulus. Specifically, if participants make use of the control, they should show a mixture of shorter and longer ITIs – which subsequently leads to high variability. On the other hand, if participants just start the trials as soon as the stimulus inviting them to do so appears, their ITIs should resemble simple RTs to a single salient and predictable stimulus onset (the fixation cross). We did not have such data from our participants but the Fixed condition from the action task offered the closest comparison. If participants were just eager to carry on through the task as quickly as possible, the coefficient of variation of their ITIs (CVITI) across both tasks should be similar to the CVRT from the Fixed condition in the action task, or even smaller, because it is a one-alternative decision, while the RT is based on a two-alternative decision.

Temporal dependency in the ITI

As mentioned in the introduction, we expect the self-paced ITIs to show temporal dependencies. Because participants were instructed to wait for every trial until they felt ready for it, their ITIs may show higher temporal dependencies than typical RTs – possibly reflecting stronger coupling to fluctuating internal states than stimulus-driven responses (the trial itself), even if these attempts did not result in better performance. To examine this, the autocorrelations and power spectra were calculated for the self-paced it is. Again, for both tasks, autocorrelations and fitted lines were compared against the Fixed condition of the action-oriented condition.

Post-error slowing in the ITI

There is a large literature showing that people are able to slow down when they see or are explicitly told that they made an error (post-error slowing*;* Rabbitt, 1966) – seemingly because of an adjustment of response caution (Dutilh et al., 2012a). When participants are making an error, they are faced with objective information that their performance-relevant internal state – and thus their decision to continue to the next trial – was suboptimal. If participants were able to make maximum use of the control based on their inner states, they could have prevented these errors from happening altogether, especially in the action task, which is very easy. However, since they were not able to use the control in this manner, they may instead slow down afterwards – resulting in post-error slowing in the ITI. This may at least indicate that our participants cared enough to adjust their behaviour in response to poor performance, even if this was ineffective in boosting their performance.

Results

ITIs show higher variability than RT

Mean ITI ranged from 243 to 1742 in the action-oriented task and from 298 to 2605 in the perception-oriented task. Similarly, to the RT data, the ITI data was log transformed as a first step, to correct for the high skew of the distributions. Figure 10A shows the distributions of the CVITI for both tasks compared with the CVRT of the Fixed condition of the action-oriented task, with accompanying Bayes Factors for the associated Paired one-sided t-tests. On both tasks, we found extreme evidence that the CVITI was much higher than the CVRT – showing that the self-paced ITIs are more variable than would be expected if they were just response times to a stimulus. This suggests that participants were using the ITI in some manner, but this did not help them to improve their subsequent performance.

ITIs may show some higher temporal dependencies than RT

For both tasks, autocorrelations and power spectra plus their fit lines were calculated on the ITIs for each participant, using the same procedure as in Test 2 in above. Bayesian Paired one-sided t-tests were conducted on the autocorrelations at lag one and on the spectral slopes – to test if the temporal dependency was higher in the ITI compared to the RT of each condition. Similarly, to Test 4 above, we first confirmed that the ITI actually contained temporal dependencies (see Supplementary Table 1). As we found evidence for this on both tasks, we then carried on with comparing the ITI to the RT.

Figure 10B shows the mean autocorrelation functions and power spectra of the ITIs from both tasks, compared to the RT of the Fixed condition of the action-oriented task. On both tasks, there was no evidence for higher temporal dependencies in the ITI compared to the RT.

Post-error slowing in the self-paced ITIs

Post-error slowing in the self-paced ITIs was calculated using the method of Dutilh et al. (2012b). To avoid unstable means due to a low number of observations, participants who made less than ten errors were excluded. For the remaining 23 participants, mean ITIs were calculated on the logged ITIs before and after each error. Bayesian paired one-sided t-tests were performed to test if post-error ITIs were on average slower than pre-error ITIs. Because participants were not given feedback throughout the main tasks, post-error slowing was only calculated for the action-oriented task, in which participants typically know when they have made an error – as opposed to the perception-oriented task, in which participants are often unsure of the correct answer.

Participants were on average 159 ms slower in their ITI after making an error (Figure 10C – analysis run on logged values) compared to just before making this error. Such difference could have two possible origins though: 1) errors may lead to ITIs larger than average on the next trial, indicative that participants have adjusted their ITI as a consequence of the error (actual post-error slowing), or 2) errors could be typically preceded by shorter ITIs and followed by regular ITIs, simply reflecting a regression to the mean. Comparing the mean pre- and post-error ITI with the

overall mean ITI shows clear support for option 1 (Bayesian one-sided paired t-test on logged values, $BF10_{post>mean} = 37.0$) and not for option 2 ($BF10_{pre<mean} = 1.7$).

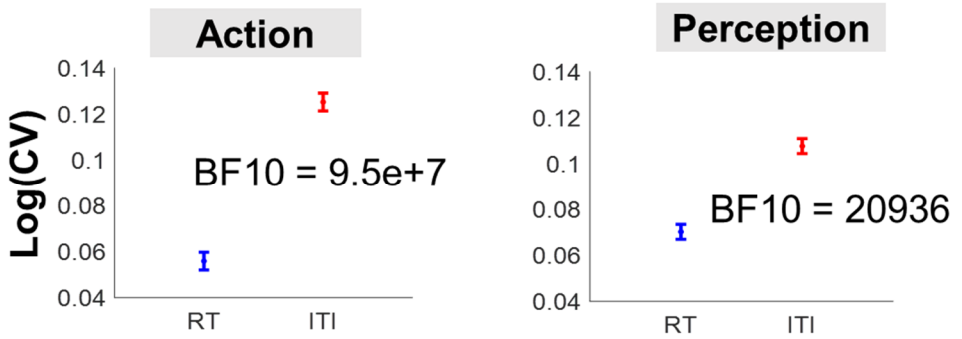
It therefore seems that our participants were able to adjust their behaviour in response to *objective* evidence that their performance was poor (see section *Motivation* in the Discussion for more discussion on this), which is interesting for two reasons. First, this contrasts with their inability to adjust their ITIs to prevent errors from occurring, i.e. presumably in response to internally-driven information that they are in a state detrimental to performance. Second, it suggests they were sufficiently motivated to act upon their performance, which is a prerequisite for the control manipulation to be relevant.

The presence of post-error slowing could suggest that participants were able and willing to make some use of the control when faced with objective information on their performance. For this to lead to improved performance though, post-error slowing on trial *n* should also result in improved performance (i.e. lower RT and higher accuracy) on trial *n+1*, as focus has suddenly gone up. Unfortunately, neither of the current tasks are suited to examine this prediction, because post-error improvements in accuracy cannot be estimated properly (Danielmeier & Ullsperger, 2011): The action-oriented task has too few errors, leading to an unreliable estimate, and the perception-oriented task contains errors of which participants are not aware, which should not lead to subsequent post-error adjustments.

However, while the ITI could potentially absorb the slowing typically seen in RT, this may not necessarily lead to improved focus. If anything, the results of Test 3 above suggest that slowing down does not necessarily improve subsequent performance. Indeed, previous literature has shown that, while post-error slowing is often seen as a strategic adjustment aimed at improving subsequent performance, post-error slowing and post-error improvement in accuracy are not necessarily found together (see Danielmeier & Ullsperger, 2011 for a review). One possible reason could be that post-error slowing partly reflects an automatic response to rare events, similar to startling in the rodent literature (Wessel & Aron, 2017), rather than a purely strategic adjustment. The observed post-error slowing in the ITIs may as such reflect a mixture of automatic responses and top-down strategies to try to refocus on the task.

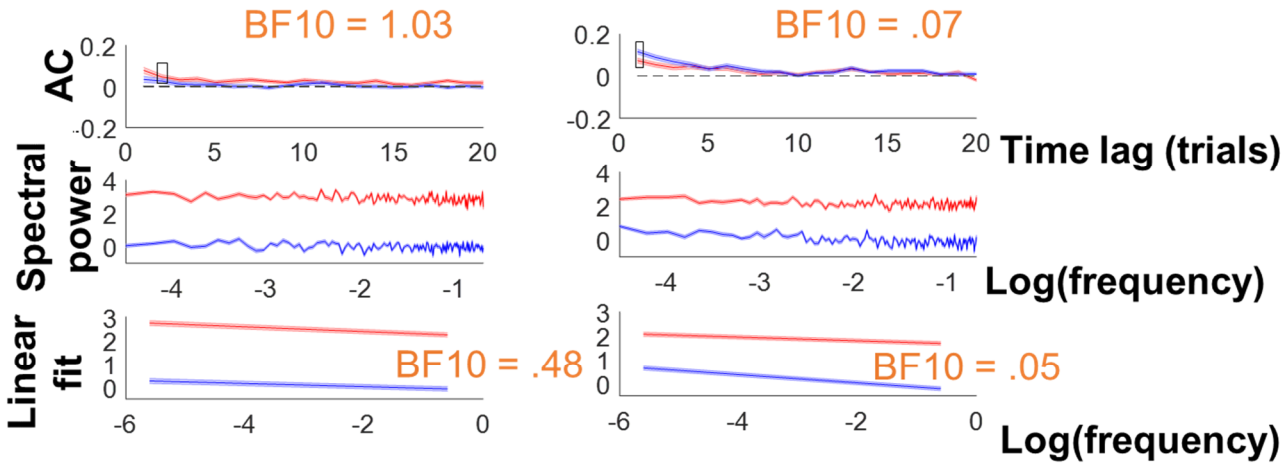
A.
Measure 1

Evidence for increased variability in self-paced ITIs compared to Fixed RT



B.
Measure 2

Evidence against increased autocorrelations and spectral slopes in self-paced ITIs compared to Fixed ITI RTs



C.
Measure 3

Evidence for post-error slowing in self-paced ITIs

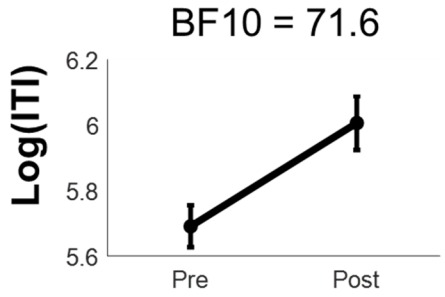


Figure 10. *Evidence across the three measures of Part 2. Measure 1 reflects the coefficient of variance for the log of the ITI (CVITI) on both tasks, compared to the coefficient of variance for the log of the RT (CVRT) of the Fixed condition on the action-oriented task. Measure 2 reflects the temporal dependency of the ITI, as measured by the autocorrelation and the fitted slope on the spectral power, compared to that of the RT of the Fixed condition in the action-oriented task. The ITI showed much higher variability than the RT, but did not show higher autocorrelations or steeper slopes. Measure 3 reflects post-error slowing found in the ITIs. Data points show the logged average self-paced ITIs in the action-oriented task before (pre) and after (post) an error, indicating that participants slowdown in their ITI after an incorrect trial. Error bars on all panels show the within-subject standard error.*

Individual differences

We noted that self-paced ITIs showed large individual differences and wondered if these could provide a key to why the control appeared useful to some participants and detrimental to others, resulting in no overall improvement. However, additional between-subject analyses did not reveal any clear links between the three ITI-characteristics (variability, temporal dependency, and post-error slowing) and the improvement in performance between the Self-paced condition and each of the forced-paced conditions. When instead looking at mean ITI, there was a consistent *negative* relationship with the improvement in performance across all three forced-paced condition: Participants who had a *shorter* mean ITI showed more improvement. As our within-subject analysis showed that longer ITIs may be markers of an overall poor mode of responding – they are followed by poorer rather than better performance, both findings could reflect that good participants indeed show less of these poor modes.

General Discussion

No improved performance or reduced variability with control

Assuming that task performance is under the influence of some internal states varying over time, we aimed to test whether people have direct access to these internal states and can use this information to improve task performance. We gave participants control over the timing of three behavioural tasks and compared their performance with conditions without such control. In all three tasks, we found that participants did not perform better when provided with control (see Figure 9 for an overview), even when questionnaires indicated high intrinsic motivation to perform the task. Furthermore, when participants took longer delays during the task, this was associated with poorer, not better, subsequent performance and increased variability. Control also did not affect temporal structures in the reaction times

When examining the time taken to move from one trial to the next in the self-paced condition (ITI), it is clear that participants do not simply rush through the task as quickly as possible. Rather, their ITIs are slower and show much higher variability than their speeded RT, as well as clear evidence of post-error slowing. As such, participants using the control in some way beyond simply and automatically responding to a fixation dot as fast as possible. Importantly, even though they appear to do 'something' with the control, it did not help them improve performance – suggesting that access to internal states is minimal at best.

Access to internal state: either limited or not directly useful

Overall, our results show that participants were not able to use the control to improve their performance and reduce their variability – suggesting that if people have some access to their performance-relevant inner states at all, this access is minimal and may not be used to noticeably improve upcoming performance. One reason why access to current performance-related states may be of little use for improving upcoming performance (500 to 1000 ms later) could be down to the difficulty of predicting future internal states from current ones. Although neural correlates of upcoming performance have been identified, these are typically very short-term and their predictive power is very low (see section "Biological underpinnings of variability

and performance" below). Although this limited predictive power could be down to technical limitations, we cannot exclude that future performance is to a large extent non-deterministic and therefore largely unpredictable even from within. A conservative interpretation of our results may therefore be that we do have some access to our performance-related internal states, but this access is 1) very limited, 2) rarely spontaneous, and therefore 3) mostly irrelevant to improving future performance.

At first, this interpretation may seem at odds with existing literature on mind wandering, which assumes people can access at least some aspects of their internal fluctuating states. However, our conservative interpretation may link with this literature in a couple of ways. First, limited access would explain why the link between behavioural performance or variability and probe-caught subjective reports of mind wandering is robust but weak. For example, over five different samples, participants who reported being fully mentally 'zoned out' from the task only showed an increase of ~3-7% in variability compared to when they were fully on task (Seli et al., 2013, Laflamme et al., 2018).

Secondly, its lack of spontaneity would match the differences between results from 'self-caught' and 'probe-caught' methods in the study of mind wandering (see Weinstein, 2017 for a review). Self-caught methods rely on the participant to report each time they are aware they are mind-wandering (and would therefore only be able to catch shallow stages of mind wandering – 'tuning out'), whereas probe-caught methods probe participants about their thoughts just prior the probe (which is, as such, always a 'post-hoc' judgement), usually at pseudo-random times during the task (and should therefore be able to catch both 'tuning out' and 'zoning out'). The self-caught method is generally not preferred, because participants often do not catch their own deteriorated states of performance (Franklin, Smallwood & Schooler, 2011; Schooler et al., 2004). Within the mind wandering literature, this inability to self-catch mind wandering has been explained by a reduction of 'meta-awareness' – such that if one is mind wandering, and performance is reduced due to a loss of attentional resources, one's meta-awareness of the mind wandering and deteriorated performance is also reduced. Indeed, assuming that unaware stages of mind wandering always follow sequentially from aware stages (Cheyne et al., 2009; Mittner et al., 2016), only limited spontaneous access during the shallow stage can explain why more severe stages happen at all, rather than being caught *before*

the episode gets more severe. Although mind wandering is a mental state and therefore requires some form of awareness (see Introduction), in these cases, the awareness may be 'post-hoc'. This inability would then relate to our third point: that our (marginal) access may be no help in improving future performance.

To draw a parallel between our findings and the mind wandering literature, prompting participants would be somewhat similar to the post-error slowing reported in the present study. Similarly, participants are able to access their task-unrelated thoughts when prompted to do so by the experimenter. In contrast, it may be much harder to spontaneously detect mind wandering and other unfavourable states, as would have been required in Experiment 2 in order to use the control when available to prevent errors and very long RT from occurring in the near future.

However, our findings are also theoretically consistent with another (more drastic) interpretation: That we do not have any access to our performance-related inner states. The correlations between behavioural variability and subjective reports of mind wandering could be caused by a third variable that underlies both, but this variable may be fully opaque to us. As often suggested in this literature, such internal states could be related to the activation in the default mode network (Christoff, Gordon, Smallwood, Smith & Schooler, 2009; Mason et al., 2007) or in task-related networks (such as the dorsal attention network; Corbetta, Patel & Shulman, 2008), or the anticorrelation between them (Kelly, Uddin, Biswal, Castellanos & Milham, 2008). As such, behavioural variability or poor performance may not be a direct consequence of mind wandering, but both would likely co-occur in time. Likewise, good performance may co-occur (more often than not) with task-related thoughts or the feeling of being ready, which would also lead to positive associations between subjective reports and behaviour.

The idea that mind wandering may not directly cause poor performance appears at first to contradict previous accounts (e.g. "mind wandering influences ongoing primary-task performance", Laflamme et al. 2018, p.1). However, such accounts may reflect the functional processes that underlie the construct of mind wandering. Previous studies have suggested that mind wandering contains a high proportion of self-oriented thoughts, and seems to play a role in future and autobiographical planning (Baird, Smallwood & Schooler, 2011; D'Argembeau, Renaud & van der Linden, 2011). In this light, mind wandering has been described as a somewhat economical phenomenon: Since the task does not seem to require

a high amount of ‘mental/cognitive resources’, they may instead be used to solve self-oriented problems in the meantime. This description is consistent with the findings of Ward & Wegner (2013), who contrasted the construct of mind wandering to that of ‘mind blanking’ (referring to a mental state in which one is void of all thoughts, both task-related and task-unrelated; also see Robison et al., 2019; Van den Driessche et al., 2017). They concluded participants seemed to find it easier to catch their own mind blanking compared to their mind wandering – and they suggested that because mind wandering may have beneficial components, it is not always necessary to ‘snap out of it’.

To summarise, our findings are consistent with both conservative (people have some access to their internal fluctuating states, but very marginally and typically irrelevant) and extreme interpretations (people have no access to their internal fluctuating states at all). While these interpretations seem at odds with common assumptions, both accounts are reconcilable with current literature.

Motivation

It also remains possible that people do have access to their internal states, but that our participants did not use this access due to a lack of ability or willingness. If so, the most apparent explanation for our results could be a lack of motivation. If their motivation was limited to going through the experiment as fast or effortlessly as possible, our participants may not have had the will to access performance-relevant information in order to improve their performance. Although we cannot reject that motivation played some role in our results, this interpretation is unlikely to explain our data. In Experiment 1, participants reported high levels of internal motivation (or were otherwise excluded). Furthermore, good performance increased chances of monetary and social rewards. In this context, it makes sense that participants, given they have access to their own internal states, act upon this access. In Experiment 2, although the task was more boring, our participants were mostly postgraduate students who were highly familiar with psychological testing – and as such, expected to show high intrinsic motivation. Moreover, if participants have access but are not acting upon it, it is likely to be reflected in fast and automatic use of the ITI. Our results show the opposite: Not only were their ITIs twice more variable than would

be expected if they simply initiated the next trial as swiftly as possible, but they also largely slowed down following an error (around 31% increase in ITI on average).

In Experiment 2, we also found that strategies in the use of ITI (as measured by its characteristics) did not correlate to improvements in the self- compared to the forced paced conditions on between-subject level. For Experiment 1, such data is unfortunately not available as there was no straightforward way to measure the self-pacing in darts throwing. Instead, we correlated the reported intrinsic motivation scores to the difference in score and CV between Self- and Fixed-paced on the last block. Similarly, no correlations were found.

Importantly, our conclusions did not depend on the task being boring or engaging, as both showed similar results. The absence of a benefit of control thus does not seem to rely on motivation. Overall, in neither experiment we can fully rule out the possibility that participants do have access but just do not act upon them. However, until such access is experimentally proven in some context, the most parsimonious conclusion is that they have none.

Changes in performance versus changes in strategy

To help us resolve ambiguities in the outcome of two of our empirical tests, EZ-diffusion model parameters were calculated for each condition (Wagenmakers et al., 2007). The first ambiguity regarded the interpretation of co-occurring RT increase and % error decrease in two of the forced-paced conditions (Replay and Shuffled Replay) in the action-oriented task. EZ-Diffusion model attributed this difference to higher boundary separation in these conditions, compared to the Self-paced and the Fixed conditions. While a change in drift rate is commonly interpreted as a change in processing efficiency, changes in boundary separation are typically interpreted as strategic changes in caution, i.e. speed-accuracy trade-off, leading us to conclude that there was no improvement in performance in the Self-paced condition. Below we discuss two counterarguments to this interpretation.

First, one could argue that the reduced boundary separations in Self-paced compared to Replay or Shuffled Replay still reflect an active effort of participants to change performance, even if it did not result in a ‘true improvement’. However, if anything, it seems more likely that active effort to change behaviour will lead to

increased boundary separation instead, as participants would be more aware of their accuracy than their speed (especially on a milliseconds scale) – i.e., it seems more concrete to aim at ‘zero errors’ than at ‘reducing speed by 50 ms’. In contrast, our results showed that boundary separation was lower in the Self-paced condition. Furthermore, this was very similar to the Fixed condition, and therefore any adjustment is not specific to the Self-paced condition.

Second, we used a simplified version of the drift diffusion model, as it has been shown to be more powerful in detecting effects in drift rate and boundary separation than more complex variants (van Ravenzwaaij, Donkin & Vandekerchove, 2017; van Ravenzwaaij & Oberauer, 2009). Importantly though, this variant does not include a parameter capturing variability in drift rates across trials. Both drift rate and drift rate variability have been associated with self-reported mind wandering, but variability was a stronger predictor (McVay & Kane, 2012, though they used a linear-ballistic accumulator model; LBA). This makes sense when considering that variability is captured by both very long RT and very short RT – the combination of which has a larger effect on variance than on mean performance. A simulation study by van Ravenzwaaij & Oberauer (2009) reported that when data are generated from the LBA and fit with the EZ-diffusion model, increased drift rate variability in the generating (LBA) model is negatively correlated with EZ’s drift rate estimates, and positively correlated with EZ’s estimates of boundary separation. One could question then if the decreased boundary separation in Self-paced actually reflects participants using control to reduce variability in their processing rates. However, data generation with a drift diffusion model shows that such a reduction in drift rate variability would lead to reduced error rates in the Self-paced relative to the Replay and Shuffled Replay conditions, whereas our results show an increase. As such, our results are more consistent with a change in boundary separation.

Altogether, the reported differences between Self-paced and Fixed versus Replay and Shuffled Replay are more likely caused by predictability of target onset than by control (see Table 2 for an overview). As such, knowing the time of trial onset seems to lead to a less cautious response pattern (lower decision threshold), in which speed is emphasised over accuracy, with no overall change in processing

efficiency. These patterns have been found previously (Miller, Sproesser & Ulrich, 2008).

Still, the presence of different speed-accuracy trade-offs may make the interpretation of the results less straightforward. Ideally, the Self-paced ITIs should be compared to ITIs that are both variable *and* predictable – but these two are mutually exclusive in a forced-paced condition. As the Self-paced condition is a unique combination of different ITI-features (see Table 2 for an overview), each feature has to be compared separately. On the one hand, this makes interpretation more difficult, as it may be expected that participants act differently in different conditions. On the other hand, the inclusion of multiple conditions remains interesting, as they all allow for different comparisons. Importantly, the central question in the current research is whether participants act differently in the Self-paced condition in a way that is systemically beneficial for their performance – with the current results contradicting this notion. Nonetheless, we may interpret the comparisons with the Replay and Shuffled Replay conditions with more caution, and focus solely on the predictable Fixed condition (which is most similar to Experiment 1). Recalculating the median value of BF_{21} from Experiment using only the Fixed vs. Self-paced comparisons gives a value of 9.9 –showing even stronger evidence against an improvement in the Self-paced condition than the median value over all conditions.

Changes in non-decision times

The second ambiguity in our results was the smaller number of very short RT and the subsequent decreased variability in the Self-paced perception-oriented task. The EZ-model suggests that, in the perception-oriented task only, non-decision times were higher for self-paced compared to the three control conditions, suggesting slower sensory or motor processes (Wagenmakers et al., 2007). It is unlikely that motor output time would be the cause, as this would be expected to be the same across tasks and conditions. One possibility is that the additional action in the self-paced condition interfered with the sensory processes, but not enough to give hindrance when the stimuli are easy to see.

In any case, unlike in Experiment 2, Experiment 1 did not have an additional action in the self-paced condition. Instead, the Fixed-paced condition in Experiment 1 featured tones, while the Self-paced condition did not. Despite these differences, we found the same results over the experiments; in Experiment 1, the forced-paced condition featured an 'additional stimulus', while in Experiment 2, the Self-paced condition featured an 'additional action', but results always favoured Hypothesis 2 over Hypothesis 1.

Neural 'quenching' in darts experiment

More specifically, it is possible that the tones in the Fixed condition reduced participants' internal variability. Previous studies have found membrane potential and firing rates in animals (Churchland et al., 2006; 2010), and electro- and magnetoencephalography, electrocorticography and fMRI signals in humans (Arazi, Censor & Dinstein; Arazi, Gonen-Yaacovi & Dinstein, 2017; He, 2013; He & Zempel, 2013; Schurger, Sarigiannidis, Naccache, Sitt, & Dehaene, 2015) are reduced after the presentation of external stimuli (though the large majority of this work has focused on visual stimuli). The magnitude of these reductions has been linked with increased performance both across trials and across individuals. Our findings could therefore be explained by two independent processes: in the Self-paced conditions, participants benefit from the control, while in the Fixed condition, participants benefit from the neural variability reductions to the tones – resulting in a lack of differences. However, these reductions typically appear 100-400 ms after target onset. In our experiment, participants were trained to throw *on* the tone, rather than as a response to it – meaning their action is performed before the neural reductions would occur. The previous ('steady-') tone was played 1500 ms prior – and to our knowledge, there is no evidence for neural reductions at this longer delay. Therefore, this alternative explanation remains quite speculative. As for now, it is more parsimonious to assume participants did not benefit from the control, rather than assume two independent but simultaneous processes.

Routines and practice in sports psychology

After finishing the darts game in Experiment 1, our participants were informally asked if they used any strategies in the Self-paced condition. Many of the participants reported that they threw *"When I felt like it"* or *"When I felt ready for it"*. However, in light of our results, the behavioural relevance of this feeling remains unclear.

While the idea of 'mentally feeling ready to perform the action' may seem intuitive to laymen, the area of sports psychology has mainly focused instead on (pre-performance/during-performance) routines as well as the consistency of these routines as means to improve performance (for reviews, see for example: Cohn, 1990; Singer, 2002; Cotterill, 2010). The key element of these routines is training automaticity as a way to enhance task attention and to decrease focus to external distractions.

It should be noted that this 'automaticity' may still entail the involvement of cognitive processes, and that sportsmen require mental flexibility for their performance (see Toner, Montero & Moran, 2015 for a review). However, the focus of these cognitive processes does not appear to be on one's own inner state, but rather with handling relevant environmental states (e.g., the amount and direction of wind when striking with a golf club) or with improving flows in their movement (while aiming to improve one's skill level). As such, internally-driven variability in sports is often described from a perspective of sensorimotor control; resulting from sources as movement timing and trajectory (e.g., Smeets et al., 2002) rather than from attentional fluctuations. Previous research has found benefits of 'external foci of attention' (e.g., focusing on the darts board) over 'internal foci of attention' (e.g., focusing on one's own movement of the arm), both on performance as well as on pre-performance (neuro)physiological states (Marchant, Clough & Crawshaw, 2007; Marchant, Clough, Crawshaw & Levy, 2009; Neumann & Piercy, 2013; Radlo, Steinberg, Singer, Barba & Melnikov, 2002).

In other words, while sports psychology has an interest in reducing variability and creating the most optimal pre-performance state, their interest does not seem to lie in 'reading inner states', but rather in training repetitive, automatic, and externally-focused states. Compared to Experiment 1, this type of training would be

more similar to the Forced-paced than the Self-paced condition. Interestingly, these 'repetitive states' also appear in other aspects of training. Within sports literature, emphasis is put on the consistency of optimal physical movements (for example, consistency in throwing in darts, Brenner, van Dam, Berkhout & Smeets, 2012; Smeets et al., 2002, or in golf, see Langdown, Bridge & Li, 2012 for a review). It is possible that people do have access to their internal states, and are able to adjust these states proactively during repetitive conditions, but to a similar extent in the self- and forced-paced conditions. This hypothesis could explain the effectiveness of training automatic states in the literature, as well as the lack of differences between self-paced and predictable forced-paced conditions in both Experiment 1 and 2. However, such monitoring system may presumably require cognitive resources and still occasionally fail in a forced-paced condition. We may therefore still expect increased performance in self-paced conditions – which the current results do not confirm.

One may wonder if the skill level in darts of our participants played a role in Experiment 1. Only one of our subjects reported playing darts about twice a week, while all the other subjects had played it a few times a year or less. Therefore, it was not possible to test the effect of skill level in our data (though the scores of the experienced participant did not seem to display a diverging pattern). However, it is important to note that, if anything, the largest numerical differences between the Self-paced and Forced-paced conditions took place in the first block, when participants may still be getting used to the rhythm of the Forced-paced condition. At the later blocks (especially block 4 and 5), the conditions are most similar, suggesting that practice makes the conditions more similar, not less.

Of course, it is still possible that professional darts players *would* be able to use the control effectively. Some empirical data can be found on the website *FiveThirtyEight*[5] on baseball scores from professional sports matches: In 2015, the Major League Baseball implemented a new rule that resulted in a shortened delay between pitches (Arthur, 2015). The effect of this rule change had varying effects – some players' performance was not affected at all, but other (mostly older) players

[5] FiveThirtyEight is a website with blog posts from data analysists, mostly focusing on politics, opinion poll aggregation, and sports data. Though the blog posts are written by professionals, it should be noted that these sources are not peer-reviewed.

performed worse after implementation. In 2016, the rule was abolished, and the older players' performance improved again (Arthur, 2016).

However, Arthur (2015) notes that in general, there is no correlation between pace and performance ($r = -.025$) and suspects the deteriorated performance was likely not caused by the changes in pace. Furthermore, even if a change in pace caused the worsened performance, it is difficult to pinpoint the driving factor – as the rule change does not constitute a well-controlled manipulation. Related to the paragraphs above, the players likely performed worse because their routines had to be altered – which would related to training automatic, repetitive, fixed states, as opposed to training to 'wait' until one is 'ready' for each pitch. As older players would likely have a more established routine, it makes sense that they are affected the most. Overall though, the baseball-data nor our current results can give a definitive answer to the question whether professionals benefit from self-pacing – and it would be interesting to replicate the current study with this population.

Training access to internal states?

The possible influence of skills and practice on performance leads us to a larger question: To what extent is it possible to train access to our own internal states? One field of research relevant here is mindfulness (meditation) training. Within this literature, people may be trained to be more mindful of their internal states – and as such, may be trained to improve their attention and performance (Brown & Ryan, 2003; Wells, 2005; Zeidan et al., 2010) and "tame mind wandering" (Morrison et al., 2014; Mrazek et al., 2013). However, reported effects tend to be moderate – for instance, Morrison et al., 2014 reported a reduction of ~8.5% in variability after a seven-hour training over seven weeks. Like attention and mind wandering, mindfulness is a very broad concept (Bergomi, Tschacher & Kupper, 2013) and could refer to a multitude of mechanisms. Furthermore, mindfulness is difficult to capture in an experimental study set-up (for instance, when picking participants or when designing a control condition). Outside the mindfulness/meditation literature, Baldwin et al. (2017) found an increase in participants' own awareness of mind wandering over the course of a five-day experiment. However, due to the highly

repetitive nature of the task, it is plausible that participants just allowed themselves to deliberately mind wander more throughout the sessions.

Temporal dependency

Across both experiments and both measures of temporal dependency, we did not find any evidence of a reduction in the self-paced compared to forced-paced conditions. This test is weak in the action-oriented task, as the evidence for the presence of structure in our RT series was often low, making any reduction hard to find. In contrast, in the perception-oriented task, we did find clear evidence for temporal structures on all RT series, along with strong evidence against a reduction in Self-paced.

Previously, Kelly et al. (2001) found reduced temporal dependency in their self-paced compared to forced-paced conditions – though the pacing here refers to the response time rather than the ITI, and is therefore not directly comparable with the current conditions. Interestingly, they mention that *"self-pacing means that the system is sampled at irregular intervals in real time, violating the assumptions of most dynamical analyses"* (p.824), and propose that conditions with fixed pacing could be better suited for measuring temporal dependency. In our action-perception task, the Fixed condition was indeed the only condition that showed clear evidence for a long-term slope structure (see Supplementary Table) – though this was not replicated in the perception-oriented task.

The temporal dependency in the RT and ITI may suggest that these are coupled to underlying fluctuating states. One commonly mentioned state is 'attention' (e.g., Irmisscher et al., 2018), which is also thought to fluctuate over time and to influence performance. However, Wagenmakers et al. (2004) noted that it is unclear *how* attention would cause the specific temporal patterns common in empirical data. Alternatively, temporal dependencies could be caused by the combination of a number of different processes with varying timescales. This is in line with findings that variability is underpinned by a number of biological processes, all with varying time scales (see below section on *Biological underpinnings of variability and performance*).

Because the mechanisms underlying temporal structure are still largely unclear, it also makes it more difficult to interpret any differences or lack of difference between conditions. In the current research, within the framework of *H1,* we assumed people would benefit from the control by mitigating against these temporally-fluctuating states – leading to increased performance and reduced temporal dependency. However, it would have been possible that participants use the control to mitigate against short-range ('moment-to-moment') fluctuations – in this case, the effect on temporal dependency would be less clear. Still, giving that we did not find a benefit of control on any of the other statistical tests either, we think for now the most straightforward interpretation is that participants cannot mitigate against their internal states at all.

Biological underpinnings of variability and performance

Above, we referred to internal states that may underlie both behavioural performance and variability. These may be reflected in fluctuations in the DMN, task-related networks, and the episodic memory network, which have been often associated with mind wandering (for a meta-analysis, see Fox et al., 2015; for reviews, see Christoff, 2012; Smallwood et al., 2012). Indeed, slow rhythms (~.05 Hz) in BOLD activity within the DMN have been associated with reaction time variability (Weissman et al., 2006). Variability in performance on detection or discrimination tasks has been related to oscillatory activity using EEG or MEG, in particular alpha (8-12Hz) power and phase, but also to beta and gamma power (Busch et al., 2009; van Dijk et al., 2008; Drewes & VanRullen, 2011; Ergenoglu et al., 2004; de Graaf et al., 2015; Hanslmayr et al., 2007; Rihs et al., 2007; Romei et al., 2008; 2010; Thut et al., 2006; VanRullen et al., 2011; Bompas et al. 2015). Interestingly, a recent study has shown that spontaneous fluctuations in alpha rhythms are partially locked to slow rhythms (~.05 Hz) in the stomach, with so called 'gastric phase' explaining about 8% of the variance in alpha (Richter et al., 2017). Heartbeat has also been found to play a role in variability in accuracy, such that detection performance is worse if stimuli are presented synchronous with one's heart beat (Salomon et al., 2016). It is largely unknown to what extent spontaneous variability within these sources could be accessible to consciousness. Whether this

knowledge could be used to improve behaviour also heavily relies on the time scale at which this variability unfolds.

Variability – a beneficial characteristic?

Within many contexts, both in our daily lives as well as in the laboratory, it may be tempting to see variability as a hindering by-product of a lack of attention which we would like to reduce as much as possible. However, variability may not necessarily be negative. Indeed, variability may ensure our behaviour is not entirely predictable to our preys and predators (Carpenter, 1999), and may facilitate exploration and novel behaviour (Shahan & Chase, 2002; see Sternad, 2018 for a review). Furthermore, variability and the resulting unpredictability of our behaviours are key to discussions about our sense of agency and beliefs of free will (see for examples: Brembs, 2011; Haggard, 2008; Koch, 2009; Tse, 2013). This is reflected in models of decision, where noise plays a crucial role (Bogacz, Brown, Moehlis, Holmes & Cohen, 2006; Bompas, Hedge & Sumner, 2017; Bompas & Sumner, 2011).

Importantly, variability is not limited to behaviour, but present throughout all levels of our central nervous system, even in very short-term fluctuations such as the firing of action potentials within a single neuron (with random noise contributing to whether the action potential will be initiated) and subsequent variability in post-synaptic response (which may similarly be affected by 'synaptic background noise'). Such fluctuations throughout various levels in our nervous system may affect trial-to-trial variability. This variability does not only occur as a response to external stimuli, but also in the absence thereof, as an intrinsic characteristic of the system. It has been argued that randomness is important for the functioning of the nervous system, rather than something that needs to be reduced or 'overcome' (Ermentrout et al., 2008; Faisal et al., 2008). All in all, variability appears to be an intrinsic and fundamental property, and as such, a large proportion of it may be not reducible at all.

Conclusion

Intuitively, it seems reasonable to think that people have some access to their own fluctuating performance-relevant inner states, and that they can use this information to improve their performance. In two separate experiments and across a series of empirical tests, we found repeated evidence against most predictions derived from this intuition. We found that, even though people varied the time they initiated a trial and reported that they threw darts only when ready, they were unable to improve their performance or reduce their variability, even when highly motivated to do so. Altogether, this suggests that if people have any access to their own inner states at all, this access is limited and not a key determinant for upcoming performance.

Supplementary Materials

In order to compare temporal dependencies across conditions, we first tested whether our RT and ITI measures actually contained temporal dependencies. Bayesian One Sample one-sided t-tests were used to test if the participants' autocorrelations at lag one (AC1) and the slopes of the linearly fitted power spectra were statistically higher than zero (see Supplementary Table 1 for the corresponding BF). In the perception-oriented task, there was clear evidence for temporal dependency on each measure. For the action-oriented task, evidence was more mixed. Still, out of ten data series (RT on four conditions plus self-paced ITIs, for two tasks), five showed clear evidence for temporal dependencies, and none of them showed clear evidence against.

Supplementary Table 1. *Bayes' Factors for the presence of temporal dependencies in the RT and self-paced ITIs, tested against two different measures: a positive autocorrelation function at lag 1 (AC1), and slope of the power spectrum.*

	Action-oriented		**Perception-oriented**	
Test	**AC1**	**Slope**	**AC1**	**slope**
SP_{RT}	3.00	1.00	859442	36211
F_{RT}	2.04	23.93	116692	53794
R_{RT}	32.45	1.21	8107	91101
SR_{RT}	.49	1.20	161	473
SP_{ITI}	174	501	222	23.98

Chapter 5

General discussion

In the current thesis, I investigated the properties and correlates of variability in behaviour. In the introduction, I discussed how humans are highly variable in their behaviour even on very simple actions, which a robot would be able to perform with near-constant precision. The largest part of this variability is endogenous – meaning we are highly variable even when all the circumstances in our environment are stable. Looking at prior research, I identified two main perspectives on behavioural variability: 1) the *intuitive* perspective, which describes variability as a result of fluctuations in our attentional and meta-cognitive states (e.g., mind wandering), and 2) the *intrinsic* perspective, which describes variability as an inherent feature of each level of our neurobiological system, from the lowest (e.g., firing rate variability in an individual neuron) to the higher (e.g., decision processes) level. These two perspectives are not necessarily irreconcilable, but their literatures rarely meet. This current thesis offers a comparison between the two perspectives over four different chapters – converging into three main questions:

1. How does endogenous variability manifests within and across individuals – i.e., would some individuals be closer to the robot's near-constant performance than others, and does this remain stable over time and over different circumstances?

2. To what extent does endogenous variability co-vary with fluctuations in our subjective experiences of attention – i.e., knowing that we have internal states that the robot does not have, to what extent are these internal states actually related to variability in behaviour?

3. To what extent can endogenous variability be reduced – i.e., could we come close to the robot's near-constant performance, or are we inevitably variable?

Question 1. How does endogenous variability manifests itself within and between individuals?

In Chapter 1, I examined how variability manifests within and between individuals in the oculomotor system when all external circumstances remain stable. Over three different experiments, the eye movements and pupil dilation of healthy participants were recorded in resting-state based paradigms, in which they were asked to fixate or to look at the screen for a couple of minutes. Results showed that oculomotor variability (as measured by variability in gaze position, variability in pupil dilation, blink rate, and (micro-) saccade rate) was consistent within individuals over time (*repeatability*) – with intervals between measures ranging from half an hour to multiple days apart. This means that on average, individuals who showed relatively low variability during the first measure in time remained relatively low on the subsequent measures, while participants who showed relatively high variability during the first measure remained highly variable throughout. Furthermore, oculomotor variability was consistent within individuals over different conditions *(generalisability*) – meaning that on average, participants who were highly variable when fixating on a dot were also highly variable when they were just looking at the screen without specific instructions.

In Chapter 2 we reached the same conclusion on a different task, showing that variability in a rhythmic manual task also shows high repeatability. We explored this further by investigating the temporal properties, and found these were also repeatable. Rather than focusing on one particular measure of dependency (as is

common in the literature), I looked at all the methods that have been commonly used across different articles – namely autocorrelation, Power Spectra Density (PSD) slopes, Detrended Fluctuation Analyses (DFA) slopes, and ARFIMA(1,d,1) models. I found evidence for consistency of temporal dependency within individuals over time – though this was variable across measures.

Previous studies have found that intra-individual variability is a consistent individual trait, both in standard cognitive tasks (Hultsch et al., 2002; Saville et al., 2011; 2012) and in oculomotor behaviour (Andrews & Coppola, 1999; Boot et al., 2009; Castelhano & Henderson, 2008; Poynter et al., 2013; Rayner et al., 2007). However, such variability typically consists of endogenous *and* exogenous variability. For example, Saville et al. (2011) examined the intra-individual variability of RT in *n*-back, go/no-go, and stop-signal tasks, Castelhano & Henderson (2008) examined different oculomotor measures during viewing tasks (e.g., viewing images of faces and scenes). In such cases, part of the variability within individuals is caused by task-features, such as condition order across trials. Furthermore, the stability within individuals may (at least partially) be driven by consistency in individual differences in information processing and response strategies.

My current findings build on these studies by examining specifically the reliability of endogenous variability – either in a task that remains stable throughout, or even in one's most basic oculomotor behaviour during near-rest. As such, we can say that some individuals are indeed systematically closer to the robot's performance than others – but no one still comes even close the robot's near-constancy. Differences between people are far from unsubstantial even on these basic tasks (for example, in the first experiment in Chapter 1, the highest measurements between individuals in oculomotor behaviour were 16-40 times larger than the lowest measurements, and in Chapter 2, the most variable participant's SD was 5.6 times larger than the least variable participant's).

How do individual differences arise?

An obvious follow-up question is whether these individual differences have a clear origin. Within the literature, emphasis has been put on the link between intra-

individual variability and ADHD (see Kofler et al., 2013 for a meta-analysis; see Tamm et al., 2012 for a review) – with the increased variability possibly driven by the combination of more attentional lapses and more impulsive responses. In both Chapter 1 and 2, I found evidence against correlations between variability and self-assessed personality traits in healthy participants – indicating that neuro-typical individuals who reported to have more ADHD tendencies, mind wandering tendencies, and impulsivity did not show more behavioural variability.

As my samples mainly consist of university students and colleagues, it is possible that we would have found positive correlations with ADHD tendencies by sampling from a more diverse population (for instance, by oversampling for extreme scores on the questionnaires). However, though such adjustments in sampling may explain extreme/clinical individual differences, it remains that those traits do not appear to reveal the origins of the substantial individual differences ubiquitous in our current samples.

It should be noted that increased intra-individual variability has not just been found in ADHD, but in a variety of neuroclinical disorders and diseases such Alzheimer's Disease (Tales et al., 2012; Tse et al., 2010), and schizophrenia, depression, and borderline disorder (Kaiser et al., 2008), but also in non-clinical cognitive aging (Hultsch et al., 2000; 2002). This may indicate that variability is a general marker of dysfunctioning or deterioration of the nervous system. Examining individual differences in neurocognitive functioning (e.g., structural and functional differences) may be a more fruitful approach to understanding individual differences in variability.

Reliability and correlates of temporal structure

The results of Chapter 2 do not only show the intra-individual reliability of the temporal structure measures – they also show that these measures correlate with performance between individuals, such that higher variability (i.e., worse task performance) was associated with higher temporal structure. However, it remains controversial how these structures originate, what their time scale is, and how they should be interpreted.

In the current study, I used the Metronome Task (MRT) because it is well-suited for measuring endogenous fluctuations. As a next step, it would be interesting to examine how the measures behave in different conditions. As a follow-up, I aim to use the current analysis pipeline on three datasets. The first dataset consists of two sessions per participant for four different tasks (Stroop, go/no-go, stop-signal, and Eriksen Flanker tasks; previously published in Hedge, Powell & Sumner, 2018). The second dataset contains data of two tasks (Eriksen Flanker and dot-motion tasks; Hedge et al., in prep), both of which participants conducted under a 'speed-focused' and an 'accuracy-focused' instruction – aiming to manipulate the speed-accuracy trade-off. The third dataset also contains a speed-accuracy manipulation (Eriksen Flanker and Stroop tasks; Hedge et al., in prep), but over two different sessions, allowing for test-retest reliability. These datasets allow for the assessment of the intra-individual consistency within (reliability) and between (generalisability) different tasks and different instructions. On the one hand, it is possible that temporal structures are different when a task is performed in a different manner. On the other hand, it has been argued that these structures are general features of the organisation of our biological system (e.g., criticality) that may be highly common no matter what one is doing. If the latter is true though, one may wonder to what extent these measures are informative for human behaviour.

Aside from experimental data, the analysis pipeline may also be conducted on simulated data (e.g., simulated data series by a linear ballistic accumulation model), in which the temporal dependency is manipulated in one of the parameters – to examine whether the temporal dependency measures can actually recover these structures.

Some of the temporal dependency measures (i.e., DFA and PSD) have not just been applied to behavioural data, but also to neuroimaging data. Temporal structures may be examined in resting-state data (for example, comparing DFA slopes in healthy versus schizophrenia patients; Nikulin, Jönsson & Brismar, 2012) or on task-related data (for example, correlating DFA slopes in behavioural hit/miss series with DFA slopes in M/EEG data recorded during task and during rest; Palva et al., 1999). As the dataset of Chapter 4 contains four resting-state sessions (pre- and post-task for both days), I could assess the reliability of the temporal structures

on resting-state MEG data over time using a wider array of time series analysis methods. Similarly, the reliability of the temporal structures in MEG data during task and during rest can be assessed over the two days, as well as the relationships between the different temporal structures across analysis methods. Possibly, the structure of a concurrent physiological measure could also be assessed, as pupil dilation was recorded throughout both sessions. As such, I would be able to examine the intra-individual correlations of the different temporal structures.

Question 2. To what extent does endogenous variability co-vary with fluctuations in meta-cognitive states?

In Chapter 3, I examined the relationships between behavioural variability, meta-cognition, and underlying neural states (as measured by oscillatory power). Over two sessions, participants performed the MRT, in which they had to press a button in synchrony with a tone, and were pseudo-randomly asked to rate their metacognitive states. During both sessions, MEG data was collected. On a behavioural level, I found that variability and subjective attentional state correlate with each – showing that when people report to be more off-task, they are more variable on the trials just before the report other (replicating previous findings; Laflamme et al., 2018; Seli et al., 2013). Furthermore, my current research is the first to show that the neural states underlying attentional state reports showed overlap with those underlying behavioural variability in the β frequency band. However, overlap was not high and not fully convincing, as participants who showed higher correlations between their behavioural variability and meta-cognitive reports did not show more overlap in the respective underlying states.

Subjective performance ratings correlated better to variability than the attentional state reports. Also, there was evidence for overlap in neural states for all the frequency bands. In contrast to the attentional state reports, the performance reports also correlated to the task-/instructions-relevant behavioural measure (absolute RT). This implies that the choice of meta-cognitive reports in an experiment should be picked with deliberation. It makes sense to include attentional

state ratings if one is particularly interested in this metacognitive experience. However, if one is interested in the best correlate to behaviour (as is often the case in topics like driving behaviour), the current results suggest that subjective experiences of performance are more interesting.

Overall, my findings highlight another difference between humans and robots: Not only do individuals exhibit endogenous variability while robots perform near-constancy, the temporal fluctuations of such variability co-varies with fluctuations in the meta-cognitive states that robots do not experience at all. Furthermore, these fluctuations partly share the same underlying neural mechanisms. However, the effect sizes of these relationships are weak (~3-5.5% shared variance between meta-cognition and behavioural variability, and ~2-5% neural overlap). Let us go back to the thought experiment about the museum and imagine that a clever genius invents a pill against inner fluctuations that cause experiences of off-taskness. When the robot is out for repair, the Board asks you to take these magic pills while you are counting the number of visitors. As you are now 100% focused on your task throughout the day, you may expect that your variability is largely reduced. However, our and previous results imply that even with these pills, at least ~95% of your variance remains – meaning you are still nowhere near the robot's performance.

It should be emphasised that the current results can only establish a correlation between subjective attentional state and behavioural variability. This can be interpreted in a number of ways. An extreme explanation of these correlations is that the poor attentional states *cause* behavioural variability – i.e., you are more variable than the robot because you are not always paying attention. Such an interpretation assumes a direct effect between attentional state and variability. Another extreme interpretation would be that the subjective experiences of off-taskness and the behavioural variability are just different markers of the same process. While both interpretations can be found in the literature, one could wonder why the shared variance between attentional state and variability is so weak if they are either the same process or one is directly causing the other. Furthermore, to confirm causal mechanisms, one needs more than correlational evidence. In this case, it is difficult to imagine what form this would take – though it has been

suggested that the induction of sad mood increases mind wandering (Smallwood et al., Philips, 2009), and that mindfulness meditation training may reduce variability (but see section *Mindfulness* below).

More conservatively, one could conclude that attentional state and behavioural variability are indirectly related – i.e., our biological systems exhibit natural fluctuations over time, that cause both variability in behaviour and variability in meta-cognitive experiences. It is possible that meta-cognitive experiences are merely an epiphenomenon of our biological system – meaning they can arise from but not influence our system. Within the study of consciousness and experience of agency, this has already been studied in more detail (Wegner & Wheatley, 1999). As such, these found correlations just indicate that the two co-vary over time. In Chapter 4, some options for such underlying states were already mentioned: they may be caused by fluctuations in default mode network activity (Christoff, Gordon, Smallwood, Smith & Schooler, 2009; Mason et al., 2007) or in task-related network activity (such as the dorsal attention network; Corbetta, Patel & Shulman, 2008), or the anticorrelation between these networks (Kelly et al., 2008), but maybe also by fluctuations in non-nervous system activity, such as stomach rhythms (Richter et al., 2017) or heart beat (Salomon et al., 2016). While this conservative interpretation seems more on par with the weak effect sizes of the correlations between attentional state and variability, the current evidence can still support both the more extreme and more conservative interpretations.

Temporal fluctuations

Although one important assumption of the intuitive perspective is that meta-cognitive states and behavioural variability co-fluctuate over time, there has been little work done on comparing their respective temporal dependencies. Irrmisscher et al. (2018) pseudo-randomly probed participants about attentional state during a meditation task (100 probes on a 5-point scale in 12 minutes of meditation), and calculated a DFA slope on the series of these ratings. They found that on average, participants who reported to be more off-task had a higher DFA slope. However, this task does not come with any behavioural data to which this slope can be compared. In the design of MacDonald et al. (2011), participants were asked about their

attentional state on each trial during a detection task – allowing for a trial-to-trial level comparison. A Fourier transform was conducted on both the ratings and detection accuracy over the experiment. These spectra showed vastly different forms, though their respective slopes were not statistically compared. However, it should be noted that the contrast of the stimulus was not constant throughout the experiment – meaning part of the fluctuations in detection accuracy are exogenously driven. Future studies may take a more rigorous approach by obtaining endogenous performance and subjective attentional state measures on each trial, and comparing the temporal structures with different measures (comparable to the approach in Chapter 2) between the objective and subjective measures.

Task-effects on meta-cognitive ratings

One potential caveat of asking for an attentional rating on each trial is that it may affect participants' ratings: If ratings are far apart, there may be more chance to 'mentally drift off', while ratings close together may disrupt the task-performance. Recently, the effects of probe frequency were empirically tested (Robison et al., 2019). Participants performed 675 SART-trials, and were presented either with 45 or with 90 pseudo-random probes throughout. They found evidence against an effect of frequency both on behavioural and subjective task measures – although evidence was weak, with BF_{10} ranging from 2.1-3.8. However, an opposite result was found by Seli, Carriere, Levene & Smilek (2013). Their participants performed 600 MRT-trials, and received minimally five to maximally 25 thought probes throughout. They found that the time between probes was positively correlated with the reported amount of off-taskness, but not with objective performance – although it seems this correlation was largely driven by a few outliers. A follow-up experiment may test the effects over a larger range of trials, as well as examine how this affects the within-subject relationship between objective and subjective measures.

In a separate experiment, Robison et al. (2019) also examined the effects of instructions – giving participants either neutral, negative (avoid mind wandering during the task), or positive (it is fine to mind wander during the task) instructions, and found moderate evidence against an effect on both the subjective ratings and

the objective performance. In my thesis, the instructions (both for Chapter 2 and 4) aligned most with the 'negative' condition: Participants were specifically instructed to stay on task and perform as well as possible. Indeed, this condition appears most interesting for studying off-taskness. First, it allows for the examination of off-taskness when participants are trying to stay on-task – mirroring typical real-life tasks. Secondly, if a participant is encouraged to mind wander throughout a task, it is unclear whether the behavioural task is actually the 'primary' or the 'secondary' task. It should be mentioned that, unlike Robison et al. (2019), I also included a reward system in the experiments. This system gives participants an intrinsic reason to stay on task. Future studies may investigate whether the combination of intrinsic ('reward') and extrinsic ('instructions') may affect the subjective ratings and/or behavioural performance.

It should be noted that current 'probe-caught' methods to capture subjective off-taskness are based on pseudo-random algorithms. An alternative approach may be to trigger thought probes when behavioural variability is very low, very high, or medium – which should respectively capture on-taskness, off-taskness, or in between reports. Instead of behavioural variability, other measures as psychophysiological or neural activity may also be used. Such an approach has been partly implemented in Henríquez, Chica, Billeke & Bartolomeo (2016), though their algorithm only presented a thought probe if an extreme RT (SD > 2) was detected. However, based on unpublished pilot data, the downside of this approach is that it may be difficult to establish a good algorithm; wrong estimates may lead to an oversampling of one subjective state.

Another task-parameter in the MRT that may be of interest is the interval between the tones. In Chapter 2, I used the original interval of 1.3 seconds (as tested by Seli et al., 2013). However, Chapter 3 features a longer interval of 3 seconds, because I was interested in the baseline neural activity in between the tones and presses (the *'event-free periods'*). Increasing the interval was necessary to get a decent event-free period, and thus to increase statistical power. Of course, one may wonder if such an increase in interval would have affected any of our results, as previous studies on tapping have found a relationship between metronome-interval and performance. In particular, it has been suggested that longer intervals cause

shifts in the tapping distribution from ‘anticipations’ to ‘reactions’ – meaning that performance becomes less predictive and more responsive of/to the tone as interval length is increased (Miyake, Onishi, Pöppel, 2004; Takano & Miyake, 2007). For example, Miyake et al. (2004) tested tapping in ten different intervals (ranging from 450 ms to 6000 ms). In the shortest interval, the majority of taps were recorded from -100 to 0 ms before tone onset, while in the longest interval, the majority of taps was recorded 150 to 300 ms after tone onset – suggesting that participants found it more and more difficult to predict the tone as the interval length increased.

However, Repp & Doggett (2007) found that when participants are specifically instructed to predict the tone and *not* take on a reactive strategy, participants succeed to do so. They speculate that the shift towards more anticipations is driven largely by instructions and motivation. In my experiments, instructions were similar to the instructions from Repp & Doggett (2007): Though I did not give participants specific instructions on when to press, they were told that *‘one helpful strategy may be to pretend you are the one causing the tone with the button press’.* Indeed, the distributions from Chapter 3 are still shifted below zero for most participants.

Still, it is possible that the task from Chapter 3 was more difficult than Chapter 2. For now, it remains unclear what the implications of this are. Even if interval duration can affect behavioural variability, it remains unclear if interval duration in the MRT can affect the *amount* of off-taskness or the *relationship* between behavioural variability and off-taskness. This raises an interesting question that has been largely unaddressed in the current thesis: To what extent task difficulty/cognitive load can affect off-taskness and its relationship with variability. It is possible that the difficulty of the task affects the off-taskness ratings themselves – that is, higher difficulty could mean less boredom and less reported off-taskness. While I did find that the off-taskness ratings in Chapter 3 were quite low, it is difficult to compare the ratings between Chapter 2 and 3 directly, because of differences in samples (Chapter 2: undergraduate students recruited from an online system who took part for course credits; Chapter 3: postgraduate students and academic peers who were familiar with participating in neuroimaging studies). For more conclusive results, one might manipulate the MRT with different duration intervals in the same

experimental group for direct comparisons. This could be extended with phase vs. anti-phase instructions (as anti-phase tapping is associated with more behavioural variability; Engström, Kelso & Holroyd, 1996).

Aside from task-parameters, the task itself could also affect the measures of interest. For example, in the MRT, it is difficult for participants to have an exact representation of how well they are performing on the task; while we may be capable of knowing whether we are very far off the tone, it is impossible to notice the difference between 20 ms off or 30 ms off. Instead, when using tasks like the SART, participants will be aware of whether they made an (omission or commission) error. As these errors are accessible to us, it may be more beneficial to use them as information for other metacognitive judgements (such as attentional state) – as such leading to higher associations between them. Future studies may investigate to what extent meta-cognitive experiences and their respective relation with performance are task-dependent. Related to this, it may also be interesting to examine to what extent the meta-cognitive ratings are influenceable by accessible information. For example, if participants are not (fully) aware of their own performance on a task, giving them false (either too positive or too negative) feedback may affect their meta-cognitive ratings. This may give us more insight into what information participants use to rate their experiences.

The concept of attention

Throughout the current thesis, I have discussed both the concept ‘mind wandering’ and the concept ‘attention’. An open question is how these concepts relate to each other. This issue has not been addressed much in the literature – and even more so, these concepts are often used interchangeably. I have briefly discussed this relationship in Chapter 4 (*Introduction,* section *Endogenous variability and its accessibility*), and speculated that a possible distinction might be found in the level of ‘awareness’: As a metacognitive process, mind wandering requires some form of awareness, even if it is post hoc, while attention does not seem to have such a requirement. Another distinction may lie in their specificities: Mind wandering appears to describe one particular form of off-taskness (task-unrelated thoughts), while attention is an umbrella term that refers to a multitude of processes – such as

(but not limited to) task-focus and high arousal, but also predictability and (cued) orientation (e.g., keeping attention to a specific spatial orientation because the target is likely to appear there). Indeed, the concept of 'attention' within cognitive neuroscience has recently been criticised (Hommel et al., 2019) for referring to too many different processes and being used simultaneously as the 'explanandum'(the phenomenon we are trying to explain) and the 'explanans' (the phenomenon we are trying to explain *with*). To advance the field, it may be necessary to tear apart the umbrella concepts, and focus on pinpointing (smaller) neurobiological processes. My current effort to distinguish different types of meta-cognitive off-taskness processes (i.e., mind wandering, mind blanking alert, and mind blanking drowsy) is one small step in this direction.

Question 3. To what extent is endogenous variability reducible?

In Chapter 4, I tested whether we can access and act upon the fluctuations in our performance-relevant states to reduce behavioural variability. While this sounds somewhat complicated, it is something that we typically assume in our daily lives. To test this prediction, I let participants play a game of darts in a self-paced and a fixed-paced condition. If people can access and act upon their performance-relevant states, they would be able to wait for the 'right' moment to throw the darts – an assumption that we hold in daily life. As they can utilise this in the self-paced but not in the fixed-paced condition, this should lead to better performance in the self-paced task. We found evidence against this prediction. Next, we tested the same prediction in two traditional computer-based psychophysics tasks, comparing a self-paced condition to three different forced-paced condition – and again found evidence against better performance in the self-paced task. Overall, these findings imply that we cannot access and act upon our performance-relevant states to reduce our variability.

Aside from testing the effects of control on variability, the darts-based task also shows the feasibility of using more fun and engaging tasks to test hypotheses. This is interesting, as most neurocognitive experiments are very boring by nature –

under the assumption that boring – 'clean' – stimuli will lead to cleaner data. However, some previous studies have investigated the neuronal variability that is induced in reaction to either typically-controlled or more natural stimuli, and found that natural stimuli do not lead to as much variability as previously assumed (Hasson, Malach & Heeger, 2010; Herikstad, Baker, Lachaux, Gray & Yen, 2011). One could also imagine that boring stimuli would induce off-taskness and/or behavioural variability in participants, as they would aim to perform 'good enough' rather than 'as well as possible'. More engaging tasks and stimuli may help participants stay involved in the task. However, while my study cannot test the effect of boring versus engaging tasks directly, it did show that the effect of control was absent in both the engaging darts-task and the boring psychophysics tasks.

Even if we are not capable of accessing our internal states to reduce variability ourselves, one may wonder this could be automated. I touched upon this in Chapter 3, when discussing that researchers within driving research are interested in the neural mechanisms of mind wandering, so that mind wandering can be detected 'online' during driving – subsequently preventing accidents. As previous literature has found an association between neural states and off-taskness and between off-taskness and variability, it has been assumed that this approach is viable. However, the fruitfulness of such approaches depends largely on the effect sizes (which have typically been unreported). As shown in Chapter 3, these effect sizes are low, and off-taskness and behavioural variability are poor markers of each other. One may argue that instead of detecting 'off-taskness' states, the online algorithm could aim to just the detect the neural states underlying poor performance. However, my results emphasise the large inter-individual differences in underlying neural mechanisms. This means that whatever mechanism is found to underlie off-taskness on the group level is uninformative for the detection of off-taskness in a new participant.

Yet another way one could aim to detect and to prevent poor future performance is by studying past performance its temporal structures. For example, it is known that RT on trial *n* is positively correlated to RT on trial *n+1* (see Chapter 2 and 3). This means that if someone shows poor performance on an action, it is more likely that the upcoming action will be poorly executed too. As these temporal

structures are reliable within individuals, the structure from a first session can give a good estimation of the structure of future sessions. Again, however, my current results suggest that effect sizes are low. Future studies may focus on the timescales on which performance is predictable, and on which temporal dependency measures are best suited for measuring structures in behaviour.

Overall, my findings suggest that: 1) we do not have the means to access fluctuations in our performance-relevant states to reduce variability, 2) it is not viable to reduce variability by detecting neural states of 'poor performance modes', and 3) if we know the temporal structure and past behaviour of an individual on action *n-1* to *n-k* (with *n-k* being the maximum amount of informative past actions), we may be able to predict behaviour on action *n* better than chance – although, given the weak effect sizes, prediction accuracy will likely be low. All in all, it seems that even with these measures, our performance will not at all come close to the robot's near-constancy – meaning our behavioural variability is largely irreducible.

Mindfulness

Prior studies have found that mindfulness meditation training reduces behavioural variability (Brown & Ryan, 2003; Wells, 2005; Zeidan et al., 2010; Morrison et al., 2014; Mrazek et al., 2013). This makes sense from the intuitive framework: Mindfulness appears to be the antipole of mind wandering – and hence, being more mindful would lead to less variability. While the effect sizes are moderate (e.g., Morrison et al. (2014) found an ~8.5% reduction in variability seven weeks of training), this is still higher than the typical effect size between subjective ratings and behavioural variability.

However, meditation studies remain difficult to conduct and to interpret due to its lack of proper control groups. The effect of training can be assessed in two ways. First, one could compare skilled meditators to unskilled control subjects. However, this means comparing groups that differ in more than just the effect of interest. Second, one could teach meditation to naïve subjects and compare them to a control group. However, it is difficult to constitute a proper control. Also, one may wonder how much training is necessary to become a (semi-) skilled meditator.

It should also be noted that these types of studies may be vulnerable to publication biases – and may thus be good candidates for preregistration.

Conclusion

With the current thesis, I examined the properties and correlates of endogenous behavioural variability. To summarise, I found that: 1) some individuals are further away from the near-constant performance of a robot than others, across different condition and time, 2) endogenous variability covaries over time with meta-cognitive states (which a robot does not experience), both on a behavioural and a neural level, and 3) the largest part of this variability appears inaccessible and irreducible to us.

While the intuitive framework typically assumes a strong and possibly direct link between meta-cognitive states and behavioural variability, the current empirical findings indicate that this link is clearly weak. While similarly weak effect sizes have been found previously in the literature, they are rarely emphasised as such. I would argue that this is a dangerous symptom of the intuitive framework: If theoretical mechanisms match our deeply-rooted intuitions, it is easy to take them for granted even if they do not fully explain the phenomena.

Adherent to the intrinsic perspective, behavioural variability may arise (at least for the largest part) from a multitude of biological fluctuations – possibly combined with randomness, though this has a high burden of proof. Rather than focusing on single biological predictors, future research may aim to combine different neural, psychophysiological, and behavioural measures to correlate to upcoming behavioural variability, in order to assess unique contributions and total explained effect sizes.

References

Aase, H., Meyer, A., & Sagvolden, T. (2006). Moment-to-moment dynamics of ADHD behaviour in South African children. *Behavioral and Brain Functions, 2*(1), 11. https://doi.org/10.1186/1744-9081-2-11

Aase, H., & Sagvolden, T. (2005). Moment-to-moment dynamics of ADHD behaviour. *Behavioral and Brain Functions, 1*(1), 12. https://doi.org/10.1186/1744-9081-1-12

Adamo, N., Baumeister, S., Hohmann, S., Wolf, I., Holz, N., Boecker, R., … Brandeis, D. (2015). Frequency-specific coupling between trial-to-trial fluctuations of neural responses and response-time variability. *Journal of Neural Transmission, 122*(8), 1197–1202. https://doi.org/10.1007/s00702-015-1382-8

Adler, L. A., Shaw, D. M., Spencer, T. J., Newcorn, J. H., Hammerness, P., Sitt, D. J., … Faraone, S. V. (2012). Preliminary Examination of the Reliability and Concurrent Validity of the Attention-Deficit/Hyperactivity Disorder Self-Report Scale v1.1 Symptom Checklist to Rate Symptoms of Attention-Deficit/Hyperactivity Disorder in Adolescents. *Journal of Child and Adolescent Psychopharmacology*, *22*(3), 238–244. https://doi.org/10.1089/cap.2011.0062

Adler, L. A., Spencer, T. J., Faraone, S. V., Kessler, R. C., Howes, M. J., Biederman, J., & Sečnik, K. (2006). Validity of pilot Adult ADHD Self- Report Scale (ASRS) to Rate Adult ADHD symptoms. *Annals of Clinical Psychiatry : Official Journal of the American Academy of Clinical Psychiatrists*, *18*(3), 145–148. https://doi.org/10.1080/10401230600801077

Akaike, H. (1998). A New Look at the Statistical Model Identification. In E. Parzen, K. Tanabe, & G. Kitagawa (Eds.), *Selected Papers of Hirotugu Akaike* (pp. 215–222). https://doi.org/10.1007/978-1-4612-1694-0_16

Andrews, T. J., & Coppola, D. M. (1999). Idiosyncratic characteristics of saccadic eye movements when viewing different visual environments. *Vision Research*, *39*(17), 2947–2953. https://doi.org/10.1016/S0042-6989(99)00019-X

Andrews-Hanna, J. R., Smallwood, J., & Spreng, R. N. (2014). The default network and self-generated thought: Component processes, dynamic control, and clinical relevance. *Annals of the New York Academy of Sciences*, *1316*(1), 29–52. https://doi.org/10.1111/nyas.12360

Arazi, A., Censor, N., & Dinstein, I. (2017). Neural Variability Quenching Predicts Individual Perceptual Abilities. *The Journal of Neuroscience, 37*(1), 97–109. https://doi.org/10.1523/JNEUROSCI.1671-16.2016

Arazi, A., Gonen-Yaacovi, G., & Dinstein, I. (2017). The Magnitude of Trial-By-Trial Neural Variability Is Reproducible over Time and across Tasks in Humans. ENeuro, 4(6). https://doi.org/10.1523/ENEURO.0292-17.2017Baird, B., Smallwood, J., & Schooler, J. W. (2011). Back to the future: Autobiographical planning and the functionality of mind-wandering. *Consciousness and Cognition*, *20*(4), 1604–1611. https://doi.org/10.1016/j.concog.2011.08.007

Arthur, R. (2015, June 12). Big Papi needs more time to think. *FiveThirtyEight.* Retrieved from: https://fivethirtyeight.com/features/big-papi-needs-more-time-to-think/ on December 18, 2019.

Arthur, R. (2016, July 28). MLB Games are slow again, and it's helping older hitters. *FiveThirtyEight*. Retrieved from: https://fivethirtyeight.com/features/mlb-games-are-slow-again-and-its-helping-older-hitters/ on December 18, 2019.

Asendorpf, J. B., Conner, M., Fruyt, F. D., Houwer, J. D., Denissen, J. J. A., Fiedler, K., … Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*(2), 108–119. https://doi.org/10.1002/per.1919

Baayen, R. H., & Milin, P. (2010). Analyzing Reaction Times. *International Journal of Psychological Research, 3*(2), 12–28.

Baer, R. A., Smith, G. T., Hopkins, J., Krietemeyer, J., & Toney, L. (2006). Using self-report assessment methods to explore facets of mindfulness. *Assessment*, *13*(1), 27–45. https://doi.org/10.1177/1073191105283504

Baillet, S. (2017). Magnetoencephalography for brain electrophysiology and imaging. *Nature Neuroscience*, *20*(3), 327–339. https://doi.org/10.1038/nn.4504

Baird, B., Smallwood, J., Lutz, A., & Schooler, J. W. (2014). The Decoupled Mind: Mind-wandering Disrupts Cortical Phase-locking to Perceptual Events. *Journal of Cognitive Neuroscience*, *26*(11), 2596–2607. https://doi.org/10.1162/jocn_a_00656

Baird, B., Smallwood, J., & Schooler, J. W. (2011). Back to the future: Autobiographical planning and the functionality of mind-wandering. *Consciousness and Cognition*, *20*(4), 1604–1611. https://doi.org/10.1016/j.concog.2011.08.007

Bakker, M., Hartgerink, C. H. J., Wicherts, J. M., & van der Maas, H. L. J. (2016). Researchers' intuitions about power in psychological research. *Psychological Science*, *27*(8), 1069–1077. https://doi.org/10.1177/0956797616647519

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*(6), 543–554. https://doi.org/10.1177/1745691612459060

Baldwin, C. L., Roberts, D. M., Barragan, D., Lee, J. D., Lerner, N., & Higgins, J. S. (2017). Detecting and Quantifying Mind Wandering during Simulated Driving. Frontiers in Human Neuroscience, 11. https://doi.org/10.3389/fnhum.2017.00406

Beck & Steer, R. A. (1993). *The Beck Anxiety Inventory*. San Antonio, TX: The Psychological Corporation.

Beck, A. T., Steer, R. A. & Brown, G.K. (1996). *The Beck Depression Inventory-II*. San Antonio, TX: Psychological Corporation.

van Beers, R. J., van der Meer, Y., & Veerman, R. M. (2013). What Autocorrelation Tells Us about Motor Variability: Insights from Dart Throwing. PLoS ONE 8(5): e64332. doi:10.1371/journal.pone.0064332

Beggs, J. M., & Timme, N. (2012). Being Critical of Criticality in the Brain. *Frontiers in Physiology*, *3*. https://doi.org/10.3389/fphys.2012.00163

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Berg, J. M., Latzman, R. D., Bliwise, N. G., & Lilienfeld, S. O. (2015). Parsing the heterogeneity of impulsivity: A meta-analytic review of the behavioral implications of the UPPS for psychopathology. *Psychological Assessment*, *27*(4), 1129–1146. https://doi.org/10.1037/pas0000111

Bergomi, C., Tschacher, W., & Kupper, Z. (2013). The Assessment of Mindfulness with Self-Report Measures: Existing Scales and Open Issues. *Mindfulness*, *4*(3), 191–202. https://doi.org/10.1007/s12671-012-0110-9

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, *113*(4), 700–765. https://doi.org/10.1037/0033-295X.113.4.700

Bompas, A., & Sumner, P. (2011). Saccadic Inhibition Reveals the Timing of Automatic and Voluntary Signals in the Human Brain. *Journal of Neuroscience*, *31*(35), 12501–12512. https://doi.org/10.1523/JNEUROSCI.2234-11.2011

Bompas, A., Hedge, C., & Sumner, P. (2017). Speeded saccadic and manual visuo-motor decisions: Distinct processes but same principles. *Cognitive Psychology*, *94*, 26–52. https://doi.org/10.1016/j.cogpsych.2017.02.002

Bompas, A., Sumner, P., Muthukumaraswamy, S. D., Singh, K. D., & Gilchrist, I. D. (2015). The contribution of pre-stimulus neural oscillatory activity to spontaneous response time variability. *NeuroImage*, *107*, 34–45. https://doi.org/10.1016/j.neuroimage.2014.11.057

Boot, W. R., Becic, E., & Kramer, A. F. (2009). Stable individual differences in search strategy? The effect of task demands and motivational factors on scanning strategy in visual search. *Journal of Vision*, *9*(3), 7. https://doi.org/10.1167/9.3.7

Box, G.E.P., Jenkins, G.M., Reinsel, G.C. and Ljung, G.M. (2016) *Time Series Analysis: Forecasting and Control*. Fifth Edition, Wiley Series in Probability and Statistics, John Wiley & Sons, Inc., Hoboken.

Braboszcz, C., & Delorme, A. (2011). Lost in thoughts: Neural markers of low alertness during mind wandering. *NeuroImage*, *54*(4), 3040–3047. https://doi.org/10.1016/j.neuroimage.2010.10.008

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436. https://doi.org/10.1163/156856897X00357

Brembs, B. (2011). Towards a scientific concept of free will as a biological trait: spontaneous actions and decision-making in invertebrates. *Proceedings of the Royal Society B: Biological Sciences*, *278*(1707), 930–939. https://doi.org/10.1098/rspb.2010.2325

Brenner, E., van Dam, M., Berkhout, S., & Smeets, J. B. J. (2012). Timing the moment of impact in fast human movements. *Acta Psychologica, 141,* 104-111. https://doi.org/10.1016/j.actpsy.2012.07.002

Brown, K. W., & Ryan, R. M. (2003). The benefits of being present: Mindfulness and its role in psychological well-being. *Journal of Personality and Social Psychology*, *84*(4), 822–848. https://doi.org/10.1037/0022-3514.84.4.822

Busch, N. A., Dubois, J., & VanRullen, R. (2009). The Phase of Ongoing EEG Oscillations Predicts Visual Perception. *Journal of Neuroscience*, *29*(24), 7869–7876. https://doi.org/10.1523/JNEUROSCI.0113-09.2009

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Carpenter, R. H. S. (1999). A neural mechanism that randomises behaviour. *Journal of Consciousness Studies*, *6*(1), 13–22. Castelhano, M. S., & Henderson, J. M. (2008). Stable individual differences across images in human saccadic eye movements. *Canadian Journal of Experimental Psychology*, *62*(1), 1–14. https://doi.org/10.1037/1196-1961.62.1.1

Castelhano, M. S., & Henderson, J. M. (2008). Stable individual differences across images in human saccadic eye movements. *Canadian Journal of Experimental Psychology, 62*(1), 1–14. https://doi.org/10.1037/1196-1961.62.1.1

Castellanos, F. X., Sonuga-Barke, E. J. S., Scheres, A., Di Martino, A., Hyde, C., & Walters, J. R. (2005). Varieties of Attention-Deficit/Hyperactivity Disorder-Related Intra-Individual Variability. *Biological Psychiatry*, *57*(11), 1416–1423. https://doi.org/10.1016/j.biopsych.2004.12.005

Cheyne, A. J., Solman, G. J. F., Carriere, J. S. A., & Smilek, D. (2009). Anatomy of an error: A bidirectional state model of task engagement/disengagement and attention-related errors. *Cognition*, *111*(1), 98–113. https://doi.org/10.1016/j.cognition.2008.12.009

Cheyne, J. A., Carriere, J. S. A., & Smilek, D. (2006). Absent-mindedness: Lapses of conscious awareness and everyday cognitive failures. *Consciousness and Cognition*, *15*(3), 578–592. https://doi.org/10.1016/j.concog.2005.11.009

Christoff, K. (2012). Undirected thought: Neural determinants and correlates. *Brain Research*, *1428*, 51–59. https://doi.org/10.1016/j.brainres.2011.09.060

Christoff, K., Gordon, A. M., Smallwood, J., Smith, R., & Schooler, J. W. (2009). Experience sampling during fMRI reveals default network and executive system contributions to mind wandering. *Proceedings of the National Academy of Sciences*, *106*(21), 8719–8724. https://doi.org/10.1073pnas.0900234106

Churchland, M. M. (2006). Neural Variability in Premotor Cortex Provides a Signature of Motor Preparation. *Journal of Neuroscience*, *26*(14), 3697–3712. https://doi.org/10.1523/JNEUROSCI.3762-05.2006

Churchland, Mark M., Yu, B. M., Cunningham, J. P., Sugrue, L. P., Cohen, M. R., Corrado, G. S., … Shenoy, K. V. (2010). Stimulus onset quenches neural variability: a widespread cortical phenomenon. *Nature Neuroscience*, *13*(3), 369–378. https://doi.org/10.1038/nn.2501

Ciuffreda, K. J., &, B. Tannen, B. (1995). *Eye movement basics for the clinician.* Mosby, St. Louis.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology*, *65*(3), 145–153. https://doi.org/10.1037/h0045186

Cohen, M. R., & Maunsell, J. H. R. (2009). Attention improves performance primarily by reducing interneuronal correlations. *Nature Neuroscience*, *12*(12), 1594–1600. https://doi.org/10.1038/nn.2439

Cohn, P. J. (1990). Preperformance Routines in Sport: Theoretical Support and Practical Applications. *The Sport Psychologist*, *4*(3), 301–312. https://doi.org/10.1123/tsp.4.3.301

Cooper, N., Burgess, A., Croft, R., & Gruzelier, J. (2006). Investigating evoked and induced electroencephalogram activity in task-related alpha power increases during an internally directed attention task. *Neuroreport*, *17*(2), 205–208. https://doi.org/10.1097/01.wnr.0000198433.29389.54

Cooper, N. R., Croft, R. J., Dominey, S. J. J., Burgess, A. P., & Gruzelier, J. H. (2003). Paradox lost? Exploring the role of alpha oscillations during externally vs. internally directed attention and the implications for idling and inhibition hypotheses. *International Journal of Psychophysiology*, *47*(1), 65–74. https://doi.org/10.1016/S0167-8760(02)00107-1

Constantine, W., & Percival, D. (2017). Fractal: A fractal time series modeling and analysis package. R package version 2.0-4. Available at: https://CRAN.R-project.org/package=fractal

Corbetta, M., Patel, G., & Shulman, G. L. (2008). The Reorienting System of the Human Brain: From Environment to Theory of Mind. *Neuron*, *58*(3), 306–324. https://doi.org/10.1016/j.neuron.2008.04.017

Cotterill, S. (2010). Pre-performance routines in sport: current understanding and future directions. *International Review of Sport and Exercise Psychology*, *3*(2), 132–153. https://doi.org/10.1080/1750984X.2010.488269

Christoff, K. (2012). Undirected thought: Neural determinants and correlates. *Brain Research*, *1428*, 51–59. https://doi.org/10.1016/j.brainres.2011.09.060

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555

Cubitt, R. P., Starmer, C., & Sugden, R. (1998). On the validity of the random lottery incentive system. *Experimental Economics*, *1*(2), 115–131. https://doi.org/10.1007/BF01669298

Danielmeier, C., & Ullsperger, M. (2011). Post-Error Adjustments. *Frontiers in Psychology*, *2*, 233. https://doi.org/10.3389/fpsyg.2011.00233

D'Argembeau, A., Renaud, O., & Linden, M. V. der. (2011). Frequency, characteristics and functions of future-oriented thoughts in daily life. *Applied Cognitive Psychology*, *25*(1), 96–103. https://doi.org/10.1002/acp.1647

Delignières, D., Lemoine, L., & Torre, K. (2004). Time intervals production in tapping and oscillatory motion. *Human Movement Science*, *23*(2), 87–103. https://doi.org/10.1016/j.humov.2004.07.001

Delignieres, D., Ramdani, S., Lemoine, L., Torre, K., Fortes, M., & Ninot, G. (2006). Fractal analyses for 'short' time series: A re-assessment of classical methods. *Journal of Mathematical Psychology*, *50*(6), 525–544. https://doi.org/10.1016/j.jmp.2006.07.004

Delignières, D., Torre, K., & Lemoine, L. (2005). Methodological issues in the application of monofractal analyses in psychological and behavioral research. *Nonlinear Dynamics, Psychology, and Life Sciences*, *9*(4), 435–461.

van Dijk, H., Schoffelen, J.-M., Oostenveld, R., & Jensen, O. (2008). Prestimulus oscillatory activity in the alpha band predicts visual discrimination ability. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *28*(8), 1816–1823. https://doi.org/10.1523/JNEUROSCI.1853-07.2008

Donders, A. R. T., van der Heijden, G. J. M. G., Stijnen, T., & Moons, K. G. M. (2006). Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*, *59*(10), 1087–1091. https://doi.org/10.1016/j.jclinepi.2006.01.014

Drewes, J., & VanRullen, R. (2011). This Is the Rhythm of Your Eyes: The Phase of Ongoing Electroencephalogram Oscillations Modulates Saccadic Reaction Time.

Journal of Neuroscience, *31*(12), 4698–4708. https://doi.org/10.1523/JNEUROSCI.4795-10.2011

van den Driessche, C., Bastian, M., Peyre, H., Stordeur, C., Acquaviva, É., Bahadori, S., … Sackur, J. (2017). Attentional Lapses in Attention-Deficit/Hyperactivity Disorder: Blank Rather Than Wandering Thoughts. *Psychological Science*, *28*(10), 1375–1386. https://doi.org/10.1177/0956797617708234

Dutilh, G., van Ravenzwaaij, D., Nieuwenhuis, S., van der Maas, H. L. J., Forstmann, B. U., & Wagenmakers, E.-J. (2012b). How to measure post-error slowing: A confound and a simple solution. *Journal of Mathematical Psychology*, *56*(3), 208–216. https://doi.org/10.1016/j.jmp.2012.04.001

Dutilh, G., Vandekerckhove, J., Forstmann, B. U., Keuleers, E., Brysbaert, M., & Wagenmakers, E.-J. (2012a). Testing theories of post-error slowing. *Attention, Perception, & Psychophysics*, *74*(2), 454–465. https://doi.org/10.3758/s13414-011-0243-2

Engbert, R., & Kliegl, R. (2003). Microsaccades uncover the orientation of covert attention. *Vision Research*, *43*(9), 1035–1045. https://doi.org/10.1016/S0042-6989(03)00084-1

Engbert, R., Sinn, P., Mergenthaler, K., & Trukenbrod, H. (2015) Microsaccade Toolbox. http://read.psych.unipotsdam.de/index.php?option=com_content&view=article&id=140:engbert-et-al-2015-microsaccade-toolbox-for-r&catid=26:publications&Itemid=34

Engström, D. A., Kelso, J. A. S., & Holroyd, T. (1996). Reaction–anticipation transitions in human perception–action patterns. *Human Movement Science, 15*, 809-832

Ergenoglu, T., Demiralp, T., Bayraktaroglu, Z., Ergen, M., Beydagi, H., & Uresin, Y. (2004). Alpha rhythm of the EEG modulates visual detection performance in humans. *Cognitive Brain Research*, *20*(3), 376–383. https://doi.org/10.1016/j.cogbrainres.2004.03.009

Ermentrout, G. B., Galán, R. F., & Urban, N. N. (2008). Reliability, synchrony and noise. *Trends in Neurosciences*, *31*(8), 428–434. https://doi.org/10.1016/j.tins.2008.06.002

Everling, S., Krappmann, P., Spantekow, A., & Flohr, H. (1997). Influence of pre-target cortical potentials on saccadic reaction times. *Experimental Brain Research*, *115*(3), 479–484. https://doi.org/10.1007/PL00005717

Farrell, S., Wagenmakers, E.-J., & Ratcliff, R. (2006). 1/f noise in human cognition: Is it ubiquitous, and what does it mean? *Psychonomic bulletin & review, 13*(4), 737–741. https://doi.org/10.3758/BF03193989

Faisal, A. A., Selen, L. P. J., & Wolpert, D. M. (2008). Noise in the nervous system. *Nature Reviews Neuroscience*, *9*(4), 292–303. https://doi.org/10.1038/nrn2258

Fox, K. C. R., Spreng, R. N., Ellamil, M., Andrews-Hanna, J. R., & Christoff, K. (2015). The wandering brain: Meta-analysis of functional neuroimaging studies of mind-wandering and related spontaneous thought processes. *NeuroImage*, *111*, 611–621. https://doi.org/10.1016/j.neuroimage.2015.02.039

Foxe, J. J., Simpson, G. V., & Ahlfors, S. P. (1998). Parieto-occipital ~10Hz activity reflects anticipatory state of visual attention mechanisms. NeuroReport, 9(17), 3929. https://doi.org/10.1097/00001756-199812010-00030

Fraley, C., Leisch, F., Maechler, M., Reisen, V., & Lemonte, A. (2006). Fracdiff: Fractionally differenced ARIMA aka ARFIMA(p,d,q) models. R package version 1.3-0. Avaliable at: https://CRAN.R-project.org/package=fracdiff

Franklin, M. S., Smallwood, J., & Schooler, J. W. (2011). Catching the mind in flight: Using behavioral indices to detect mindless reading in real time. *Psychonomic Bulletin & Review*, *18*(5), 992–997. https://doi.org/10.3758/s13423-011-0109-6

Fried, M., Tsitsiashvili, E., Bonneh, Y. S., Sterkin, A., Wygnanski-Jaffe, T., Epstein, T., & Polat, U. (2014). ADHD subjects fail to suppress eye blinks and microsaccades while anticipating visual stimuli but recover with medication. *Vision Research*, *101*, 62–72. https://doi.org/10.1016/j.visres.2014.05.004

Fu, K.-M. G., Foxe, J. J., Murray, M. M., Higgins, B. A., Javitt, D. C., & Schroeder, C. E. (2001). Attention-dependent suppression of distracter visual input can be

cross-modally cued as indexed by anticipatory parieto–occipital alpha-band oscillations. *Cognitive Brain Research*, *12*(1), 145–152. https://doi.org/10.1016/S0926-6410(01)00034-9

Geurts, H. M., Grasman, R. P. P. P., Verté, S., Oosterlaan, J., Roeyers, H., van Kammen, S. M., & Sergeant, J. A. (2008). Intra-individual variability in ADHD, autism spectrum disorders and Tourette's syndrome. *Neuropsychologia*, *46*(13), 3030–3041. https://doi.org/10.1016/j.neuropsychologia.2008.06.013

Giambra, L. M. (1980). Sex Differences in Daydreaming and Related Mental Activity from the Late Teens to the Early Nineties. *The International Journal of Aging and Human Development*, *10*(1), 1–34. https://doi.org/10.2190/01BD-RFNE-W34G-9ECA

Giambra, L. M. (1995). A Laboratory Method for Investigating Influences on Switching Attention to Task-Unrelated Imagery and Thought. *Consciousness and Cognition*, *4*(1), 1–21. https://doi.org/10.1006/ccog.1995.1001

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, *102*, 74–78. https://doi.org/10.1016/j.paid.2016.06.069

Gilden, D. L. (2001). Cognitive Emissions of 1/f Noise. *Psychological Review*, *108*(1), 33–56. https://doi.org/10.I037//0033-295X.108.1.33

Gilden, D. L., & Hancock, H. (2007). Response Variability in Attention-Deficit Disorders. *Psychological Science*, *18*(9), 796–802. https://doi.org/10.1111/j.1467-9280.2007.01982.x

Gilden, D. L., Thornton, T., & Mallon, M. W. (1995). 1/f noise in human cognition. *Science*, *267*(5205), 1837–1839. https://doi.org/10.1126/science.7892611

Gilden, D. L., & Wilson, S. G. (1995). Streaks in skilled performance*. Psychonomic Bulletin & Review, 2*(2), 260-265. https://doi.org/10.3758/BF03210967

Gonzalez Andino, S. L., Michel, C. M., Thut, G., Landis, T., & Grave de Peralta, R. (2005). Prediction of response speed by anticipatory high-frequency (gamma band) oscillations in the human brain. *Human Brain Mapping, 24*(1), 50-58. https://doi.org/10.1002/hbm.20056

de Graaf, T. A. de, Gross, J., Paterson, G., Rusch, T., Sack, A. T., & Thut, G. (2013). Alpha-Band Rhythms in Visual Task Performance: Phase-Locking by Rhythmic Sensory Stimulation. *PLOS ONE*, *8*(3), e60035. https://doi.org/10.1371/journal.pone.0060035

Gruberger, M., Simon, E. B., Levkovitz, Y., Zangen, A., & Hendler, T. (2011). Towards a Neuroscience of Mind-Wandering. *Frontiers in Human Neuroscience*, *5*. https://doi.org/10.3389/fnhum.2011.00056

Haggard, P. (2008). Human volition: towards a neuroscience of will. *Nature Reviews Neuroscience*, *9*(12), 934–946. https://doi.org/10.1038/nrn2497

Hamm, J. P., Dyckman, K. A., Ethridge, L. E., McDowell, J. E., & Clementz, B. A. (2010). Preparatory Activations across a Distributed Cortical Network Determine Production of Express Saccades in Humans. *Journal of Neuroscience*, *30*(21), 7350–7357. https://doi.org/10.1523/JNEUROSCI.0785-10.2010

Hamm, J. P., Sabatinelli, D., & Clementz, B. A. (2012). Alpha oscillations and the control of voluntary saccadic behavior. *Experimental Brain Research, 221*(2), 123-128, https://doi.org/10.1007/s00221-012-3167-8

Hanslmayr, S., Aslan, A., Staudigl, T., Klimesch, W., Herrmann, C. S., & Bäuml, K.-H. (2007). Prestimulus oscillations predict visual perception performance between and within subjects. *NeuroImage*, *37*(4), 1465–1473. https://doi.org/10.1016/j.neuroimage.2007.07.011

He, B. J. (2013). Spontaneous and Task-Evoked Brain Activity Negatively Interact. *Journal of Neuroscience*, *33*(11), 4672–4682. https://doi.org/10.1523/JNEUROSCI.2922-12.2013

He, B. J., & Zempel, J. M. (2013). Average Is Optimal: An Inverted-U Relationship between Trial-to-Trial Brain Activity and Behavioral Performance. *PLoS Computational Biology*, *9*(11). https://doi.org/10.1371/journal.pcbi.1003348

Hommel, B., Chapman, C. S., Cisek, P., Neyedli, H. F., Song, J.-H., & Welsh, T. N. (2019). No one knows what attention is. *Attention, Perception, & Psychophysics, 81*(7), 2288-2303. https://doi.org/10.3758/s13414-019-01846-w

Huang, M. X., Mosher, J. C., & Leahy, R. M. (1999). A sensor-weighted overlapping-sphere head model and exhaustive head model comparison for MEG. *Physics in Medicine and Biology*, *44*(2), 423–440. https://doi.org/10.1088/0031-9155/44/2/010

Huber, M. E., Kuznetsov, N., & Sternad, D. (2016). Persistence of reduced neuromotor noise in long-term motor skill learning. *Journal of Neurophysiology, 116*(6), 2922-2935. https://doi.org/ 10.1152/jn.00263.2016

Hultsch, D. F., MacDonald, S. W. S., & Dixon, R. A. (2002). Variability in reaction time performance of younger and older adults. *The Journals of Gerontology: Series B*, *57*(2), P101–P115. https://doi.org/10.1093/geronb/57.2.P101

Hultsch, D. F., MacDonald, S. W. S., Hunter, M. A., Levy-Bencheton, J., & Strauss, E. (2000). Intraindividual variability in cognitive performance in older adults: Comparison of adults with mild dementia, adults with arthritis, and healthy adults. *Neuropsychology*, *14*(4), 588–598. https://doi.org/10.1037/0894-4105.14.4.588

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L.,O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E., & Yasmeen, F. (2018). Forecast: Forecasting functions for time series and linear models. R package version 8.4. Available: http://pkg.robjhyndman.com/forecast.

Hyndman, R.J., Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software, 26*(3), 1-22. https://doi.org/10.18637/jss.v027.i03

JASP Team (2017). JASP (Version 0.8.5).

Jin, C. Y., Borst, J. P., & van Vugt, M. K. (2019). Predicting task-general mind-wandering with EEG. *Cognitive, Affective, & Behavioral Neuroscience*. https://doi.org/10.3758/s13415-019-00707-1

Johnson, K. A., Kelly, S. P., Bellgrove, M. A., Barry, E., Cox, M., Gill, M., & Robertson, I. H. (2007). Response variability in Attention Deficit Hyperactivity Disorder: Evidence for neuropsychological heterogeneity. *Neuropsychologia*, *45*(4), 630–638. https://doi.org/10.1016/j.neuropsychologia.2006.03.034

Irrmischer, M., van der Wal, C. N., Mansvelder, H. D., & Linkenkaer-Hansen, K. (2018). Negative mood and mind wandering increase long-range temporal correlations in attention fluctuations. *PLOS ONE*, *13*(5), e0196907. https://doi.org/10.1371/journal.pone.0196907

Kaiser, S., Roth, A., Rentrop, M., Friederich, H.-C., Bender, S., & Weisbrod, M. (2008). Intra-individual reaction time variability in schizophrenia, depression and borderline personality disorder. *Brain and Cognition, 66*(1), 73–82. https://doi.org/10.1016/j.bandc.2007.05.007

Kam, J. W. Y., Dao, E., Blinn, P., Krigolson, O. E., Boyd, L. A., & Handy, T. C. (2012). Mind wandering and motor control: Off-task thinking disrupts the online adjustment of behavior. *Frontiers in Human Neuroscience*, *6*. https://doi.org/10.3389/fnhum.2012.00329

Kam, J. W. Y., Dao, E., Farley, J., Fitzpatrick, K., Smallwood, J., Schooler, J. W., & Handy, T. C. (2011). Slow Fluctuations in Attentional Control of Sensory Cortex. *Journal of Cognitive Neuroscience*, *23*(2), 460–470. https://doi.org/10.1162/jocn.2010.21443

Kantelhardt, J. W., Koscielny-Bunde, E., Rego, H. H. A., Havlin, S., & Bunde, A. (2001). Detecting long-range correlations with detrended fluctuation analysis. *Physica A: Statistical Mechanics and Its Applications*, *295*(3), 441–454. https://doi.org/10.1016/S0378-4371(01)00144-3

Karalunas, S. L., Geurts, H. M., Konrad, K., Bender, S., & Nigg, J. T. (2014). Annual Research Review: Reaction time variability in ADHD and autism spectrum disorders: measurement and mechanisms of a proposed trans-diagnostic phenotype. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *55*(6), 685–710. https://doi.org/10.1111/jcpp.12217

Karalunas, S. L., Huang-Pollock, C. L., & Nigg, J. T. (2013). Is reaction time variability in ADHD mainly at low frequencies? *Journal of Child Psychology and Psychiatry*, *54*(5), 536–544. https://doi.org/10.1111/jcpp.12028

Kelly, A., Heathcote, A., Heath, R., & Longstaff, M. (2001). Response-Time Dynamics: Evidence for Linear and Low-Dimensional Nonlinear Structure in

Human Choice Sequences. *The Quarterly Journal of Experimental Psychology Section A*, *54*(3), 805–840. https://doi.org/10.1080/713755987

Kelly, A. M. C., Uddin, L. Q., Biswal, B. B., Castellanos, F. X., & Milham, M. P. (2008). Competition between functional brain networks mediates behavioral variability. *NeuroImage*, *39*(1), 527–537. https://doi.org/10.1016/j.neuroimage.2007.08.008

Kessler, R. C., Adler, L., Ames, M., Demler, O., Faraone, S., Hiripi, E., … Walters, E. E. (2005). The World Health Organization adult ADHD self-report scale (ASRS): a short screening scale for use in the general population. *Psychological Medicine*, *35*(2), 245–256. https://doi.org/10.1017/S0033291704002892

Kleiner, M., Brainard, D. H., & Pelli, D. (2007). What's new in Psychtoolbox-3? *Perception*, *36*(14), 1–16. https://doi.org/10.1068/v070821

Koch, C. (2009). Free Will, Physics, Biology, and the Brain. In N. Murphy, G. R. R. Ellis, & T. O'Connor (Red.), *Downward Causation and the Neurobiology of Free Will* (pp. 31–52). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-03205-9_2

Kofler, M. J., Rapport, M. D., Sarver, D. E., Raiker, J. S., Orban, S. A., Friedman, L. M., & Kolomeyer, E. G. (2013). Reaction time variability in ADHD: a meta-analytic review of 319 studies. *Clinical Psychology Review*, *33*(6), 795–811. https://doi.org/10.1016/j.cpr.2013.06.001

Kontsevich, L. L., & Tyler, C. W. (1999). Bayesian adaptive estimation of psychometric slope and threshold. *Vision Research*, *39*(16), 2729–2737. https://doi.org/10.1016/S0042-6989(98)00285-5

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Krzemiński, D., Kamiński, M., Marchewka, A., & Bola, M. (2017). Breakdown of long-range temporal correlations in brain oscillations during general anesthesia. *NeuroImage*, *159*, 146–158. https://doi.org/10.1016/j.neuroimage.2017.07.047

Kucyi, A., Esterman, M., Riley, C. S., & Valera, E. M. (2016). Spontaneous default network activity reflects behavioral variability independent of mind-wandering. *Proceedings of the National Academy of Sciences*, *113*(48), 13899–13904. https://doi.org/10.1073/pnas.1611743113

Langdown, B. L., Bridge, M. & Li, F.-X. (2012). Movement variability in the golf swing. *Sports Biomechanics, 11*(2), 273-287. https://doi.org/10.1080/14763141.2011.650187

Laflamme, P., Seli, P., & Smilek, D. (2018). Validating a visual version of the metronome response task. *Behavior Research Methods*, 1–12. https://doi.org/10.3758/s13428-018-1020-0

Lau, M. A., Bishop, S. R., Segal, Z. V., Buis, T., Anderson, N. D., Carlson, L., … Devins, G. (2006). The Toronto Mindfulness Scale: development and validation. *Journal of Clinical Psychology*, *62*(12), 1445–1467. https://doi.org/10.1002/jclp.20326

Ledberg, A., Montagnini, A., Coppola, R., & Bressler, S. L. (2012). Reduced Variability of Ongoing and Evoked Cortical Activity Leads to Improved Behavioral Performance. *PLOS ONE*, *7*(8), e43166. https://doi.org/10.1371/journal.pone.0043166

Lemoine, L., Torre, K., & Delignières, D. (2006). Testing for the presence of 1/f noise in continuation tapping data. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *60*(4), 247–257. https://doi.org/10.1037/cjep2006023

Lynam DR, Smith GT, Whiteside SP, Cyders MA. *The UPPS-P: Assessing five personality pathways to impulsive behavior* (Technical Report) West Lafayette: Purdue University; 2006.

Macdonald, J. S. P., Mathan, S., & Yeung, N. (2011). Trial-by-Trial Variations in Subjective Attentional State are Reflected in Ongoing Prestimulus EEG Alpha Oscillations. *Frontiers in Psychology*, *2*. https://doi.org/10.3389/fpsyg.2011.00082

Madison, G. (2004). Fractal modeling of human isochronous serial interval production. *Biological Cybernetics*, *90*(2), 105–112. https://doi.org/10.1007/s00422-003-0453-3

Marchant, D. C., Clough, P. J., & Crawshaw, M. (2007). The effects of attentional focusing strategies on novice dart throwing performance and Their task experiences. *International Journal of Sport and Exercise Psychology*, *5*(3), 291–303. https://doi.org/10.1080/1612197X.2007.9671837

Marchant, D. C., Clough, P. J., Crawshaw, M., & Levy, A. (2009). Novice motor skill performance and task experience is influenced by attentional focusing instructions and instruction preferences. *International Journal of Sport and Exercise Psychology*, *7*(4), 488–502. https://doi.org/10.1080/1612197X.2009.9671921

Marom, S., & Wallach, A. (2011). Relational dynamics in perception: impacts on trial-to-trial variation. *Frontiers in Computational Neuroscience*, *5*, 16. https://doi.org/10.3389/fncom.2011.00016

Marszalek, J. M., Barber, C., Kohlhart, J., & Holmes, C. B. (2011). Sample size in psychological research over the past 30 years. *Perceptual and Motor Skills*, *112*(2), 331–348. https://doi.org/10.2466/03.11.PMS.112.2.331-348

Martinez-Conde, S., Otero-Millan, J., & Macknik, S. L. (2013). The impact of microsaccades on vision: towards a unified theory of saccadic function. *Nature Reviews Neuroscience*, *14*(2), 83–96. https://doi.org/10.1038/nrn3405

Mason, M. F., Norton, M. I., van Horn, J. D., Wegner, D. M., Grafton, S. T., & Macrae, C. N. (2007). Wandering Minds: The Default Network and Stimulus-Independent Thought. *Science*, *315*(5810), 393–395. https://doi.org/10.1126/science.1131295

Masquelier, T. (2013). Neural variability, or lack thereof. *Frontiers in Computational Neuroscience*, *7*, 1–7. https://doi.org/10.3389/fncom.2013.00007

Mayeux, R. (2004). Biomarkers: potential uses and limitations. *NeuroRx: The Journal of the American Society for Experimental NeuroTherapeutics*, *1*(2), 182–188. https://doi.org/10.1602/neurorx.1.2.182

Miyake, Y., Onishi, Y. & Pöppel, E. (2004). Two types of anticipation in synchronization tapping. *Acta Neurobiologiae Experimentalis, 64*(3), 415-426.

Mazaheri, A., Nieuwenhuis, I. L. C., Dijk, H. van, & Jensen, O. (2009). Prestimulus alpha and mu activity predicts failure to inhibit motor responses. Human Brain Mapping, 30(6), 1791–1800. https://doi.org/10.1002/hbm.20763

McAuley, E., Wraith, S., & Duncan, T. E. (1991). Self-Efficacy, Perceptions of Success, and Intrinsic Motivation for Exercise1. *Journal of Applied Social Psychology*, *21*(2), 139–155. https://doi.org/10.1111/j.1559-1816.1991.tb00493.x

McClelland, G. H. (2000). Increasing statistical power without increasing sample size. *American Psychologist*, *55*(8), 963–964. https://doi.org/10.1037/0003-066X.55.8.963

McVay, J. C., & Kane, M. J. (2012). Drifting from Slow to “D’oh!” Working Memory Capacity and Mind Wandering Predict Extreme Reaction Times and Executive-Control Errors. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *38*(3), 525–549. https://doi.org/10.1037/a0025896

Miller, D. J., Derefinko, K. J., Lynam, D. R., Milich, R., & Fillmore, M. T. (2010). Impulsivity and Attention Deficit-Hyperactivity Disorder: subtype classification using the UPPS Impulsive Behavior Scale. *Journal of Psychopathology and Behavioral Assessment*, *32*(3), 323–332. https://doi.org/10.1007/s10862-009-9155-z

Miller, J., Sproesser, G., & Ulrich, R. (2008). Constant versus variable response signal delays in speed-accuracy trade-offs: Effects of advance preparation for processing time. *Perception & Psychophysics*, *70*(5), 878–886. https://doi.org/10.3758/PP.70.5.878

Mitchell, J. F., Sundberg, K. A., & Reynolds, J. H. (2007). Differential Attention-Dependent Response Modulation across Cell Classes in Macaque Visual Area V4. *Neuron*, *55*(1), 131–141. https://doi.org/10.1016/j.neuron.2007.06.018

Mittner, M., Hawkins, G. E., Boekel, W., & Forstmann, B. U. (2016). A Neural Model of Mind Wandering. *Trends in Cognitive Sciences*, *20*(8), 570–578. https://doi.org/10.1016/j.tics.2016.06.004

Mittner, M., Boekel, W., Tucker, A. M., Turner, B. M., Heathcote, A., & Forstmann, B. U. (2014). When the Brain Takes a Break: A Model-Based Analysis of Mind Wandering. *The Journal of Neuroscience*, *34*(49), 16286–16295. https://doi.org/10.1523/JNEUROSCI.2062-14.2014

Mo, J., Liu, Y., Huang, H., & Ding, M. (2013). Coupling between visual alpha oscillations and default mode activity. *NeuroImage*, *68*, 112–118. https://doi.org/10.1016/j.neuroimage.2012.11.058

Moerel, M., De Martino, F., & Formisano, E. (2014). An anatomical and functional topography of human auditory cortical areas. *Frontiers in Neuroscience*, *8*. https://doi.org/10.3389/fnins.2014.00225

Monto, S., Palva, S., Voipio, J., & Palva, J. M. (2008). Very Slow EEG Fluctuations Predict the Dynamics of Stimulus Detection and Oscillation Amplitudes in Humans. *Journal of Neuroscience,* *28*(33), 8268–8272. https://doi.org/10.1523/JNEUROSCI.1910-08.2008

Morrison, A. B., Goolsarran, M., Rogers, S. L., & Jha, A. P. (2014). Taming a wandering attention: short-form mindfulness training in student cohorts. *Frontiers in Human Neuroscience*, *7*. https://doi.org/10.3389/fnhum.2013.00897

Mrazek, M. D., Smallwood, J., & Schooler, J. W. (2012). Mindfulness and mind-wandering: finding convergence through opposing constructs. *Emotion, 12(*3), 442–448. https://doi.org/10.1037/a0026678

Neumann, D. L., & Piercy, A. (2013). The Effect of Different Attentional Strategies on Physiological and Psychological States During Running. *Australian Psychologist*, *48*(5), 329–337. https://doi.org/10.1111/ap.12015

Otero-Millan, J., Macknik, S. L., Langston, R. E., & Martinez-Conde, S. (2013). An oculomotor continuum from exploration to fixation. *Proceedings of the National Academy of Sciences*, *110*(15), 6175–6180. https://doi.org/10.1073/pnas.1222715110

Otero-Millan, Jorge, Troncoso, X. G., Macknik, S. L., Serrano-Pedraza, I., & Martinez-Conde, S. (2008). Saccades and microsaccades during visual fixation, exploration, and search: Foundations for a common saccadic generator. *Journal of Vision*, *8*(14), 21–21. https://doi.org/10.1167/8.14.21

Panagiotidi, M., Overton, P., & Stafford, T. (2017). Increased microsaccade rate in individuals with ADHD traits. *Journal of Eye Movement Research*, *10*(1). https://doi.org/10.16910/10.1.6

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442. https://doi.org/10.1163/156856897X00366

Peng, C. -K., Havlin, S., Stanley, H. E., & Goldberger, A. L. (1995). Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *5*(1), 82–87. https://doi.org/10.1063/1.166141

Perfetti, B., Moisello, C., Landsness, E. C., Kvint, S., Pruski, A., Onofrj, M., Tononi, G., & Ghilardi, M. F. (2011). Temporal evolution of oscillatory activity predicts performance in a choice-reaction time reaching task. *Journal of Neurophysiology, 105*(1), 18-27. https://doi.org/10.1152/jn.00778.2010

Poynter, W., Barber, M., Inman, J., & Wiggins, C. (2013). Individuals exhibit idiosyncratic eye-movement behavior profiles across tasks. *Vision Research*, *89*, 32–38. https://doi.org/10.1016/j.visres.2013.07.002

Qin, J., Perdoni, C., & He, B. (2011). Dissociation of Subjectively Reported and Behaviorally Indexed Mind Wandering by EEG Rhythmic Activity. *PLoS ONE*, *6*(9), e23124. https://doi.org/10.1371/journal.pone.0023124

R Core Team (2013). R: A language and environment for statistical computing. R Foundation or Statistical Computing, Vienna, Austria. URL http://www.R-project.org

Rabbitt, P. M. (1966). Errors and error correction in choice-response tasks. *Journal of Experimental Psychology*, *71*(2), 264–272. https://doi.org/10.1037/h0022853

Radlo, S. J., Steinberg, G. M., Singer, R. N., Barba, D. A., & Melnikov, A. (2002). The influence of an attentional focus strategy on alpha brain wave activity, heart rate, and dart-throwing performance. *International Journal of Sport Psychology*, *33*(2), 205–217.

Ratcliff, R. (1978). A theory of memory retrieval. Psychological Review, 85(2), 59-108. https://doi.org/10.1037/0033-295X.85.2.59

van Ravenzwaaij, D., Donkin, C., & Vandekerckhove, J. (2017). The EZ diffusion model provides a powerful test of simple empirical effects. *Psychonomic Bulletin & Review, 24*(2), 547-556. https://doi.org/10.3758/s13423-016-1081-y

van Ravenzwaaij, D., & Oberauer, K. (2009). How to use the diffusion model: Parameter recovery of three methods: EZ, fast-dm, and DMAT. *Journal of Mathematical Psychology, 53*(6), 463-473. https://doi.org/10.1016/j.jmp.2009.09.004

Rayner, K., Li, X., Williams, C. C., Cave, K. R., & Well, A. D. (2007). Eye movements during information processing tasks: Individual differences and cultural effects. *Vision Research*, *47*(21), 2714–2726. https://doi.org/10.1016/j.visres.2007.05.007

Repp, B. H. & Doggett, R. (2007). Tapping to a very slow beat: A comparison of musicians and nonmusicians. *Music Perception, 24*(4), 367-376. https://doi.org/:10.1525/MP.2007.24.4.367

Reuter, M., Kirsch, P., & Hennig, J. (2006). Inferring candidate genes for Attention Deficit Hyperactivity Disorder (ADHD) assessed by the World Health Organization Adult ADHD Self-Report Scale (ASRS). *Journal of Neural Transmission*, *113*(7), 929–938. https://doi.org/10.1007/s00702-005-0366-5

Richter, C. G., Babo-Rebelo, M., Schwartz, D., & Tallon-Baudry, C. (2017). Phase-amplitude coupling at the organism level: The amplitude of spontaneous alpha rhythm fluctuations varies with the phase of the infra-slow gastric basal rhythm. *NeuroImage*, *146*, 951–958. https://doi.org/10.1016/j.neuroimage.2016.08.043

Rihs, T. A., Michel, C. M., & Thut, G. (2007). Mechanisms of selective inhibition in visual spatial attention are indexed by α-band EEG synchronization. *European Journal of Neuroscience*, *25*(2), 603–610. https://doi.org/10.1111/j.1460-9568.2007.05278.x

Robinson, S. E., & Vrba, J. (1999). Functional neuroimaging by synthetic aperture magnetometry. *Functional Neuroimaging by Synthetic Aperture Magnetometry (SAM)*, 302–305. Retrieved from Scopus.

Robison, M. K., Miller, A. L., & Unsworth, N. (2019). Examining the effects of probe frequency, response options, and framing within the thought-probe method. *Behavior Research Methods*, *51*(1), 398–408. https://doi.org/10.3758/s13428-019-01212-6

Rolfs, M. (2009). Microsaccades: Small steps on a long way. *Vision Research*, *49*(20), 2415–2441. https://doi.org/10.1016/j.visres.2009.08.010

Rolfs, M. (2009). Microsaccades: Small steps on a long way. *Vision Research*, *49*(20), 2415–2441. https://doi.org/10.1016/j.visres.2009.08.010

Romei, V., Gross, J., & Thut, G. (2010). On the Role of Prestimulus Alpha Rhythms over Occipito-Parietal Areas in Visual Input Regulation: Correlation or Causation? *Journal of Neuroscience*, *30*(25), 8692–8697. https://doi.org/10.1523/JNEUROSCI.0160-10.2010

Romei, V., Rihs, T., Brodbeck, V., & Thut, G. (2008). Resting electroencephalogram alpha-power over posterior sites indexes baseline visual cortex excitability: *NeuroReport*, *19*(2), 203–208. https://doi.org/10.1097/WNR.0b013e3282f454c4

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*(2), 301–308. https://doi.org/10.3758/s13423-014-0595-4

Rouder, J. N., & Haaf, J. M. (2018). Power, dominance, and constraint: A note on the appeal of different design traditions. *Advances in Methods and Practices in Psychological Science*, *1*(1), 19–26. https://doi.org/10.1177/2515245917745058

Salomon, R., Ronchi, R., Dönz, J., Bello-Ruiz, J., Herbelin, B., Martet, R., Faivre, N., Schaller, K., & Blanke, O. (2016). The Insula Mediates Access to Awareness of Visual Stimuli Presented Synchronously to the Heartbeat. *Journal of Neuroscience*, *36*(18), 5115–5127. https://doi.org/10.1523/JNEUROSCI.4262-15.2016

Salthouse, T. A. (2012). Psychometric properties of within-person across-session variability in accuracy of cognitive performance. *Assessment*, *19*(4), 494–501. https://doi.org/10.1177/1073191112438744

Saville, C. W. N., Pawling, R., Trullinger, M., Daley, D., Intriligator, J., & Klein, C. (2011). On the stability of instability: Optimising the reliability of intra-subject

variability of reaction times. *Personality and Individual Differences*, *51*(2), 148–153. https://doi.org/10.1016/j.paid.2011.03.034

Saville, C. W. N., Shikhare, S., Iyengar, S., Daley, D., Intriligator, J., Boehm, S. G., … Klein, C. (2012). Is reaction time variability consistent across sensory modalities? Insights from latent variable analysis of single-trial P3b latencies. *Biological Psychology*, *91*(2), 275–282. https://doi.org/10.1016/j.biopsycho.2012.07.006

Schmiedek, F., Oberauer, K., Wilhelm, O., Süss, H. M., & Wittmann, W. W. (2007). Individual differences in components of reaction time distributions and their relations to working memory and intelligence. *Journal of Experimental Psychology. General*, *136*(3), 414–429. https://doi.org/10.1037/0096-3445.136.3.414

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, *47*(5), 609–612. https://doi.org/10.1016/j.jrp.2013.05.009

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*, *22*(2), 322–339. https://doi.org/10.1037/met0000061

Schooler, J. W., Reichle, E. D., & Halpern, D. V. (2004). Zoning Out while Reading: Evidence for Dissociations between Experience and Metaconsciousness. In D. T. Levin (Red.), *Thinking and seeing: visual metacognition in adults and children*. Cambridge, Mass: MIT Press.

Schurger, A., Sarigiannidis, I., Naccache, L., Sitt, J. D., & Dehaene, S. (2015). Cortical activity is more stable when sensory stimuli are consciously perceived. *Proceedings of the National Academy of Sciences*, *112*(16), E2083–E2092. https://doi.org/10.1073/pnas.1418730112

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Sedlmeier, P.& Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *105*(2), 309–316. https://doi.org/10.1037/0033-2909.105.2.309

Seli, P., Cheyne, J. A., & Smilek, D. (2013). Wandering minds and wavering rhythms: linking mind wandering and behavioral variability. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(1), 1–5. https://doi.org/10.1037/a0030954

Seli, P., Smallwood, J., Cheyne, J. A., & Smilek, D. (2015). On the relation of mind wandering and ADHD symptomatology. *Psychonomic Bulletin & Review*, *22*(3), 629–636. https://doi.org/10.3758/s13423-014-0793-0

Shahan, T. A., & Chase, P. N. (2002). Novelty, stimulus control, and operant variability. *The Behavior Analyst*, *25*(2), 175–190. https://doi.org/10.1007/BF03392056

Shaw, G. A., & Giambra, L. (1993). Task-unrelated thoughts of college students diagnosed as hyperactive in childhood. *Developmental Neuropsychology*, *9*(1), 17–30. https://doi.org/10.1080/87565649309540541

Shew, W. L., & Plenz, D. (2013). The Functional Benefits of Criticality in the Cortex. *The Neuroscientist*, *19*(1), 88–100. https://doi.org/10.1177/1073858412445487

Simola, J., Zhigalov, A., Morales-Muñoz, I., Palva, J. M., & Palva, S. (2017). Critical dynamics of endogenous fluctuations predict cognitive flexibility in the Go/NoGo task. *Scientific Reports*, *7*. https://doi.org/10.1038/s41598-017-02750-9

Singer, R. N. (2002). Preperformance State, Routines, and Automaticity: What Does It Take to Realize Expertise in Self-Paced Events? *Journal of Sport and Exercise Psychology*, *24*(4), 359–375. https://doi.org/10.1123/jsep.24.4.359

Singer, J. L., & Antrobus, J. S. (1963). A Factor-Analytic Study of Daydreaming and Conceptually-Related Cognitive and Personality Variables. *Perceptual and Motor Skills*, *17*(1), 187–209. https://doi.org/10.2466/pms.1963.17.1.187

Smallwood, J., Beach, E., Schooler, J. W., & Handy, T. C. (2008). Going AWOL in the Brain: Mind Wandering Reduces Cortical Analysis of External Events. *Journal*

of Cognitive Neuroscience, *20*(3), 458–469. https://doi.org/10.1162/jocn.2008.20037

Smallwood, J., Brown, K., Baird, B., & Schooler, J. W. (2012). Cooperation between the default mode network and the frontal–parietal network in the production of an internal train of thought. *Brain Research*, *1428*, 60–70. https://doi.org/10.1016/j.brainres.2011.03.072

Smallwood, J., Fitzgerald, A., Miles, L. K., & Phillips, L. H. (2009). Shifting moods, wandering minds: Negative moods lead the mind to wander. *Emotion*, *9*(2), 271–276. https://doi.org/10.1037/a0014855

Smallwood, J., & Schooler, J. W. (2015). The Science of Mind Wandering: Empirically Navigating the Stream of Consciousness. *Annual Review of Psychology*, *66*(1), 487–518. https://doi.org/10.1146/annurev-psych-010814-015331

Smeets, J. B. J., Frens, M. A., & Brenner, E. (2002). Throwing darts: timing is not the limiting factor. *Experimental Brain Research, 144,* 268-274.

Smallwood, J., McSpadden, M., & Schooler, J. W. (2007). The lights are on but no one's home: Meta-awareness and the decoupling of attention when the mind wanders. *Psychonomic Bulletin & Review*, *14*(3), 527–533. https://doi.org/10.3758/BF03194102

Smith, G. (2003). Horseshoe pitchers' hot hands. *Psychonomic Bulletin & Review, 10*(3), 753-758. https://doi.org/10.3758/BF03196542

Smith, P. L., & Little, D. R. (2018). Small is beautiful: In defense of the small-N design. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-018-1451-8

Stadnitski, T. (2012). Measuring Fractality. *Frontiers in Physiology*, *3*. https://doi.org/10.3389/fphys.2012.00127

Sternad, D. (2018). It's not (only) the mean that matters: variability, noise and exploration in skill learning. *Current Opinion in Behavioral Sciences*, *20*, 183–195. https://doi.org/10.1016/j.cobeha.2018.01.004

Stins, J. F., Yaari, G., Wijmer, K., Burger, J. F., & Beek, P. J. (2018). Evidence for Sequential Performance Effects in Professional Darts. *Frontiers in Psychology, 9,* 591. https://doi.org/10.3389/fpsyg.2018.00591

Takano, K. & Miyake, Y. (2007). Two types of phase correction mechanism involved in synchronized tapping. *Neuroscience Letters, 417,* 196-200. https://doi.org/10.1016/j.neulet.2007.02.044

Tales, A., Leonards, U., Bompas, A., Snowden, R. J., Philips, M., Porter, G., … Bayer, A. (2012). Intra-Individual Reaction Time Variability in Amnestic Mild Cognitive Impairment: A Precursor to Dementia? *Journal of Alzheimer's Disease, 32*(2), 457–466. *https://doi.org/10.3233/JAD-2012-120505*

Tamm, L., Narad, M. E., Antonini, T. N., O'Brien, K. M., Hawk, L. W., & Epstein, J. N. (2012). Reaction Time Variability in ADHD: A Review. *Neurotherapeutics*, *9*(3), 500–508. https://doi.org/10.1007/s13311-012-0138-5

Tauer, J. M., & Harackiewicz, J. M. (2004). The Effects of Cooperation and Competition on Intrinsic Motivation and Performance. *Journal of Personality and Social Psychology*, *86*(6), 849–861. https://doi.org/10.1037/0022-3514.86.6.849

The MathWorks, Inc. (Release 2016a). MATLAB 9. Natick, Massachusetts, United States.

Thome, J., Ehlis, A.-C., Fallgatter, A. J., Krauel, K., Lange, K. W., Riederer, P., … Gerlach, M. (2012). Biomarkers for attention-deficit/hyperactivity disorder (ADHD). A consensus report of the WFSBP task force on biological markers and the World Federation of ADHD. *The World Journal of Biological Psychiatry: The Official Journal of the World Federation of Societies of Biological Psychiatry*, *13*(5), 379–400. https://doi.org/10.3109/15622975.2012.690535

Thomson, D. R., Seli, P., Besner, D., & Smilek, D. (2014). On the link between mind wandering and task performance over time. *Consciousness and Cognition*, *27*, 14–26. https://doi.org/10.1016/j.concog.2014.04.001

Thornton, T. L., & Gilden, D. L. (2005). Provenance of correlations in psychological data. *Psychonomic Bulletin & Review*, *12*(3), 409–441. https://doi.org/10.3758/BF03193785

Thut, G., Nietzel, A., Brandt, S. A., & Pascual-Leone, A. (2006). α-Band Electroencephalographic Activity over Occipital Cortex Indexes Visuospatial Attention Bias and Predicts Visual Target Detection. *Journal of Neuroscience*, *26*(37), 9494–9502. https://doi.org/10.1523/JNEUROSCI.0875-06.2006

Toner, J., Montero, B. G., & Moran, A. (2015). Considering the role of cognitive control in expert performance. *Phenomenology and the Cognitive Sciences, 14*(4), 1127-1144. http://dx.doi.org/10.1007/s11097-014-9407-6

Torre, K., Balasubramaniam, R., Rheaume, N., Lemoine, L., & Zelaznik, H. N. (2011). Long-range correlation properties in motor timing are individual and task specific. *Psychonomic Bulletin & Review*, *18*(2), 339–346. https://doi.org/10.3758/s13423-011-0049-1

Tse, P. (2013). *The Neural Basis of Free Will: Criterial Causation*. MIT Press.

Tse, C.-S., Balota, D. A., Yap, M. J., Duchek, J. M., & McCabe, D. P. (2010). Effects of healthy aging and early stage dementia of the Alzheimer's type on components of response time distributions in three attention tasks. *Neuropsychology, 24*(3), 300–315. https://doi.org/10.1037/a0018274

Unsworth, N., & Robison, M. K. (2016). Pupillary correlates of lapses of sustained attention. *Cognitive, Affective, & Behavioral Neuroscience*, *16*(4), 601–615. https://doi.org/10.3758/s13415-016-0417-4

Unsworth, N., & Robison, M. K. (2018). Tracking arousal state and mind wandering with pupillometry. *Cognitive, Affective, & Behavioral Neuroscience*, *18*(4), 638–664. https://doi.org/10.3758/s13415-018-0594-4

Valsecchi, M., Betta, E., & Turatto, M. (2007). Visual oddballs induce prolonged microsaccadic inhibition. *Experimental Brain Research*, *177*(2), 196–208. https://doi.org/10.1007/s00221-006-0665-6

Van Orden, G. C., Holden, J. G., & Turvey, M. T. (2003). Self-organization of cognitive performance. *Journal of Experimental Psychology: General*, *132*(3), 331–350. https://doi.org/10.1037/0096-3445.132.3.331

VanRullen, R., Busch, N., Drewes, J., & Dubois, J. (2011). Ongoing EEG Phase as a Trial-by-Trial Predictor of Perceptual and Attentional Variability. *Frontiers in Psychology*, *2*, 1–9. https://doi.org/10.3389/fpsyg.2011.00060

van Veen, B. D., Drongelen, W. V., Yuchtman, M., & Suzuki, A. (1997). Localization of brain electrical activity via linearly constrained minimum variance spatial filtering. *IEEE Transactions on Biomedical Engineering*, *44*(9), 867–880. https://doi.org/10.1109/10.623056

Vrba, J., & Robinson, S. E. (2001). Signal Processing in Magnetoencephalography. *Methods*, *25*(2), 249–271. https://doi.org/10.1006/meth.2001.1238

Wagenmakers, E.-J., Farrell, S., & Ratcliff, R. (2004). Estimation and interpretation of 1/fα noise in human cognition. *Psychonomic Bulletin & Review*, *11*(4), 579–615. https://doi.org/10.3758/BF03196615

Wagenmakers, E.-J., Farrell, S., & Ratcliff, R. (2005). Human Cognition and a Pile of Sand: A Discussion on Serial Correlations and Self-Organized Criticality. *Journal of Experimental Psychology: General*, *134*(1), 108–116. https://doi.org/10.1037/0096-3445.134.1.108

Wagenmakers, E.-J., Maas, H. L. J. van der, & Farrell, S. (2012). Abstract Concepts Require Concrete Models: Why Cognitive Scientists Have Not Yet Embraced Nonlinearly Coupled, Dynamical, Self-Organized Critical, Synergistic, Scale-Free, Exquisitely Context-Sensitive, Interaction-Dominant, Multifractal, Interdependent Brain-Body-Niche Systems. *Topics in Cognitive Science*, *4*(1), 87–93. https://doi.org/10.1111/j.1756-8765.2011.01164.x

Wagenmakers, E.-J., Van Der Maas, H. L. J., & Grasman, R. P. P. P. (2007). An EZ-diffusion model for response time and accuracy. *Psychonomic Bulletin & Review*, *14*(1), 3–22. https://doi.org/10.3758/BF03194023

Ward, A. F., & Wegner, D. M. (2013). Mind-blanking: When the mind goes away. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00650

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063–1070.

Weinstein, Y. (2017). Mind-wandering, how do I measure thee with probes? Let me count the ways. *Behavior Research Methods*, *50*(2), 642–661. https://doi.org/10.3758/s13428-017-0891-9

Weissman, D. H., Roberts, K. C., Visscher, K. M., & Woldorff, M. G. (2006). The neural bases of momentary lapses in attention. *Nature Neuroscience, 9*(7), 971-978. https://doi.org/10.1038/nn1727

Wells, A. (2005). Detached Mindfulness In Cognitive Therapy: A Metacognitive Analysis And Ten Techniques. *Journal of Rational-Emotive & Cognitive-Behavior Therapy*, *23*(4), 337–355. https://doi.org/10.1007/s10942-005-0018-6

Wessel, J. R., & Aron, A. R. (2017). On the Globality of Motor Suppression: Unexpected Events and Their Influence on Behavior and Cognition. *Neuron*, *93*(2), 259–280. https://doi.org/10.1016/j.neuron.2016.12.013

Wetzels, R., & Wagenmakers, E.-J. (2012). A default Bayesian hypothesis test for correlations and partial correlations. *Psychonomic Bulletin & Review*, *19*(6), 1057–1064. https://doi.org/10.3758/s13423-012-0295-x

Whiteside, S. P., & Lynam, D. R. (2001). The Five Factor Model and impulsivity: using a structural model of personality to understand impulsivity. *Personality and Individual Differences*, *30*(4), 669–689. https://doi.org/10.1016/S0191-8869(00)00064-7

Willcutt, E. G., Doyle, A. E., Nigg, J. T., Faraone, S. V., & Pennington, B. F. (2005). Validity of the Executive Function Theory of Attention-Deficit/Hyperactivity Disorder: a meta-analytic review. *Biological Psychiatry*, *57*(11), 1336–1346. https://doi.org/10.1016/j.biopsych.2005.02.006

Winterstein, B. P., Silvia, P. J., Kwapil, T. R., Kaufman, J. C., Reiter-Palmon, R., & Wigert, B. (2011). Brief assessment of schizotypy: Developing short forms of the Wisconsin Schizotypy Scales. *Personality and Individual Differences*, *51*(8), 920–924. https://doi.org/10.1016/j.paid.2011.07.027

Zeidan, F., Johnson, S. K., Diamond, B. J., David, Z., & Goolkasian, P. (2010). Mindfulness meditation improves cognition: Evidence of brief mental training. *Consciousness and Cognition*, *19*(2), 597–605. https://doi.org/10.1016/j.concog.2010.03.014

Zhang, Y., Wang, X., Bressler, S. L., Chen, Y., & Ding, M. (2008). Prestimulus Cortical Activity is Correlated with Speed of Visuomotor Processing. *Journal of Cognitive Neuroscience*, *20*(10), 1915–1925. https://doi.org/10.1162/jocn.2008.20132