Temporal structure in sensorimotor variability: a stable trait, but what for?

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, and this variability is a robust marker of individual differences. Variability is not fully random, but often exhibits temporal dependencies. However, their measurement and interpretation come with several controversies, and their potential benefit for studying individual differences remains unclear. Two recent studies have linked individual differences in temporal structures to task performance on the same data, but with contrasting results. Here we use the metronome task, requiring participants to press a button in synchrony with a tone, widely used to study sensorimotor variability. We first investigate the intra-individual repeatability of the most common measures of temporal structures. Secondly, we examine inter-individual correlates of these measures with: 1) task performance assessed from the same data series, 2) meta-cognitive ratings of on-taskness from thought probes occasionally presented throughout the task, and 3) self-assessed traits commonly assumed to relate to attention or sensorimotor control (ADHD tendencies, mind wandering, and impulsivity). Autocorrelation at lag 1 and Power Spectra Density slope showed high repeatability and good correlations with task performance. Detrended Fluctuation Analysis slope showed the same pattern, but less reliably. The long-term component (d) of the ARFIMA(1,d,1) model shows poor repeatability and no correlation to performance. None of these measures showed external validity when correlated to either mean subjective attentional state or self-assessed traits between participants. In sum, some measures of serial dependencies may be stable individual traits, but their usefulness to study individual differences in other constructs typically associated with variability in performance remains elusive.

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

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 over time – 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). The causes of such intrinsic variability in performance remain largely mysterious, attracting both physiological and cognitive interpretations. It is clear though that intrinsic variability reflects more than measurement noise: it is a stable individual trait that transfers across tasks and modalities (Hultsch, MacDonald & Dixon, 2002; Hultsch, MacDonald, Hunter, Levy-Bencheton & Strauss, 2000; Saville et al., 2011; 2012), evolves with age and neurodegenerative disorders (e.g., Tales et al., 2012; Tse, Balota, Yap, Duchek & MacCabe, 2010), and is considered a robust marker of Attention-Deficit and/or Hyperactivity Disorder (ADHD; see Kofler et al., 2013 for a meta-analysis; see Tamm et al., 2012 for a review).

Intrinsic variability in RT is often measured by calculating the standard deviation (SD) or coefficient of variability (SD divided by mean) over whole sessions (e.g., Perquin et al., 2020). However, these measures ignore serial dependencies within the

RT. Indeed, trials are not independent over time, but often show autocorrelations that persists over many trials (see for instance Gilden, 2001; Torre, Vergotte, Viel, Perrey & Dupeyron, 2019; Torre & Wagenmakers, 2009; Van Orden, Holden & Turvey, 2003; Wagenmakers, Farrell & Ratcliff, 2004). It has been suggested that these temporal structures consistently differ between individuals (Gilden & Hancock, 2007; Madison, 2004; Simola, Zhigalov, Morales-Muñoz, Palva & Palva, 2017; Torre, Balasubramaniam, Rheaume, Lemoine & Zalaznik, 2011), but this has not been investigated systematically. More specifically, it has been proposed that they reflect the ability for the brain to flexibly adapt and should be a marker of brain health (Simola et al., 2017; Torre et al., 2019). Related to this (although not in direct support to this claim), the temporal structures of *brain activity* data (e.g., oscillatory power at rest) have been shown to be reduced in schizophrenia (Nikulin, Jönsson & Brismar, 2012; Slezin et al., 2009; Sun et al,. 2014) and epilepsy patients (Adda & Benoudnine, 2016) compared to healthy controls.

In parallel to this argument, intrinsic variability is also often thought to be causally related to occasional disengagement from the task (Laflamme, Seli & Smilek, 2018; Seli, Cheyne & Smilek, 2013; Thomson, Seli, Besner & Smilek, 2014), and both are known to be increased in ADHD patients (ADHD and off-taskness: Shaw & Giambra, 1993; ADHD and variability: Kofler et al., 2013; Tamm et al., 2012). One may therefore expect individual differences in the strength of serial dependencies to be informative of people's ability to sustain attention and high performance over time. The purpose of the present research was therefore to link individual differences in variability, and more specifically serial dependencies, with subjective feeling of on-taskness and self-reported personality traits associated with ADHD.

Investigating temporal structures

Time series analysis methods

Before addressing individual differences in temporal structure, we first introduce the methods that have been commonly used to study temporal structure in behaviour: 1) autocorrelation, 2) Power Spectrum Density (PSD), 3) Detrended Fluctuation Analysis (DFA), and 4) autoregressive fractionally integrated moving-average (ARFIMA) models. 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. Autocorrelation 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 given 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. RT data typically show positive autocorrelations at short lags (Figure 1). For comparison, a completely random process in which the observations are fully independent from each other (i.e., white noise) would show no autocorrelation at any lag.

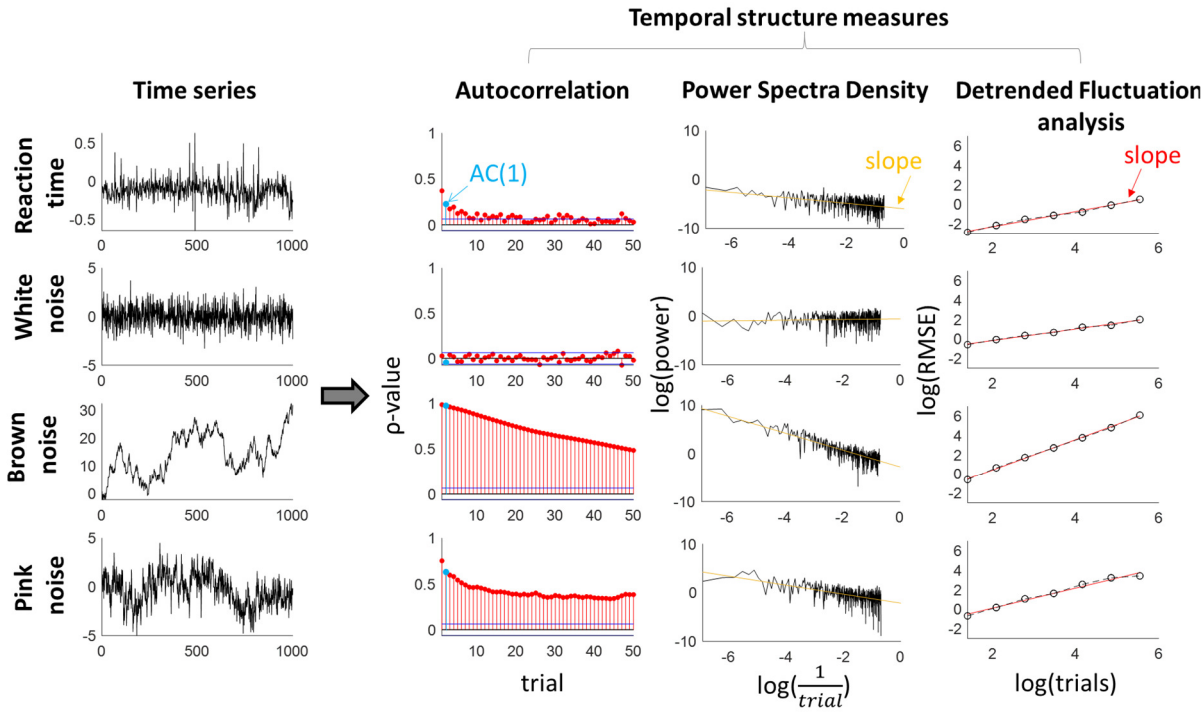

Figure 1. Examples of time series data over 1000 samples (left) and their corresponding temporal structures from lag 0 to lag 50. Shown are a reaction time series (showing small but clear temporal dependency), a white noise series (no temporal dependency), a brown noise series (i.e., a random walk, very 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 equivalently, the power spectrum can be calculated with a Fourier transform on the autocorrelation function (Box et al., 2016). In a variety of natural measures, including typical RT series, the frequency *f* and power *S(f)* are proportional:

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

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

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.

In white noise, every observation *n* on time series *x* consists solely of a random error term ε:

$$x_n = \varepsilon_n$$

In contrast, 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 (AC1) 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.

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'*, characterised by an α of one (although α between .5-1.5 are still 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. It should be noted that, while the regression line is linearly fit, there are no means to test the appropriateness of this fit (Torre, Delignières & Lemoine, 2007; Wagenmakers, Farrell & Ratcliff, 2005).

Detrended Fluctuation Analysis. Alternatively, Detrended Fluctuation Analysis (DFA; Peng, Havlin, Stanley, & Goldberger, 1995) has been used to analyse temporal structure in RT. While PSD is meant for stationary time series – time series that have a constant mean and variance throughout (e.g., white noise) – DFA is supposed to be more robust against non-stationarity (e.g., brown noise; Stadnitski, 2012). DFA has gained popularity in cognitive neuroscience in the recent years (e.g., Irrmischer, van der Wal & Linkenkaer-Hansen, 2018; Simola et al., 2017).

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. Note that if the window sizes are logarithmically spaced, this puts less emphasis on the shorter time scales compared to PSD. 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 slope 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.

ARFIMA models. Although PSD and DFA have been popular for analysing RT, the methods have an important limitation: they have difficulties differentiating long- from short-term dependencies. As pointed out in prior literature (see Wagenmakers et al., 2004 for a first detailed exposition of the problem), non-zero (or for DFA, non-.5) slopes are not necessarily indicative of long-term dependencies. Indeed, although short-term dependencies should theoretically lead to shallower slopes, in practice they can resemble pink noise. The use of autoregressive fractionally integrated moving-average (ARFIMA) models has been suggested (Wagenmakers et al., 2004; Torre et al., 2007) as an alternative to statistically *test* the benefit of a long-term parameter. The ARFIMA model is an extension of the ARIMA model, which consists of a combination of three processes, as detailed below.

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. If a time series is not stationary, 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 values for d are 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 – which means that 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. In short, critical systems supposedly reflect an optimal balance between predictability and randomness. In Physics, a system that has converged to the border between order and chaos is called critical, and natural fluctuations within such metastable systems exhibit a $1/f$ spectrum (Thorston & Gilden, 2005). It has been argued that neural networks self-organise to operate around the critical point, allowing maximum sensitivity to perturbation (e.g., from sensory inputs) without activity imploding (see e.g., Beggs & Timme, 2012; Shew & Plenz, 2013 for detailed reviews). To the extent that these theories apply to human behaviour and cognition (note the big conceptual jump), the presence of 1/f noise in behavioural time series could be taken to imply that 1) the brain-body system approaches criticality and 2) the system is affected by very slow fluctuations that "cause a cascade of energy dissipation at all length scales" (Bak, Tang & Wiesenfeld, 1987).

This cascading means that critical systems display some amount of correlation (here we focus on temporal correlations), with trials close in time showing the strongest correlation, which decreases as the time lag increases. As the dependency over time 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. Within the literature, this is often referred to as a 'power law' – the PSD 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; it has no specific time scales. 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.

Behaviour over time also shows 1/f noise and it has therefore been argued 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 remains a controversial topic (see Farrell, Wagenmakers & Ratcliff, 2006; Wagenmakers et al., 2004; 2005; Wagenmakers, van der Maas & Farrell, 2012 for critiques) – and may be dependent on the 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. This existence is clear though in specific tasks, including the task used in the present article: finger tapping in synchrony with a tone (Torre, Balasubramaniam & Delignières, 2010). The present research therefore addresses the usefulness of this research framework in understanding individual differences.

Predicting behaviour. Even if the temporal structures do not relate to criticality, one may still agree that RT series 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, and a recent article outlines a method to assess the feasibility of this approach to forecasting human behaviour on the next trial (Wagenmakers et al., 2016). Aside from a theoretical interest, such behavioural forecasting may also have a practical use: Given that behavioural variability fluctuates over time and occasionally fluctuates to extremely poor responses (very long RT, errors, omissions), it would be desirable 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. The present article is set to explore this aspect.

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 locally with fluctuations in performance (e.g., consistency during synchronised tapping), which deteriorates 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 subjectively feel less on-task. It has been argued that increased fluctuations from on-taskness to off-taskness are reflected in increased temporal structures (Irrmischer et al., 2018). Indeed, a mechanistic model has been proposed, where the combination of short-term dependencies (first order autoregressive term) and (comparatively slower) alternation between two response modes or strategies suffice to capture empirically observed temporal dependencies in synchronised tapping (Torre & Delignières, 2008; Torre & Wagenmakers, 2009; Torre

et al., 2010, see also Bastian & Sackur, 2013). Although it is tempting to associate the two modes to on-task and off-task subjective feelings, they can also be conceived as two states of an internal parameter (e.g., response threshold) which may or may not have meta-cognitive counterparts.

If temporal structures are indeed related to the pace at which such an internal variable fluctuates, they may be different in people who show low consistency in task performance. ADHD has previously been associated with higher RT variability, which have been attributed to 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 (Kofler et al., 2013; Tamm et al., 2012). 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). While these studies do not report the PSD slope (nor any of the time series analyses mentioned above), they found increased power in the low frequency of the spectrum (< 1.5 Hz) in ADHD patients, although it is unclear whether these would translate into higher PSD slopes for ADHD patients (see Discussion more details).

Individual differences in temporal dependencies

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), with 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. They found that participants with steeper DFA slopes had longer mean target RT (r = .72) – indicating they 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 (the study did not include a pre-manipulation baseline measure of performance). 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 the same theoretical viewpoint: brains which operate closer to the critical point show higher long-term correlations. On the one hand, Simola et al. (2017) interpret their positive correlation between temporal structure and performance as evidence that criticality allows for the mental flexibility needed to perform the Go/No-Go task. On the other hand, Irrmischer et al. (2017) interpret their negative correlation as evidence that criticality allows for

the successful dynamics of switching attention from task-related to task-unrelated thoughts. 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. We come back to this in Discussion.

A few studies have looked at the intra-individual repeatability of temporal dependency in task performance. 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 correlations between the two – suggesting that temporal structures are not consistent over different tasks. Separately for each task, they computed a Cronbach's α on the seven sessions. 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 AC1, the PSD, or the DFA between time 1 and 2, but do not report test-retest repeatability.[2]

Current research

Here, we examine individual differences in the properties of temporal dependencies, to assess: 1) to what extent these structures repeat in individuals over time, 2) how

[2] It is perfectly 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.

these structures relate to objective and subjective task measures, and 3) how these structures relate to self-assessed personality traits.

We used the Metronome Task (MRT; Seli et al., 2013), in which participants are instructed to press a button in synchrony with a regular tone. 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 relies minimally on the external environment while still providing behavioural measures and is therefore 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 an *online* measure of attentional state (i.e., measured during the experiment), which is known to correlate locally to fluctuations in RT. Last, the main performance measure is straightforward to interpret. Tasks such as the Go/No-Go or the CTET require both withholding and responding, providing 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 produces only few commission errors (she does not respond when she shouldn’t) but also many omission errors (she also does not respond when she should), it is unclear whether this constitutes good or poor performance. This complexity allows arguably too much flexibility in results interpretation. In contrast, the metronome task produces one main measure (asynchronies, or Rhythmic RT) with only very few omissions, that can be ignored. In a preliminary results section, we first validate the existence of temporal dependencies, including long-range correlations, in MRT data series.

Intra-individual repeatability of temporal dependency

For temporal dependencies to be informative for any individual differences, they need to be a stable trait, i.e., show consistency within individuals. In the main results section, we then report the intra-individual repeatability of a range of temporal dependency measures (AC1, PSD slope, DFA slope, and ARFIMA parameters) in 25 participants over two sessions of the MRT (conducted about 40 minutes apart). The measures were calculated separately for each session (time 1 vs time 2). For comparison, the same analysis was applied to mean performance and subjective attentional state ratings.

Correlates of temporal dependency

Last, we report the extent to which temporal structures in MRT data series relate to individual differences in: 1) task performance measures calculated on the same data, 2) subjective reports of attentional state from thought probes, and 3) (self-assessed) personality traits from questionnaires (ADHD tendencies, mind wandering tendencies and impulsivity). To investigate this, we examined the temporal structure measures in 84 participants.

Methods

Participants

84 healthy participants (69 female, 14 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

All 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, Kirsch & Hennig, 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 (Lynam, Smith, Whiteside & Cyders, 2006; Whiteside & Lynam, 2001). 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, Clark & Tellegen, 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.

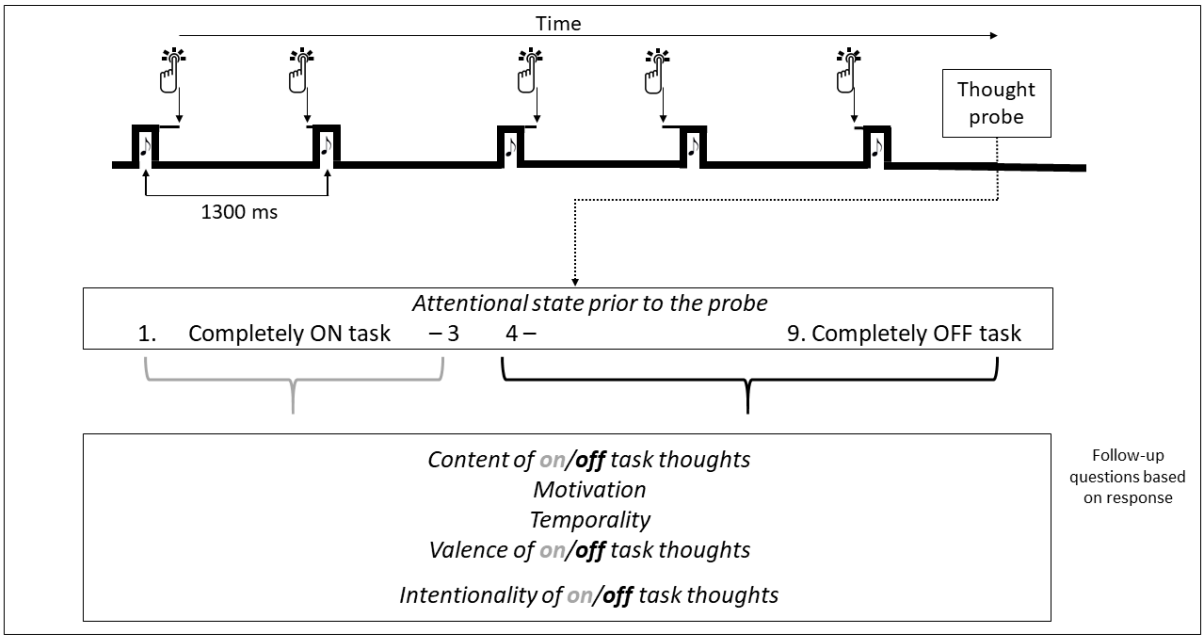

Figure 2. Overview of the task and thought probes.

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 questionnaires. In total, this took about 1.5 hours. Of the 83

participants, 25 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 lasts 1300 ms. In the middle of the trial (650 ms after onset), a short tone is 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. Rhythmic 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'). 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 five 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.

Two measures of overall performance were calculated on each RT series (RT_1, RT_2, …, RT_{1050}): the standard deviation of the RT (SD, reflecting consistency), and the mean of the absolute RT (|RT|, reflecting distance to the tone). As the distributions for both measures were highly skewed on the group level, both were log-transformed. As expected, the two measures were highly positively correlated ($r = .81$, $\log(BF_{10}) = 38.1$).

Two measures of subjective attentional state were calculated on each thought probe rating series: the mean and SD of the 21 ratings. These subjective measures were also correlated ($r = .38$, $BF_{10} = 53.1$). Across individuals, there was no evidence

that the subjective measures correlated with any of the objective measures (r ranging from 0.05 to 0.20, BF01 ranging 1.5 to 6.6). For completion, we also report the standard within-subject analysis (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 (median Kendall's $\tau = .09$, $BF_{10} = 585$). We report this now as this was not the purpose of the present study. We note that, consistent with previous studies, local correlations between performance and attention ratings are highly significant, but the shared variance is weak, consistent with the idea that most of the variability in performance escapes consciousness (see Perquin et al., 2020 for empirical implications of this idea).

The AC1 (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 weight of the two short-term (first-order) parameters AR and MA (referred to as φ_1 and θ_1 in Introduction).

Bayesian statistics were conducted in JASP (JASP Team, 2017), using equal prior probabilities for each model and 10000 Monte Carlo simulation iterations. We report the Bayes Factor (BF_{10}), which indicates 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, a BF_{10} of 3 in a correlation means that the likelihood for the data is 3 times larger under the correlation hypothesis than under the null-hypothesis, and would start to be interpreted as evidence in favour of a correlation. On the other hand, a BF_{10} of 0.33 means that the data is ($\frac{1}{.33}$) three times more likely under the null than under the stability hypothesis, which would start to be interpreted

as evidence for the null hypothesis. Any BF_{10} between 0.33 and 3 is typically referred as indeterminate.

Preliminary Results: temporal dependencies in MRT data

Presence of serial dependencies in MRT data

Before examining the intra- and inter-individual correlates of temporal structures in RT series, we first tested whether these series unambiguously showed such temporal structures. Bayesian One Sample t-tests were conducted on AC1 and PSD slopes – to test if they were statistically different from zero – and DFA slopes – to test if they were statistically different from 0.5. This was done separately for the participants from the first session (78 participants) and from the second session (21 participants). We found extreme evidence in favour of the existence of temporal structures (Table 1). We performed the same analysis on the ARFIMA(1,d,1) parameters and found extreme evidence for d values being above zero (see below for formal comparison between ARFIMA and ARMA model). The AR and MA weights were also clearly above zero on the larger sample who contributed to session 1, but not on the smaller sample who carried on to session 2.

Table 1. Bayes' Factors in favour of the existence of temporal structures in the RT in the different measures: 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). Green and red fonts indicate clear evidence against and for the null respectively.

	AC1*	Spectral slope*	DFA slope**	AR*	MA*	d*
Session 1	10^{74}	10^{79}	10^{78}	10^{5}	5	10^{57}
Session 2	10^{16}	10^{17}	10^{19}	0.025	0.032	10^{15}

* $BF_{value>0}$, ** $BF_{value>.5}$

Long-term dependencies

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, the difference in Akaike information criterion (AIC) between the best ARMA(1,1) and ARFIMA(1,d,1) models was calculated (following the procedure of Wagenmakers et al., 2004). Figure 3 shows this difference for each participant, ordered according to the value of their d parameter. Values above 0 indicate a better (lower) AIC for the ARFIMA model, while values below 0 indicate a better AIC for the ARMA model. In practice, however, only differences larger than 2 are taken as clear support for one model over the other (blue area in Figure 3; Wagenmakers et al., 2004).

Amongst the 78 analysed participants, the long-term model was clearly favoured for 59 of them (~75%). This was particularly the case in those participants with larger d values, as indicated by a positive correlation between d and the difference in AIC. When using the Bayesian information criterion (BIC) instead, a more conservative goodness-of-fit measure (as recommended by Torre et al., 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 for all participants (red area on Figure 3). We conclude that long-term correlations were likely presents in a majority of participants, but still note this was not the case for all of them. Below we assess the stability of these individual differences and explore their potential relevance.

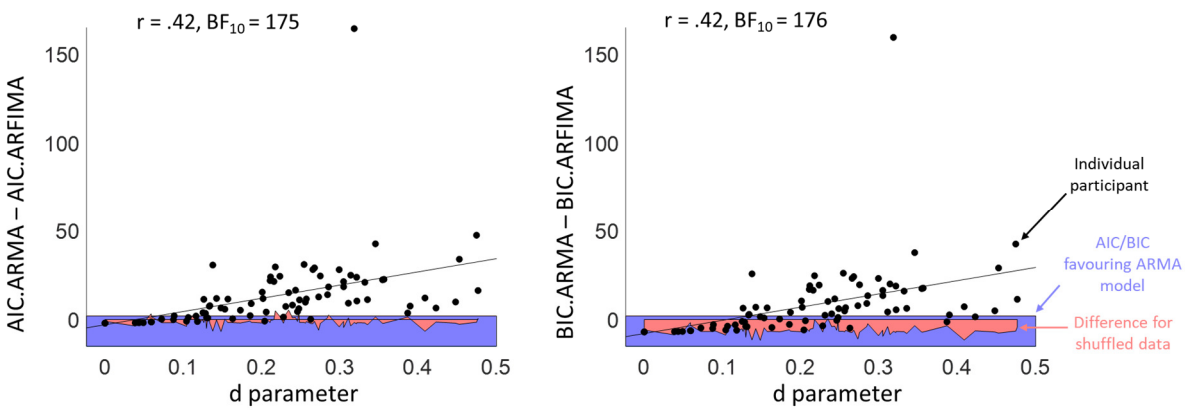

Figure 3. Difference in AIC (left) and BIC (right) between the ARMA(1,1) and ARFIMA(1,d,1) models, with each dot representing one individual subject. For points above the blue-shaded area (difference score of 2), the AIC/BIC clearly favours the ARFIMA(1,d,1) model – indicating that the d parameter adds 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.

Main Results: stability and validity of individual differences in temporal dependencies

Test-retest 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 4).

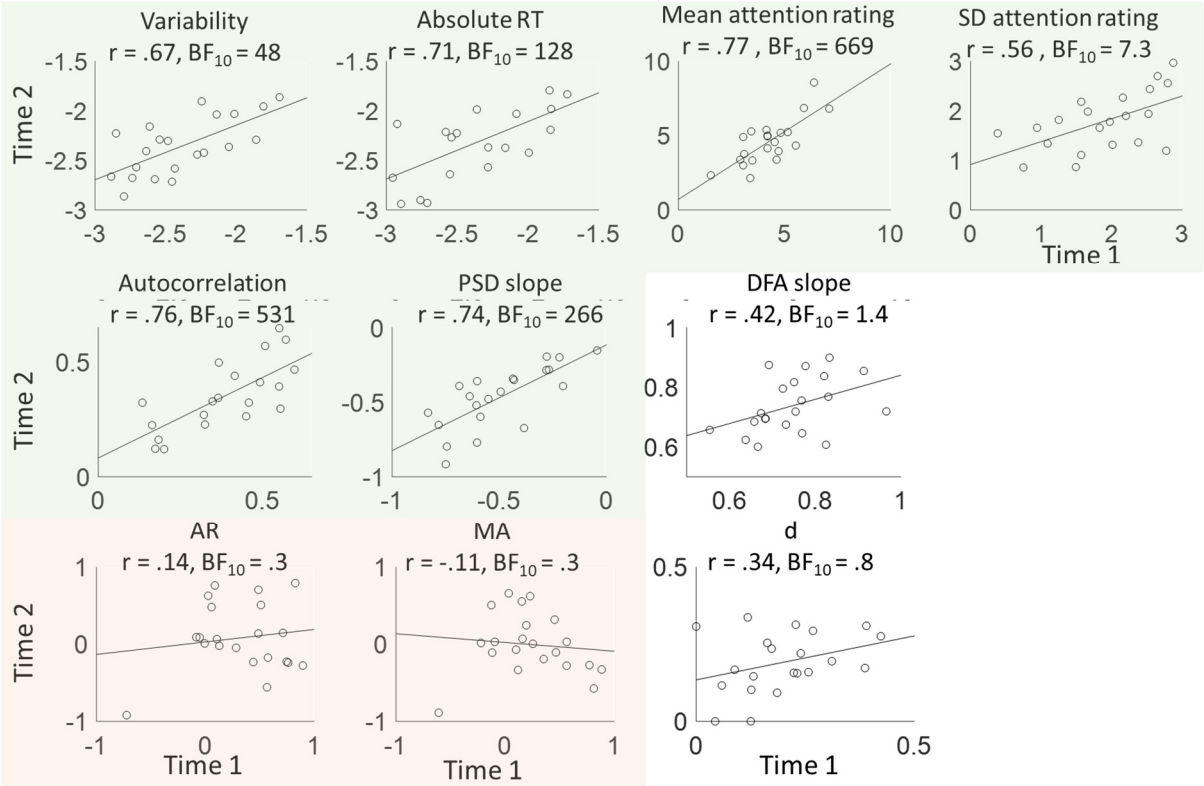

Figure 4. Within-subject correlations between MRT sessions 1 and 2 for performance (RT variability and mean absolute RT, both logged), subjective attentional state ratings (mean and variability), and temporal dependency measures (AC1, PSD slope, DFA slope, and the three ARFIMA(1,d,1) parameters – AR, MA, and d). Corresponding Bayes Factors above 3 are shaded in green (indicating clear evidence in favour of that correlation), while Bayes Factors below 0.3 are shaded in red (indicating clear evidence against that correlation). For illustrative purposes, the absolute value of the

PSD is displayed – meaning that for all temporal dependency measures, higher values indicate more dependency.

Overall, measures of performance (variability and mean absolute RT) and subjective attentional state ratings (mean and variability) showed high repeatability over time. Looking at the temporal structure measures, AC1 and PSD were the most reliable (equally high as the performance measures) and were the only ones strong enough to qualify unambiguously as stable traits. The short-term ARFIMA parameters were unreliable, while the DFA slope and d parameter fell in-between, with only indeterminate evidence for stability, and correlation coefficients that would be qualified as borderline between poor and medium.

Between-subject correlates of temporal dependency

Bayesian Pearson correlation analyses were conducted between each temporal dependency measure and: 1) RT variability and absolute mean RT (Figure 5, left), 2) the mean and SD of the attentional state ratings (Figure 5, right), and 3) the questionnaire scores (Figure 7). AC1, PSD slopes and DFA slopes all increased as performance decreased (Figure 5-6, this is detailed below). However, we found evidence against correlations of temporal dependencies with both attentional state ratings and questionnaires.

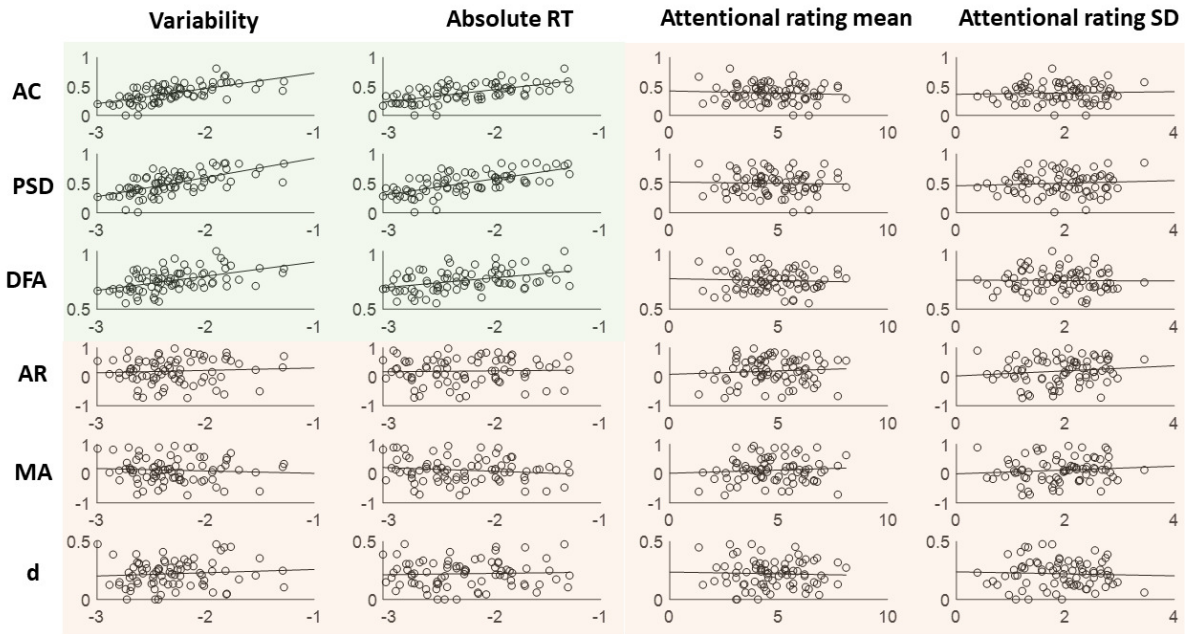

Figure 5. Between-subject correlations between temporal dependency and MRT measures. Convention are the same as in Figure 3. Performance correlates moderately to strongly with temporal dependency measures, with the exception of the ARFIMA(1,d,1) parameters. However, Bayes Factors show evidence against 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. Green and red fonts indicate clear evidence against and for the null respectively.

	RT Variability	Absolute RT	Mean attentional state	SD attentional state
AC	.62 (10^7)	.62 (10^7)	-.07 (.17)	.05 (.15)
PSD	.64 (10^7)	.66 (10^8)	-.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)

RT variability is the most common measure of performance in the MRT task. We found that participants who performed well (low DS) on the task displayed on average low temporal dependency (as indicated by strong correlations with AC1, PSD and DFA slopes). Crucially, these correlations cannot depend on the variance of the time series, as we correct for this by subtracting the shuffled data series. In the absence of correlation with the d parameter, we cannot conclude that the relationship between RT variability and temporal dependencies are carried by long-range correlations.

Figure 6 shows these dynamics in more detail over four different participants. Good performance (left column in Figure 6), as indicated by low variability, was associated with a low AC1, that appears to 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 (right column), as indicated by a high SD, was associated with high AC1, that appears to decay only slowly over the next lags, as well as with relatively steep PSD and DFA slopes. Average performance (respectively showing SD around median and mean values) showed intermediate temporal structures.

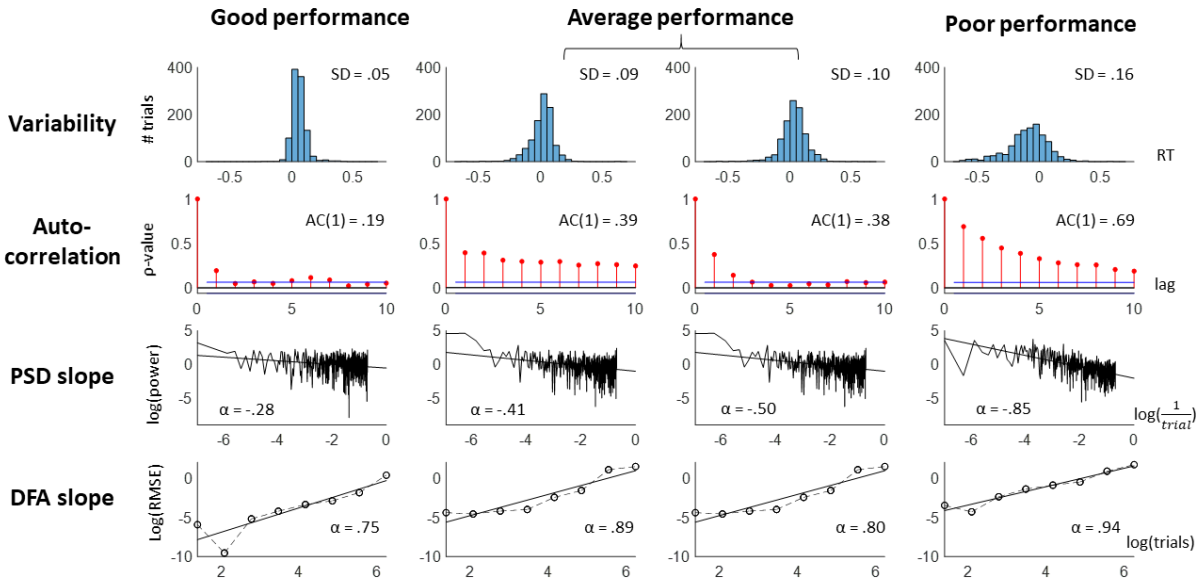

Figure 6. Examples from four participants over the temporal dependency measures, representing (from left to right) good, close-to-median, close-to-mean, and poor performance. Good performance was associated with low temporal dependencies, as reflected in low autocorrelation and shallow PSD and DFA slopes, while poor performance was associated with high temporal dependencies, as reflected in high autocorrelation and steep PSD and DFA slopes.

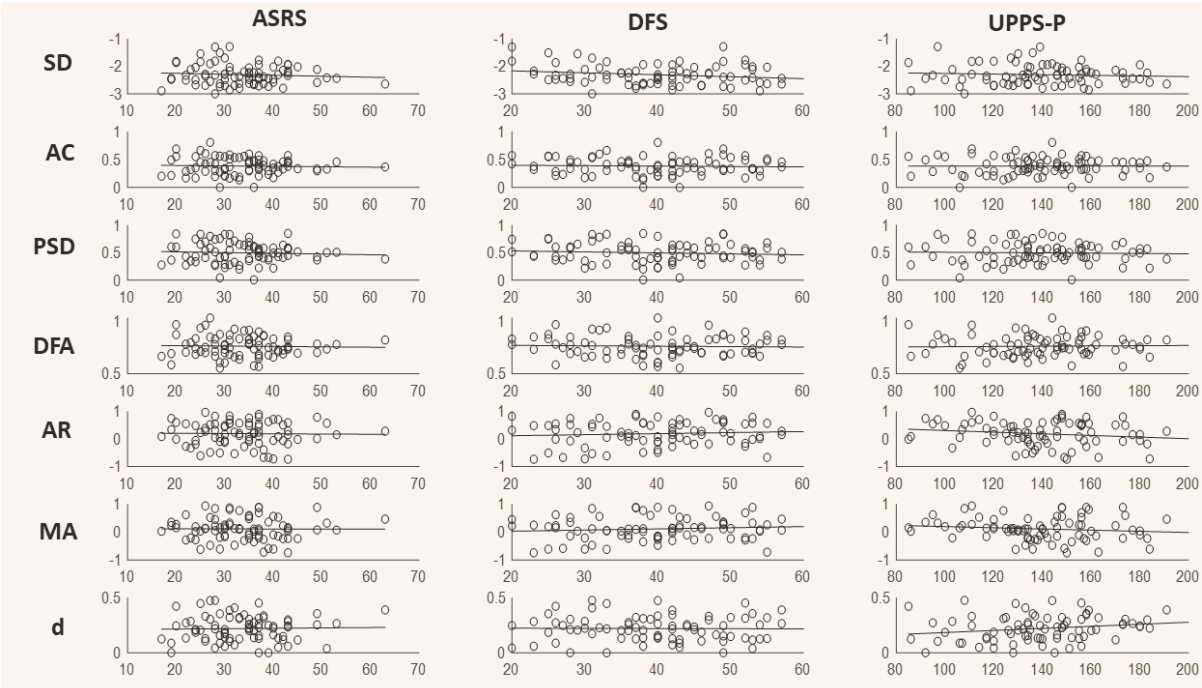

Figure 7. Between-subject correlations between the temporal dependency measures and self-assessed personality traits. Convention are the same as in Figure 3 and 4. None of the correlation pairs were supported.

Table 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)

Table 4. Pearson r-values with corresponding BF_{10} between the different measures of temporal dependency. Green and red fonts indicate clear evidence against and for the null respectively.

	PSD	DFA	AR	MA	d
AC(1)	.94 (10^{35})	.82 (10^{16})	.08 (.18)	-.09 (.19)	.30 (5.27)
PSD slope	-	.71 (10^{10})	.16 (.37)	-.06 (.16)	.14 (.31)
DFA slope	-	-	-.11 (.23)	-.12 (.24)	.58 (10^{5})
AR	-	-	-	.93 (10^{30})	-.27 (2.53)
MA	-	-	-	-	-.01 (.14)

Inter-measure correlations

Bayesian Pearson correlations across the different temporal dependency measures (performed on the data from session 1) revealed that the measures that showed some within-individual repeatability (AC1, PSD slope, DFA slope and d) correlated with each other (Table 4).

Counterintuitively maybe, although the AR and MA parameters from the ARFIMA model were not repeatable, they still correlated highly with each other (but not to any of the other measures).

Discussion

We provide the first systematic exploration of the repeatability and inter-individual correlates of temporal structures in endogenous RT across the most commonly used methods. Our results are summarised in Figure 8, where reliable variables are shown in bold, and the proximity between two variables captures the strength of their correlation. The figure also highlights notable absent correlations. These may be expected based on literature – such as the relationship between ADHD tendencies and RT variability – or on the supposed similarity between the measures – such as within-task attentional state ratings and general mind wandering (DFA) tendencies.

We found that temporal dependency showed repeatability over time, though this was dependent on which measure was used: AC1 and PSD were highly repeatable, the DFA slope and the long-term dependency ARFIMA term only moderately so, and the short-term ARFIMA parameters not at all. Similarly, we found that the objective measures of performance and the subjective measures of attentional state were highly repeatable over time (as are the questionnaires, see: Giambra, 1980; Perquin & Bompas, 2019). The temporal dependency measures (with exception of the ARFIMA(1,d,1) parameters) did correlate with performance – such that good performance was associated with less temporal structures. However, there was Bayesian evidence against correlations of the temporal dependency measures with both subjective attentional state and self-assessed personality traits. In Figure 8, this is reflected by ‘cluster forming’ of the different variables: A cluster of questionnaire-measures, a cluster of attentional state ratings, and a cluster of behavioural RT and

its features (with exception of the AR and MA parameters) – indicating that similar types of measures resemble each other, but do not translate to other types of measures.

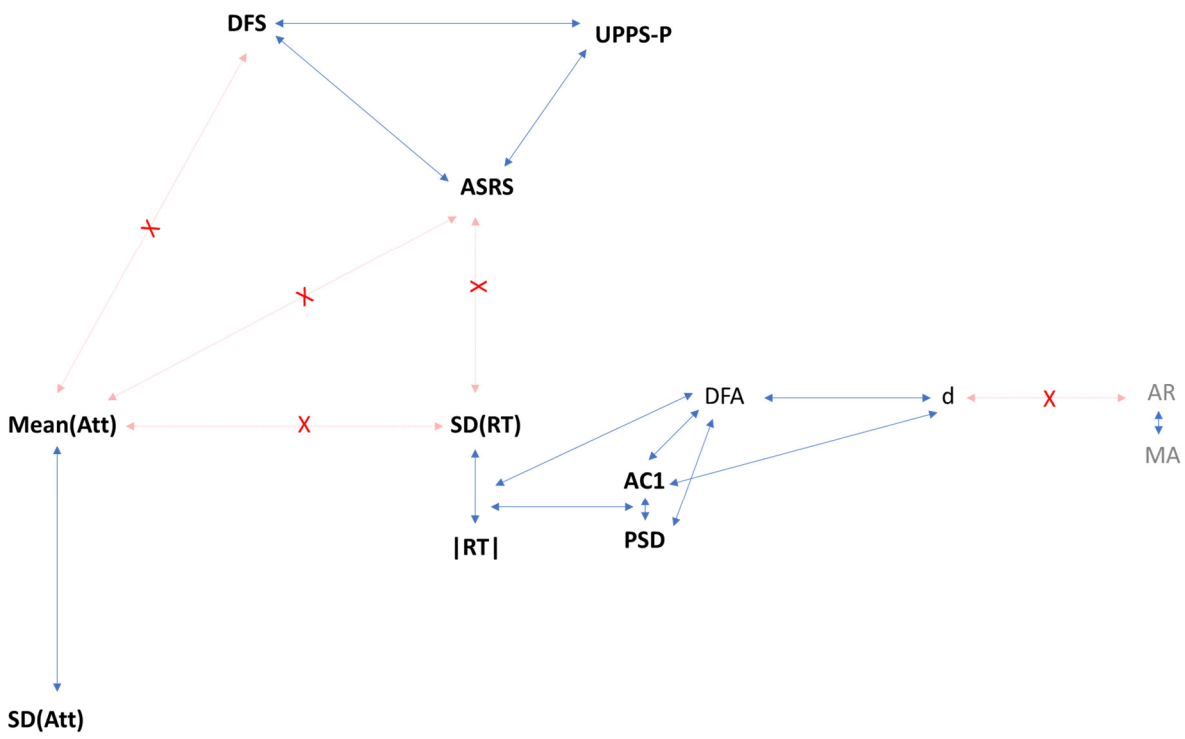

Figure 8. Repeatability and relationships between all measures. Most of the measures show at least good intra-individual reliability (bold), with the exception of the ARFIMA(1,d,1) parameters and the DFA slope. Correlated variables are linked by a blue arrow with a length of 1 – r, with r being the Pearson's r correlation coefficient (see Tables 2-4), i.e. strongly correlated variables are shown close together. The most notable absences of correlations are flagged with a red dotted line and red cross. The variables form clear clusters: one cluster with the performance measures, one cluster with the subjective attentional state ratings, and one cluster with the questionnaires. This indicates that, while measures of a similar construct correlate well together (e.g.,

personality traits, meta-cognition), they do not correlate to other constructs – reflecting poor external validity.

Intra-individual repeatability of temporal dependencies

Our conclusions that DFA slopes show moderate repeatability at best is consistent with previous work comparing slopes across two tasks (Torre et al., 2011). The Cronbach's α we report (0.59) is also in the same range as Torre et al. (2011). Our results indicate that AC1 and PSD measures were more reliable (with respectively ~58 and 53% of shared variance across sessions, while the DFA slopes only share ~18% of variance). To our knowledge, this is the first time the repeatability of the MRT measures is investigated. Consistency in MRT performance measures were high, and overall similar to other tasks involving RT (Hultsch et al., 2002; Saville et al., 2011; Saville et al., 2012). Mean subjective attentional states were the most reliable (~61% shared variance across sessions), although ratings did not correlate with any other measures across individuals (except for the standard deviation of these attentional ratings).

Temporal dependencies, performance and attentional state

The ACF, PSD, and DFA measures all increased with poorer performance. Superficially, this appears in line with Irrmischer et al. (2018) and at odds with Simola et al. (2017). It is possible though that our results are consistent with both studies, if we assume that, what temporal dependencies actually reflect, is not task performance, but fluctuations/flexibility. We may speculate that participants with low mental flexibility

perform better on the Metronome Task, as they stick to the consistent action throughout.

From the perspective of the mechanistic model mentioned in the Introduction (Torre & Delignières, 2008; Torre & Wagenmakers, 2009), one could expect that the relationship between performance and temporal dependencies across participants may strongly depend on the task, and may not be linear. For instance, an engaging task over the course of which one high-performance state dominates would be characterised by low temporal dependencies (because alternations between states are scarce). In this scenario, individuals impacted by more low-performance episodes would see their overall performance decrease but exhibit more temporal structure. In contrast, a task allowing on average for an even split between the two modes may lead to a quadratic relationship, whereby any departure from the average leads to less temporal dependencies. Of course, it remains entirely possible that such a two-mode model is not suitable to capture variability in many cognitive tasks.

In the current data, the monotonic increase of temporal structure with RT variability is consistent with the following (highly speculative) reasoning: good performance in the MRT task is associated with the dominance of one high-performance mode, while increasingly poorer performance is driven by more switching to a low-performance mode. The *within*-individual local correlation between RT variability and attention ratings suggests that high and low modes could be somehow associated with different subjective feelings of on-taskness. However, this correlation is weak, and the within-subject distributions of subjective reports were not bimodal. Moreover, we found evidence against correlations between temporal dependency and the *variability* of subjective attention ratings. Therefore, although the idea of mode-

switching between on-task and off-task states is seducing, it appears unlikely to be an accurate description of our data.

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. 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 may have led to different conclusions.

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 the high variability group showed less temporal structure, i.e. resembled white noise to a higher extent than the low variability group. However, aside from methodological issues with the analysis method (see Farrell et al.,2006 for a critique), several problems with the design 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

relationship between questionnaire scores and temporal RT structures was not directly assessed, but inferred indirectly via the low and high variability groups. 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. 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 (although we did not observe this in the present data in healthy individuals), 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 compare peak frequencies across groups. Although these studies typically found increased power in ADHD patients, they have only focused on low frequencies (< 1.5 Hz). It remains unknown if this would translate in higher *slopes* (which would mainly happen if lower frequencies are increased but higher frequencies are not).

In contrast, ADHD in children has 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. This makes conclusions from learning tasks hard to generalise to other tasks designed to focus on endogenous variability alone.

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 AC1 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) gives 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. However, it should be noted that, as the ARFIMA parameters were at best poorly 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. However, the AR and MA weights from the ARMA model (hence only two parameters) were not more reliable than from the ARFIMA model. Only when each of the parameters were fitted separately (e.g., fitting an ARFIMA(0,d,0) to estimate *d*) did they show high repeatability across sessions (AR: $r = 0.77$, $BF_{10} = 689$; MA: $r = 0.77$, $BF_{10} = 538$; d: $r = 0.77$, $BF_{10} = 620$). Furthermore, the single AR weights were the exact same values as AC1 – 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.

While AC1 and PSD have the highest repeatability, the DFA may come with most 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 against the window size (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, which means that 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.

For example, in the current dataset, we calculated PSD over all the possible frequencies. However, log-log spectra of behaviour tend not to follow a fully approximately linear downwards slope, but instead increase in power at the highest frequencies – resulting in a small curve at the high end of the spectrum (see Torre & Delignières, 2008 Figure 1 for instance). Some previous studies therefore exclude the highest frequencies, to fit the slope only on the linear part of the spectrum. For

example, one could exclude 1/8 of the highest frequencies (e.g., Torre et al., 2019). Similar approaches have been taken in DFA extraction – for instance, by taking a minimum window size of 10 trials (Torre et al., 2019). Both these approaches aim to make the measures as sensitive as possible. We recalculated our PSD using this rule, and reran the within- and between-subject analyses – but found no substantial differences in findings. Using a similar approach, we recalculated the DFA measures on windows of 16-512 trials (instead of 4-512), and found that the correlation coefficient for within-subject reliability *decreased* substantially (r = .27, with Bayesian evidence also decreased but still in the indeterminate range: BF_{10} = .54). Furthermore, DFA no longer correlated with performance (RT: r = .14, BF_{01} = 3.4; SD: r = .19, BF_{01} = 1.7) – suggesting that the short-scale temporal dependencies are most informative of individual differences in performance.

Ideally, it should thus be reported how any analysis choices were decided upon, and potentially, 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. Three methods have been discussed to deal with this issue: 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 results comparison. 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 – which means that 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 may introduce or mask 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 with each individual's median RT, and 2) replacing with an RT of 650 ms (reflecting the maximum time a participant had to respond). Both of these methods

gave mostly similar response patterns, with 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 was more reliable, now showing moderate 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 measures. In particular, AC1 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 AC1 with self-assessed ADHD, mind wandering, or impulsivity, previous studies have hinted at differences when particularly studying clinical cases of ADHD versus healthy controls.

Of course, repeatability is not the only construct of interest in neurocognitive variables. Most of the measures did correlate to performance – and may still be useful for capturing moment-to-moment fluctuations in the data. However, as previously noted (Stadnitski, 2012), small changes in the analysis pipeline may lead to substantial changes, highlighting the importance of transparent reporting.

Still, the underlying mechanisms driving the relationship between performance and temporal structure remain unknown. 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 clear falsifiable predictions (see for instance 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. Similarly, some may argue that all participants should exhibit temporal structure to the same extent, because all neural systems should have spontaneously converged towards criticality. We note that this was clearly not the case in our data, where some participants exhibited clear temporal structure, while others exhibited none. Unfortunately though, such repeatable individual difference failed to show external validity. Future research may therefore aim to directly manipulate the temporal structures with different experimental conditions (see Perquin et al., 2020 for one such an attempt), rather than just measure them – to get a clear idea of their underlying neural-cognitive mechanisms.

Context paragraph

Within daily life, people often make intuitive assumptions about fluctuations in their attentional state and in their performance, as well as about how these fluctuations relate to each other. For example, during a task, one may feel to be in a high-performing state, in which one feels highly focused and performing very well over some period of time spanning multiple repeated actions or trials. However, empirical support in favour of such assumptions remains weak. While serial correlations are commonly found in performance, their interpretation is controversial, and their link to performance and meta-cognition is rarely addressed. Furthermore, while the literature on mind wandering has commonly linked within-subject fluctuations of performance and subjective attentional state, these links show low effect sizes and seem to be overinterpreted. With this current work, we aimed to empirically investigate these assumptions with a design that could capture different elements (e.g., temporal structures, performance, subjective attentional state). Furthermore, we were motivated by an interest in the temporal structure measures themselves, aiming to figure out how they can be used to study behaviour and individual differences – and therefore using a multi-measure approach to capture the temporal structures.

Acknowledgements

We would like to thank Rachel Draper, Laura Daniells, Laura Fleetwood, and Joel Bentley for their help with collecting the data. We are grateful to Simon Farrell for sharing his code for the time series analyses.

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

Adda, A. & Benoudnine, H. (2016). Detrended fluctuation analysis of EEG recordings for epileptic seizure detection. *International Conference on Bio-engineering for Smart Technologies (bioSMART),* 1-4.

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

Bastian, M., & Sackur, J. (2013). Mind wandering at the fingertips: automatic parsing of subjective states based on response time variability. *Frontiers in Psychology, 4*:573. https://doi.org/10.3389/fpsyg.2013.00573

Bak, P., Tang, C., & Wiesenfeld, K. (1987). Self-organised criticality: An explanation of the 1/f noise. *Physical Review Letters, 59*(4), 381. https://doi.org/10.1103/PhysRevLett.59.381

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

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

Delignières, 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.

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

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

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 metaanalytic review of 319 studies. *Clinical Psychology Review,* *33*(6), 795–811. https://doi.org/10.1016/j.cpr.2013.06.001

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, D. R., Smith, G. T., Whiteside, S. P., & Cyders, M. A. (2006). *The UPPS-P: Assessing five personality pathways to impulsive behavior* (Technical Report) West Lafayette: Purdue University.

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

Nikulin, V. V., Jönsson, E. G., & Brismar, T. (2012). Attenuation of long-range temporal correlations in the amplitude dynamics of alpha and beta neuronal oscillations in patients with schizophrenia. *NeuroImage, 61*(1), 162-169. https://doi.org/10.1016/j.neuroimage.2012.03.008

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

Perquin, M. N. & Bompas, A. (2019). Reliability and correlates of intraindividual variability in the oculomotor system. *Journal of Eye Movement Research, 12*(6):11. https://doi.org/1 0.16910/jemr.12.6.11

Perquin, M. N., Yang, J., Teufel, C., Sumner, P., Hedge, C., & Bompas, A. (2020). Inability to improve performance with control shows limited access to inner states. *Journal of Experimental Psychology: General, 149*(2), 249-274. http://dx.doi.org/10.1037/xge0000641

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

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

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

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, 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

Slezin, V. B., Korsakova, E. A., Dytjatkovsky, M. A., Schultz, E. A., Arystova, T. A., & Siivola, J. R. (2009). Multifractal analysis as an aid in the diagnostics of mental disorders. *Nordic Journal of Psychiatry, 61*(5). 339-342. https://doi.org/10.1080/08039480701643175

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

Sun, J., Tang, Y., Lim, K. O., Wang, J., Tong, S. Li, H., & He, B. (2014). Abnormal dynamics of EEG oscillations in schizophrenia patients on multiple time scales. *IEEE Transactions on Biomedical Engineering, 61*(6), 1756-1764. https://doi.org/10.1109/TBME.2014.2306424

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

The MathWorks, Inc. (Release 2015a). MATLAB 8. 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., & Delignières, D. (2010). Oscillating in Synchrony with a Metronome: Serial Dependence, Limit Cycle Dynamics, and Modeling. *Motor Control, 14*(3)*,* 323-343. https://doi.org/10.1123/mcj.14.3.323

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. (2008). Unraveling the finding of 1/ f^{β} noise in self-paced and synchronized tapping: a unifying mechanistic model. *Biological Cybernetics, 99*(2), 159-170. https://doi.org/10.1007/s00422-008-0247-8

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

Torre, K., Vergotte, G., Viel, É., Perrey, S., & Dupeyron, A. (2019). Fractal properties in sensorimotor variability unveil internal adaptations of the organism before symptomatic functional decline. *Scientific Reports, 9*:15736. https://doi.org/10.1038/s41598-019-52091-y

Torre, K., & Wagenmakers, E.-J. (2009). Theories and models of 1/f^{β} noise in human movement science. *Human Movement Science, 28*(3), 279-318. https://doi.org/10.1016/j.humov.2009.01.001

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

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/f^{α} 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., Grünwald, P., & Steyvers, M. (2016). Accumulative prediction error and the selection of time series models. *Journal of Mathematical Psychology, 50*(2)*,* 149-166. https://doi.org/10.1016/j.jmp.2006.01.004

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