Examining temporal structures in reaction time

Marlou Nadine Perquin[1] & Aline Bompas[1]

1. CUBRIC, School of Psychology, Cardiff University, Maindy Road, CF24 4HQ, Cardiff, Wales, UK

Corresponding author:

Marlou Nadine Perquin

CUBRIC, Maindy Road

CF24 4HQ, Cardiff, Wales, UK

+44 (0)2920870365

perquinmn@cardiff.ac.uk

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, 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).

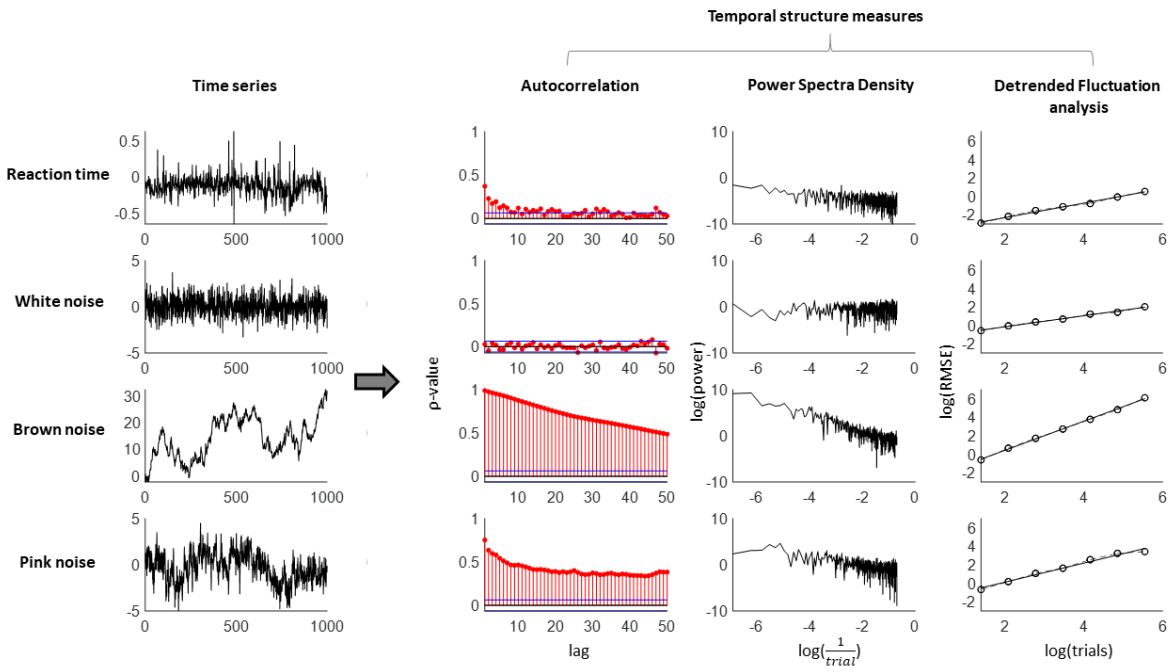

Figure 1. Examples of time series data over 1000 samples (left) and their corresponding temporal structures from lag 0 to lag 50. Shown are a reaction time series (showing small but clear temporal dependency), a white noise series (no temporal dependency), a brown noise series (i.e., a random walk, very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

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[1] – 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:

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

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

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.

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).

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.

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.[2]

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.

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

[2] 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.

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, 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 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.

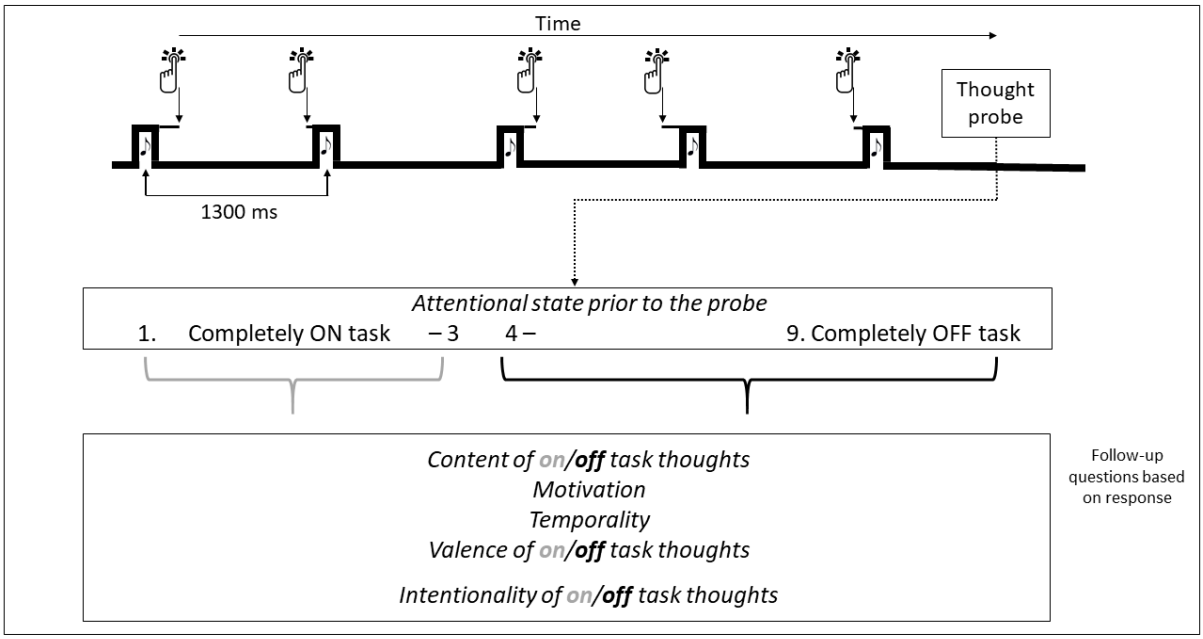

Figure 2. Overview of the task and thought probes.

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 τ = .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. Before running any analyses, we first tested whether the temporal dependency measures were statistically different from zero on the group level. We found extreme evidence for the presence of temporal dependency for nine out of (2 sessions * 6 measures) twelve measures, showing there is indeed substantial structure in RT on the group level (see Supplementary Materials for details).

Results

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.

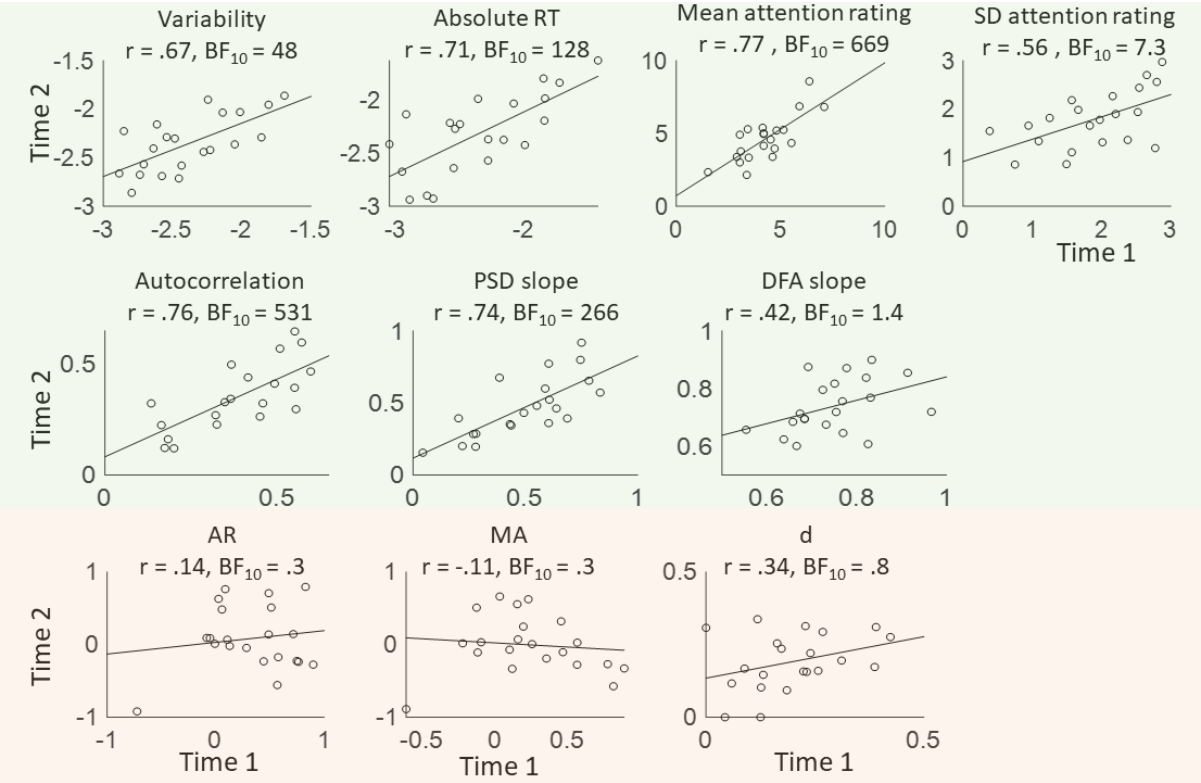

Figure 3. Within-subject correlations between MRT session 1 and 2 for performance (variability and mean absolute RT), subjective attentional state ratings (mean and variability), and temporal dependency measures (autocorrelation at lag 1, Power Spectrum Density (PSD) slope, Detrended Fluctuation Analysis (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.

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.

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.

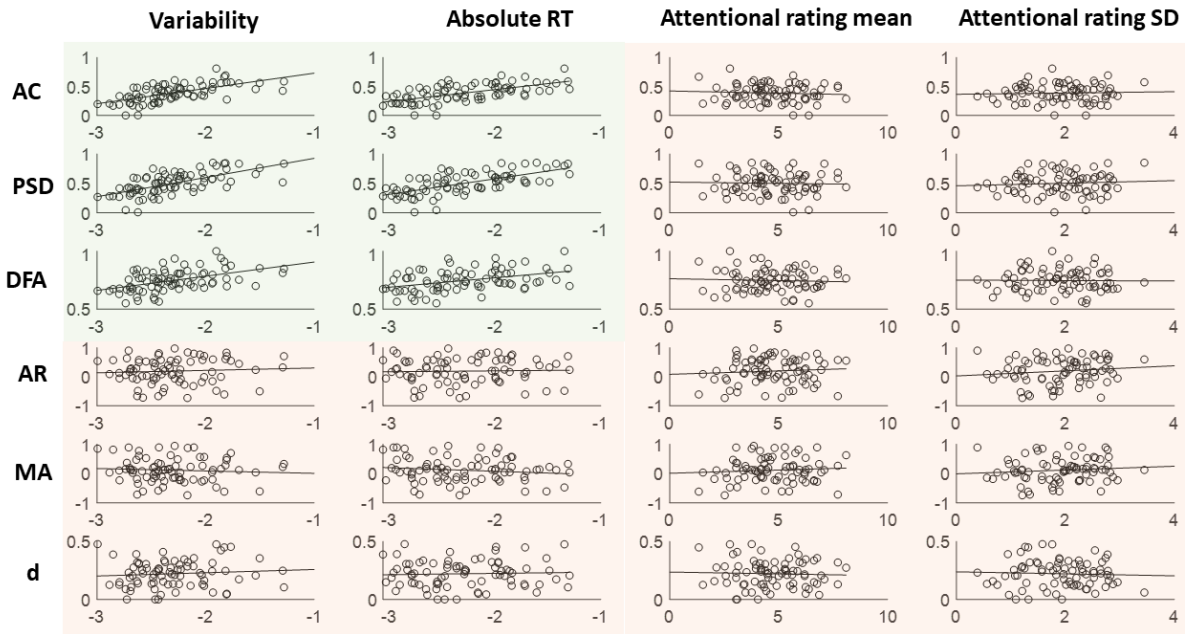

Figure 4. Between-subject correlations between performance and temporal dependency (left column for variability, and middle column for absolute RT), and subjective attentional state and temporal dependency (right column). Convention are the same as in Figure 3. Performance correlates moderately to strong 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 1. 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 1. 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.

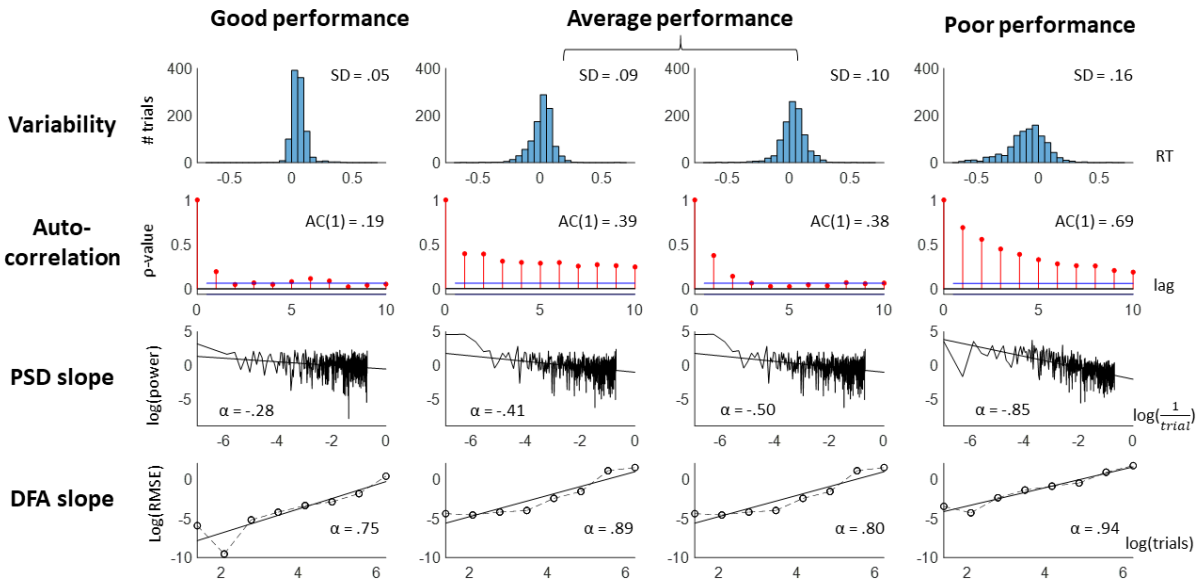

Figure 5. Examples from four participants over the temporal dependency measures, representing (from left to right) good, close-to-median, close-to-mean, and poor performance. Good performance was associated with low temporal dependencies, as reflected in low autocorrelation and shallow PSD and DFA slopes, 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 1. 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 1. 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.

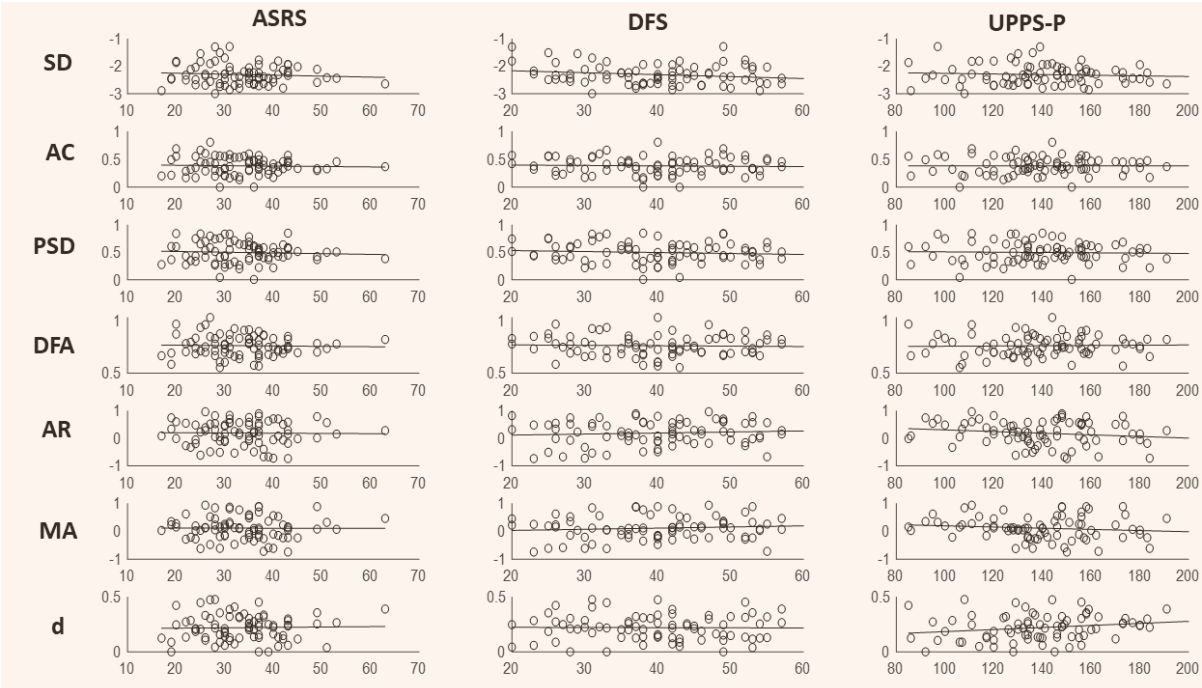

Figure 6. Between-subject correlations between the temporal dependency measures and self-assessed personality traits. Convention are the same as in Figure 3 and 4. None of the correlation pairs were supported.

Table 2. 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)

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

	PSD	**DFA**	**d**	**AR**	**MA**
Autocorr.	.94 (2.1e+35)	.82 (8.7e+16)	.30 (5.27)	.08 (.18)	-.09 (.19)
PSD slope	-	.71 (3.7e+10)	.14 (.31)	.16 (.37)	-.06 (.16)
DFA slope	-	-	.58 (8.4e+5)	-.11 (.23)	-.12 (.24)
d	-	-	-	-.27 (2.53)	-.01 (.14)
AR	-	-	-	-	.93 (5.6e+30)
MA	-	-	-	-	-

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 3 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.

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.

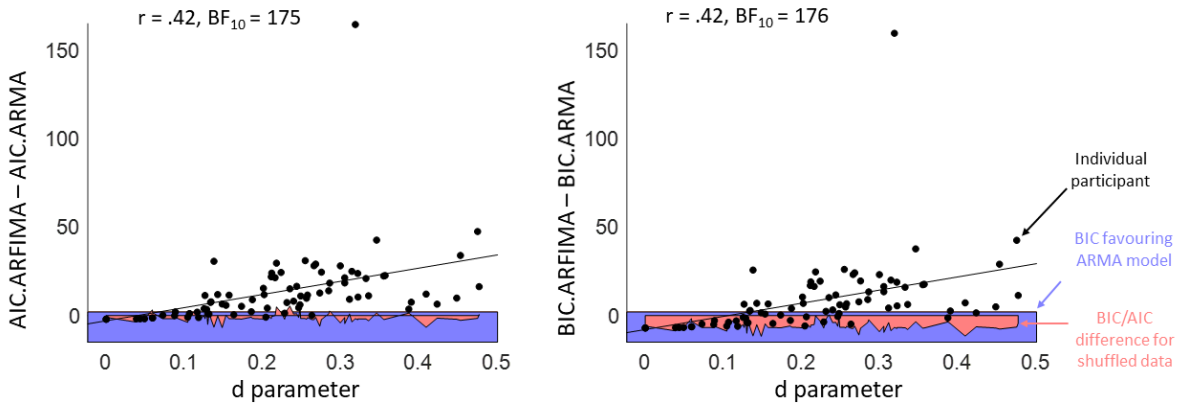

Figure 7. Shown are the difference in AIC (left) and BIC (right) between the ARMA(1,1) model and the ARFIMA(1,d,1) model, with each dot representing one individual subject. For points above the blue-shaded area (difference score of 2), the AIC/BIC clearly favours the ARFIMA(1,d,1) model – indicating 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.

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). 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, 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.

Supplementary Materials

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 existence of temporal structures – Supplementary 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).

Supplementary 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

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

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

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

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.

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

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.

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

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

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

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

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.

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

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

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

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

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).

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

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

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

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

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

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.

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

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

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

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

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

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

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

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

Stadnitski, T. (2012). Measuring Fractality. *Frontiers in Physiology*, *3*. https://doi.org/10.3389/fphys.2012.00127

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

The MathWorks, Inc. (Release 2016a). MATLAB 9. Natick, Massachusetts, United States.

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

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

Torre, K., Delignières, D., & Lemoine, L. (2007). Detection of long-range dependence and estimation of fractal exponents through ARFIMA modelling. British Journal of Mathematical and Statistical Psychology, 60(1), 85–106. https://doi.org/10.1348/000711005X89513

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

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

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.

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

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