UiT The Arctic University of Norway

Faculty of Health Sciences, Institute for Psychology

Feedback Effects on Mind Wandering:

A Series of Online Experiments

Krister Karlsen

Supervisor: Prof. Matthias Mittner

Master's Thesis in Psychology, May 2022

Feedback Effects on Mind Wandering:

A Series of Online Experiments

Student: Krister Karlsen

Supervisor: Matthias Mittner

PSY-3900

Master's Thesis in Psychology

Institute for Psychology

UiT The Arctic University of Norway

Tromsø, May 2022

Forord

Veien til denne masteroppgaven ble til gjennom flere tilfeldige ytre faktorer. For det første har jeg alltid ønsket å kunne skape noe eget gjennom å programmere noe fra bunnen av. For det andre oppsto det en pandemi ved starten av masteren som gjorde møte blant deltagere, veiledere og studenter mer utfordrende, til sist startet jeg i en jobb som vekselvis krevde mye av min tid. Etter et møte med min veileder fant vi i sammen ut at jeg kunne undersøke mulighetene for å gjøre et nettbasert eksperiment. På så måte oppnådde jeg fleksibilitet i min arbeidstid. Følgelig fant jeg ut at nettbasert eksperimentering var en effektiv metode for å samle data og dette muliggjorde innsamling av data til så mange som seks eksperiment. Korte møter med veileder over zoom hjalp meg med å forstå både programmeringsspråket til eksperimentene og analysespråket som vi benyttet i dataanalyseverktøyet R. Masteren har gitt meg mange fine utfordringer og jeg er utrolig glad for alt den har lært meg. Både hvordan programmere egne eksperiment, samle data online, benytte R og utføre diagnostikk på data var nytt for meg. Kunne ikke ha gjort dette uten Matthias, min veileder, som har pushet med til å følge min plan og ikke slippe opp selv om det eksempelvis hadde holdt med færre studier. Han har fått meg til å føle meg verdsatt og dratt meg inn i diskusjoner i den kognitive nevrovitenskap gruppen. Han har vært uvurderlig når det kom til å diskutere både dataanalyser og i programmering av eksperimentene. Takk Matthias.

Krister Karlsen

Student

Matthias Mittner

Veileder

Abstract

Replicating in-lab experiments online can ensure scientific progress when physical contact is discouraged, like during the covid-19 pandemic. In this thesis, we replicated the results from Boayue et al. (2021) in-lab Mind Wandering (MW) experiment online. The task uses the Finger-Tapping Random Sequence Generation Task, a sustained attention task, equating MW with Task Unrelated Thoughts (TUTs). In addition to collecting self-reported TUTs, the FT-RSGT continuously collects Behavioural Variability (BV) and Approximate Entropy (AE), which are both related to MW. We replicated Boayue et al. (2021) in-lab results showing that we can reliably conduct MW experiments online. Moreover, by using six different versions of the task, we investigated whether giving different types of feedback to the participants could improve their task focus. The task versions were: (1) Identical to the lab-based task. (2) Performance feedback training. (3) Intermittently delivered performance feedback throughout the experiment. (4) Non-specific feedback. (5) Camera monitoring feedback and (6) progression feedback. We consistently found that specific performance feedback increased the global on-task focus as measured by our behavioural indices (AE and BV) relative to non-specific motivational feedback, leaving self-reported MW unaffected. On the other hand, progression and camera feedback increased the magnitude of the subjectively reported MW while leaving task performance unchanged. This dissociation could result from participants' exaggeration when surveilled and their novelty with the self-rating. We conclude that, during online experiments, researchers may want to incorporate performance feedback to increase behavioural indices. These insights may apply to other situations where increased task performance is desired. All data, experimental materials, and pre-registrations are available at the Open Science Framework (https://osf.io/wjvk2).

Keywords: mind wandering, performance feedback, task unrelated thoughts, sustained attention, online experiment

Sammendrag

For å sikre vitenskapelig fremgang er det ønskelig å kunne replikere laboratorieeksperimenter gjennom nettbaserte eksperiment som ikke krever fysisk tilstedeværelse, dette spesielt i tider hvor fysisk kontakt er frarådet, som under covid-19 pandemien. Vi replikerer et tidligere tankevandring laboratorium eksperiment av Boayue og kollegaer (2021) slik at det kan kjøres over nettet. Oppgaven som benyttes er en vedvarende oppmerksomhet oppgave. Selvrapportering gjennom tilfeldig fordelte spørsmål i eksperimentet måler deltagernes opplevde tankevandring. Tanker som kategoriseres som ikke oppgaverelatert kategoriseres som tankevandring. I tillegg til selvrapporteringene bruker oppgaven behavioristiske mål som indikerer grad av tankevandring. Disse er tilfeldighets- og variabilitet mål som angir tilfeldighet og presisjonen i deltagers ytelse. Høy tilfeldighet og presisjon indikerer oppgavefokus og lite tankevandring. Vi replikerte laboratorium resultatene i vår nettbaserte studie og utvidet eksperimentet med å inkludere tilbakemeldinger som kan øke oppgavefokus. Vi publiserer seks versjoner av eksperimentet; (1) identisk som laboratorium eksperimentet. (2) Oppgave ytelses trening. (3) Ytelses tilbakemelding gjennom eksperimentet. (4) Motiverende ikke ytelsesbasert tilbakemelding. (5) Kamera overvåkning tilbakemelding og (6), progresjonstilbakemelding. Resultatene viser at oppgave fokuset øker ved bruk av spesifikk ytelse tilbakemelding gjennom eksperimentet uten å påvirke selvrapportert tenkevandring. Motsatt finner vi at elementer som ikke informerer om oppgaveytelsen distraherer deltagerne og resulterer i økt selvrapportering av MW. Uoverensstemmelsen mellom de behavioristiske indikatorene for tankevandring og selv rapportene kan resultere fra deltagers' tendens til å overdrive svar når de overvåkes eller mulig fra deres manglende erfaring med å selv-rapportere grad av fokus. Vi konkluderer med at ytelses basert tilbakemelding kan benyttes for å øke oppgaveytelsen. For å sikre vitenskapelig åpenhet er all eksperimentell data tilgjengelig på "Open Science Framework" (https://osf.io/wjvk2).

Feedback Effects on Mind Wandering: An Online Experimental Series

During the Covid-19 pandemic, the World Health Organization recommended that governments apply safety precautions intended to limit transmission of the virus, including limiting peer-to-peer physical meetings. During surges in infections, governments can even deny universities and scientists to perform physical in-lab experiments to reduce the spread of the disease and prevent the national healthcare service from collapsing (World Health Organization, 2020). One way to preserve scientific progress in times like these involves conducting behavioural experiments online. However, this raises questions about the validity of experimental data collected online. For example, there have been concerns about varying reaction times (Crump et al., 2013) as participants online use their personal computers, including different browsers, software, and internet speeds.

Continuing to gain knowledge during restrictions on physical meetings, overcoming concerns about replication crisis (Simmons et al., 2011) and validity of experiments, existing in-lab experimental tasks can be transformed to fit the online condition to investigate whether the results from the in-lab experiment replicate in the online setting. Consequently, there has been a surge in online studies successfully replicating different in lab-experiments in the online condition (Buso et al., 2021; Claypoole et al., 2018; Crump et al., 2013; Dandurand et al., 2008; de Leeuw, 2015; Nussenbaum et al., 2020; Ratcliff & Hendrickson, 2021). We contribute to this verification process by transforming an in-lab task designed to investigate the behavioural effects of mind wandering (MW) and test whether the in-lab results are replicable in the online condition. Extending the in-lab task, we develop performance feedback to increase task comprehension. Moreover, we test the participants under six different conditions displaying different feedback. We collect both behavioural indices of MW and self-reports.

MW is a natural part of being human. Thus, most people intuitively understand the concept. However, defining the phenomenon scientifically into a term that captures the whole complexity included in the intuitive understanding has shown to be challenging. Consequently, there exist multiple definitions that capture different aspects of it. Examples that have received attention in the literature are Task Unrelated Thoughts (TUTs), Unintentional Thought, Stimuli-Independent Thought, Stimuli-Independent Task-Unrelated Thought, or Unguided Thought (Seli et al., 2018). For example, the term Unintentional Thought would not categorize a thought as MW if, during a task, a participant would deliberately try to figure out what to have for dinner or generally intentionally engage in other thoughts (while it would fall into the definition of task-unrelated thoughts). Consequently, this definition seems less suited as a definition for experimental tasks as it fails to capture such intentional attentional failures (Seli et al., 2018). We circumvent this challenge by categorizing all thoughts unrelated to our experimental task as MW. Thus, we use TUTs to define MW – even though this comes with its own set of problems (Smallwood & Schooler, 2006).

When we fail to keep our attention directed at the task at hand, our thoughts become task unrelated (Smallwood & Schooler, 2006). One early method to investigate MW that is still in widespread use today is the experience sampling method (ESM; Csikszentmihalyi et al., 1997). ESM involves repeatedly prompting participants with one or several questions concerning their thoughts and experiences over time, for example, over days or weeks. Using ESM, Killingsworth & Gilbert (2010) found that people engage in MW (defined as TUTs) approximately 50% over the duration of a day, while Kane et al. (2007) found that the mean time spent in an MW state was 30%. Most participants, when told about this finding, were not surprised. Fascinatingly, this indicates that people can go as far as to only pay attention to the external environment – or the task at hand – half of the day.

Known drawbacks with engaging in MW are that it disturbs our performance on an external task, and MW is therefore especially detrimental to tasks requiring sustained attention. For example, during an in-person lecture (Farley et al., 2013), online lecture (Szpunar et al., 2013), or reading (Dixon & Bortolussi, 2013; Jackson & Balota, 2012; Unsworth & McMillan, 2013), engaging in MW reduces the total amount of comprehension.

On the other side, two influential literature reviews (Mooneyham & Schooler, 2013; Smallwood & Andrews-Hanna, 2013) clarify that MW is vital in autobiographical planning and creative problem-solving. When the current task is perceived as uninteresting, unimportant, or too complex, the mind frees executive resources to solve other perceived problems. Then, when the mind starts to wander, it takes our attention through different thoughts, called attentional cycling. Consequently, the attention is directed at current and relevant goals in our life. However, Wilson et al. (2014) found that people would rather shock themselves than be left in an empty room to themself. Accordingly, MW can give us a mental break, freedom from the "unbearable" or "boring" here and now. Still, this mental break should not be necessary if the individual manages to keep the full attention to the here and now and not reflect on irrelevant information, as when in flow (Csikszentmihalyi, 2008). During flow, the thoughts are entirely directed to the task at hand, and there would not be a need for a mental break as the full awareness of the moment would feel rewarding in itself (Csikszentmihalyi, 2008).

Taking the experiment to the online condition has some additional drawbacks that are hard to overcome. For example, it is challenging to use specialized equipment, making it hard – or impossible – to collect neurological data needed if we were to use triangulation for identifying MW episodes (Smallwood & Schooler, 2006). Additionally, we lose a great deal of control (Reips, 2000), not knowing how many distractions are in the participant's environment and if they leave the computer – for example, getting a friend to continue while

he makes himself a coffee. It is even possible for the participants to ignore the instructions entirely and watch TV simultaneously as they participate (Chandler et al., 2014). It would be tough to identify such behaviour from the experimental data alone.

On the other side, convenient access to a broad and diverse group that is representative, at least, of the online population makes online studies attractive. In addition, online experimental scripts include everything, from instruction to the code behind the task, making it quite transparent (for the current project, see https://osf.io/wjvk2). Consequently, other researchers who want to replicate the experiment can copy the script and run it through a new sample. It is also easy to add tasks or even set up the experiment in the lab or analyze the data in alternative ways.

In many cases, people might not know that their thoughts have started to drift. They lack meta-awareness (Schooler, 2002). This lack of meta-awareness is a challenge when collecting self-reports, as during ESM. Clearly, participants need to be aware of what they were thinking about directly before the prompt to rate their thoughts (Smallwood & Schooler, 2006). A commonly reported experience is that during reading, the mind wanders without the reader even noticing (Schooler, 2002). Consequently, scientists have been looking for other measures in addition to self-reports that indicate when participants are likely to engage in MW. Several studies suggest that there are behavioural and neural indicators of the switching from task-focus to TUTs (Boayue et at., 2021; Kucyi et al., 2017; Seli et al., 2013; Teasdale et al., 1995). For this reason, Smallwood & Schooler (2006) recommends triangulation as a method for investigating MW. Triangulation uses data from different modalities, i.e., self-reported probes, behavioural measures, and neurophysiological or neuroimaging measures, to infer the participant's state of mind. We use self-reported thought probes and two behavioural measures in our online studies.

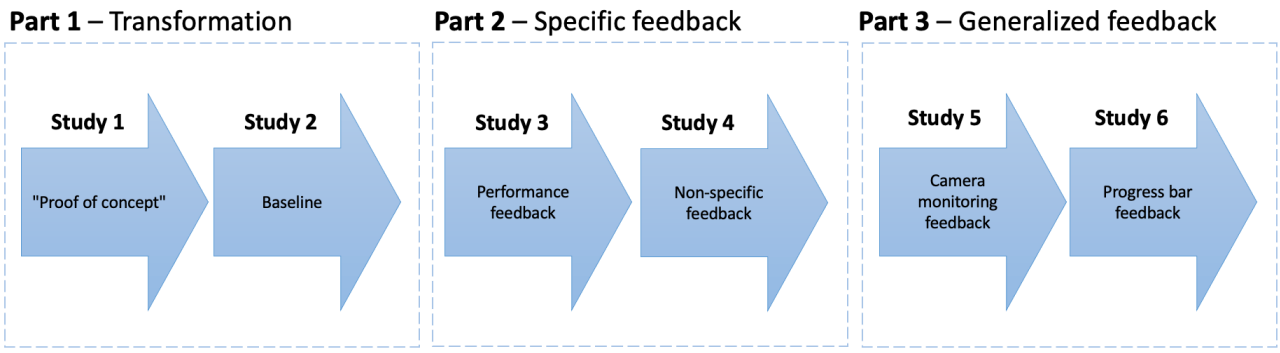
To date, most behavioural experiments studying MW have employed the Sustained Attention to Response Task (SART; Robertson et al., 1997), and several reliable behavioural patterns have been identified (e.g., Cheyne et al., 2009; for review, see Smallwood & Schooler, 2006). The SART is a go-nogo task, meaning that participants must oscillate between responding and withholding a response depending on the presented stimuli. In the SART, participants are instructed to press the space bar for all displayed numbers except for the number three. When a "3" appears on the screen, they must withhold the keypress. Hence, if their mind wanders at this critical moment, participants tend to continue pressing even though the number three appears. Therefore, when a participant misses the "3", this indicates MW at that time. However, "3" appears very infrequent. Thus, this is not very efficient in measuring TUTs as data is only obtained every time "3" appears. To improve on the SART, the recently developed Finger-Tapping Randomized Sequence Generation Task (FT-RSGT; Boayue et al., 2021) continually collects behavioural data indicating MW as well as intermittently probing participants to self-report their thoughts (Boayue et al., 2021).

The FT-RSG task combines a random number generation task (Baddeley et al., 1998; Towse, 1998) and a finger tapping task (Kucyi et al., 2017; Seli al., 2013). The randomness within the number generation is measured using Approximate Entropy (AE), and the precision of the finger-tapping is measured using Behavioural Variability (BV). Both indices (AE and BV) have been related to executive control/functioning and MW (Boayue et al., 2021; Seli et al., 2013; Teasdale et al., 1995). Consequently, higher AE and lower BV values indicate MW.

In the current thesis, we build on the FT-RSGT developed and tested in the lab and transform the task to fit into the online condition, thus overcoming challenges with restrictions on peer-to-peer in-lab meetings during the Covid-19 pandemic. Moreover, our experiments are divided into three parts; Part 1 takes the in-lab FT-RSG task and transforms

it into an online experiment. Part 2 investigates whether the feedback needs to reflect the task performance or merely encourage participants. Part 3 investigates if additional feedback improves or decreases task focus. See Figure 1 for an overview of the studies.

Figure 1



Note. Overview of the studies displaying the workflow from study 1 to study 6.

In part 1, we focus on adapting the in-lab FT-RSGT into an online version. Study 1 serves as a "proof of concept" and was published to investigate whether the previously found results regarding the relation between behavioural performance in the FT-RSGT and MW would replicate in an online version of the task. Hence, study 1 was kept as identical as possible to the original task (Boayue et al., 2021). Nevertheless, minor differences like a comprehension quiz, description of randomness, video demonstrations, and audio tests were automated and incorporated online to increase the likelihood of reliability. Additionally, we provided statements reminding participants of a bonus payment for adequate performance and a warning stating erroneous performance was recorded. These statements were provided to increase motivation.

In study 2, we introduce performance feedback during the training session only ("baseline study") to increase our participant's understanding of the task. Thus, the performance feedback training works as a hurdle to be overcome. It should clarify the task and prepare participants for the experiment. The remaining four studies (study 3 to study 6)

keep the same performance feedback during the training session and explore the effects of other, additional feedback. Moreover, future studies can also apply this performance feedback training in the lab condition to improve task comprehension.

The performance feedback was meant to (1) ensure that the participants understood how to perform the task, thus, receive a high score, (2) allow the participant to try different performance tactics and observe how the performance feedback changes in return, and (3) encourage them to perform well in the following experimental task. Crucially, study 2 did not implement any feedback once the training session was completed. Therefore, we used study 2 as a baseline to compare the remaining studies to; all remaining studies used the same performance training as study 2 in the training session before the onset of the experimental session that contained different feedback interventions.

In part 2, as Crump and colleagues (2013) proposed that performance feedback might reduce MW, we continued using the specific performance feedback intermittently throughout the whole experiment session in study 3. Specifically, the specific feedback was displayed in three visual-analogue scales reflecting how well participants performed concerning BV, AE, and both measures combined relative to optimal behaviour. When participants intermittently gain access to their ongoing performance, this should allow them to continuously modify their responses to receive a higher score. As a control, we implemented non-specific feedback in study 4 and measured if just any feedback could improve the behavioural score on the task. Hence, in study 4, we delivered non-specific feedback encouraging them to keep on performing well. This was to control for a potentially unspecific motivational effect of the performance feedback. Importantly, this non-specific feedback did not reflect task performance and was provided at the same frequency as the performance feedback.

In part 3, we investigated how feeling monitored, or awareness of progression influenced the probe answers (Hróbjartsson et al., 2012; Mortenson, Sixsmith & Woolrych,

2015; Villar, Callegaro & Yang, 2013). Being under surveillance can potentially influence people's behaviour on tasks (Mortenson et al., 2015) and Hróbjartsson and colleagues (2012) found that people who feel observed exaggerated their subjective answers in experimental trials. Hence, we investigate whether a global sense of being controlled (implemented by a supposed surveillance through the participants' webcam, study 5) improves focus on the task or reduces MW. Initially, the participant's face was displayed on the screen, suggesting that their eye movements were recorded. During the remainder of the task, the integrated camera light was kept on (while no eye movements were recorded).

Additionally, people report preferring the availability of progression information when taking online surveys, and this is usually thought of as a motivating factor encouraging participants to keep going (Villar et al., 2013). However, the display of a progress bar does not seem to significantly affect dropout rates if the progress bar is not manipulated to change progression speed (Villar et al., 2013). Moreover, we know that participants pay attention to available progression information based on differing dropout rates when manipulating the displayed progression speed. We induce a generally heightened sense of being in control of the situation (implemented by a continuously visible progress bar, study 6) and measure if this improves focus on the task or reduces MW. We compared the results with the earlier studies and measured whether this feedback, camera, or progression, increased perceived MW.

In summary, testing and replicating MW studies in the online condition contributes to validating the online experimentation methodology. Using this methodology makes the process of developing new knowledge more efficient additionally, it is reliable when peer-to-peer meetings are restricted. Recruitment and testing of participants online are far more efficient and therefore lower the bar for conducting replication studies that are only rarely done in lab settings (Palan & Schitter, 2018).

The critical effects studied in this thesis are between-subject effects due to the different feedback interventions. However, we also expect the within-study effects of how MW relates to the behavioural performance indices collected in the FT-RSGT to be stable across studies. Hence, we expect to replicate the previous in-lab study results six times in the online setting. Additionally, similar results across studies would endorse online experimentation for replication purposes. Our within-study hypotheses are:

H1. The MW probe response is expected to decrease with AE. I.e., lower randomness (AE) predicts off-task focus (coefficient for AE < 0).

H2. The MW probe response is expected to increase with BV. I.e., increased variability (BV) predicts on-task focus (coefficient for BV > 0).

H3. A block effect. I.e., MW increases with time on task (coefficient of block> 0).

Between studies, we expect our behavioural measurements (AE and BV) to be improved when using performance feedback (in study 3). For study 5 and 6, we expect increased self-reported MW.

1 General Methods

1.1 Participant Selection Pool

By exploiting existing recruitment platforms that allow the testing of participants efficiently, we have the opportunity to test and collect new data over night – or even within a few hours. Using these platforms, it is possible to use the anonymized ID assigned to each registered participant to conduct a longitudinal study or include them in a blocklist blocking them from multiple participation in similar studies (Crump et al., 2013).

Specifically, we used Prolific (www.prolific.co), an online crowdsourcing platform specializing in scientific studies (Palan & Schitter, 2018). Members on this site must go through a three-step authorization (number, e-mail, and ID, e.g., driver's license). This would

make it impossible to establish multiple accounts, or dummy accounts, to trick the system and join studies multiple times to earn more money. Moreover, members make themselves eligible by answering questions about themselves – this works as pre-screening for us. When we publish the studies, we specify our inclusion and exclusion criteria in prolific. As a result, prolific displays how many eligible members they have even before each study is published. After publishing, the eligible members have it displayed on their prolific home page, and prolific send an e-mail telling them a study is available.

We only recruited people who had answered the following questions on Prolific's pre-screening page: (1) Fluent in English, (2) aged from 18 to 50, (3) normal or corrected eyesight. All participants self-identified as healthy adults. Additionally, they answered affirmatively: (4) No mental illness/condition, (5) no mild cognitive impairment or dementia, and (6) no mental illness daily impact. Prolific automatically excluded participants who had already joined earlier online studies using FT-RSGT published by our group. If participants failed to complete the study, prolific was set to automatically recruit new participants until a total sample of 40 participants was completed in each study. Also, we used a blocklist to make sure the same participant did not participate in multiple studies using the same task, thus, resulting in unique and independent samples across studies.

Baseline payment for joining the studies was advertised to be 3 £ (GBP) and additionally we instructed that if they performed adequately, they would receive a bonus of 2 £. For ethical reasons, every participant not excluded due to our criteria received the bonus. The total duration of each experiment was calculated to be about 30 minutes.

Lastly, we checked the resulting data files for people who did not comply with our instructions. We excluded participants based on the following criteria: (1) Paused tapping buttons three or more times during the experiment, (2) tapped the buttons at another frequency than once at each beep, (3) switched windows away from their browser and the

experiment more than ten times during the experiment, (4) spent time away from the experiment for more than 10 minutes once the experiment was started. Some switching between windows and time away from the experiment was allowed to allow participants to switch off e-mail notifications and get ready for the study.

Before participants who seemed to fulfill these exclusion criteria were excluded, we investigated their datafile. They were not excluded if the switching, inconsistent tapping, stopping, or time away from the experiment occurred during the instructional session. Moreover, if only short switching between blocks occurred, they were not excluded. Participants were informed that short brakes were allowed between blocks, no longer than one minute. This mirrors a typical protocol employed in the lab. However, participants who fulfilled the exclusion criteria were excluded, and new participants were recruited until a final sample of N = 40 was reached for each of the experiments.

1.2 Finger-Tapping Random Sequence Generation Task

To measure the neuro-behavioural signature of MW the FT-RSG task uses a metronome sound – or beep – at a predefined Inter-Stimulus Interval (ISI). Because we sought to collect as much data as possible over a short duration, we wanted the ISI to be as short as possible without losing the data quality. Out of a set of different ISI's tested, Boayue et al. (2021) found 750ms to be the optimal ISI, allowing participants time enough to produce a random sequence and not too long to lose focus between the beeps. Therefore, an ISI of 750ms is also used in our online MW experiments.

Simultaneously as the FT-RSGT presents a beep every 750ms the participants are instructed to press – or tap – one of two buttons on their computer keyboard, either "f" or "j" (coded 0 and 1 for analysis). As the task's name implies, these two buttons are to be pressed in a randomized order, thereby producing a sequence of f and j's in a randomized order. The

task usually does not start before participants have received some explanation of randomness. We used the example of repeatedly tossing a coin.

1.2.1 Explanation of Randomness

When tossing a fair coin, the outcome of one toss is entirely independent of the previous tosses. Even if three heads in a row have been observed, no information about the next toss is available, and another head is just as likely as a tail. Participants were informed that switching the heads and tails from this explanation with the buttons f and j on the keyboard would give them information on how to produce a random sequence. Flipping a coin in our head is, of course, impossible, and we have to rely on approximate computations to produce typical, "random" sequences. A fundamental characteristic of a random sequence is that all sub-sequences of different lengths have an equal probability of occurring. Hence, we can continuously try to produce sequences to avoid repeating patterns of different lengths. To achieve this, we have to continuously calculate and update our representation of the previously produced tapping sequences, which requires mental effort. Interestingly, once task focus is lost, we tend to produce an automatic sequence that is easily predictable (Teasdale et al., 1995). Due to this relationship, a non-random sequence indicates MW and low task focus, while a random sequence indicates task focus.

To illustrate what a random sequence is can be quite challenging because any observed sequences might result from a purely random process. Under this assumption, all possible observed sequences have equal probability (Pincus & Kalman, 1997). However, there are much fewer sequences with a clear structure, and these can therefore be considered less random. For example, tapping buttons every other time, f-j-f-j, and so forth, does not appear very random as one tap is predictable based on the previous taps. On the other hand, the sequence f-j-f-f-j-f-j-j-j may be more likely to come from a random process. We emphasized this in a demonstration video in the online experiment. Additionally, we included

three questions about this in a "comprehension quiz" to ensure participants understood the explanation. Participants had to answer all questions correctly before they were allowed to progress. In case of wrong answers, hints pop up, and they could try again.

1.2.2 Approximate Entropy

Because all possible sequences could result from random tapping measuring randomness is challenging. We measured randomness in terms of the approximate entropy measure (AE; Pincus, 1991). Approximate Entropy (AE) is a statistic defined for any sequence that measures the irregularity within the sequence (Pincus, 1991; Pincus & Singer, 1996). Higher AE values indicate higher irregularity and hence, higher randomness. AE is described as "the logarithmic frequency with which blocks of length m that are close together remain close together for blocks augmented by one position" (Pincus & Singer, 1996; Pincus & Kalman, 1997). In our task, the minimum block length was 40 (i.e., a minimum of 40 stimuli between two thought probes), and the maximum number of beeps was 80. Therefore, we can calculate the entropy for the sequence of taps occurring during the 40 – 80 taps preceding thought-probes. The first five beeps are never included in the calculation of AE because participants are allowed some time to get used to the pace of the beeps.

According to the definition, when using the statistic AE, we need to define the length of the sub-sequences (blocks of numbers) consisting of f and j, coded as 0 and 1, before calculating the randomness of the whole sequence. The parameter m indicates the length of these sub-sequences. Boayue and colleagues (2021) found the optimal sub-sequence length during the FT-RSG task to be m = 2 (zero included) when the ISI was set to 750ms. Hence, we look at a sub-sequence distribution of triplets, e.g., [0,0,0], [1,1,1], [0,1,0], [1,0,1] and so on. Moreover, the raw AE number is transformed according to equation 1:

$$AE_{trans} = Randomness = -\log\left(\log(2) - AE_{raw}\right) \quad (1)$$

1.2.3 Behavioural Variability

Behavioural Variability (BV) is a second measure indexing MW emerging from the FT-RSGT and is calculated as the log-transformed standard deviation of the inter-tap intervals. BV is a measurement of behavioural precision and is the statistic displayed during the precision feedback. Basically, it reflects how precisely or synchronously the participant is "tapping" the button, matching the sound of the ongoing metronome. Because we calculate the standard deviation of the taps, low BV scores reflect rhythmic tapping, and participants tap the button more or less synchronously with the beep. Contrasting, a higher BV score reflects irregularity in the tapping frequency (Seli et al., 2013).

1.2.4 Performance Training

The training session consists of five short AE (randomness) training blocks (see Figure 2A), five BV (precision) training blocks (see Figure 2B), and five training blocks displaying both performance scales and total performance (see Figure 2C). The performance feedback feeds the participants on their performance on the AE and BV measurements. The scores were presented as an intuitive visual analogue scale ranging from zero to one hundred to make the performance feedback understandable. Participants were instructed to try and achieve as high a value on the feedback scale as possible where the highest number, "100", reflected a full score. The raw AE values were automatically normalized, as stated above in equation 1, and our script multiplied the normalized AE value by 100 to display the participants estimated performance on the scale ranging from 0% random to 100% random.

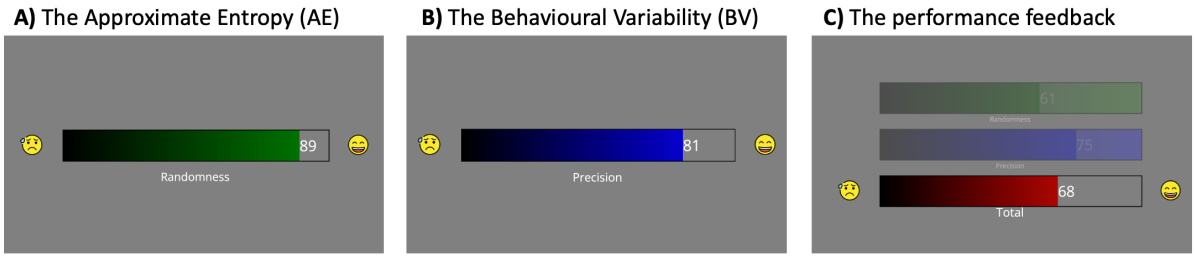
Participants received a full score on the BV scale when the Inter Stimuli Interval (ISI) matched the Inter Trial Interval (ITI) perfectly. A full score on the BV scale was achieved when the Standard Deviation (SD) was a maximum of 30ms. As the ISI is set to 750ms, participants could (typically) tap the buttons at an interval between 720 – 780ms and still

receive a full score. They received the worst score of zero if the SD deviated more than 200ms. Between the SD = 200ms and SD = 30ms, the score linearly increases from zero to hundred percent.

The total feedback score displayed both the randomness score and the precision score in the background and a highlighted total task performance score, see Figure 2C. This total feedback scale is what we refer to as the combined "performance feedback". This total score is calculated as the mean of the previously explained randomness and precision score. Suppose the participant performs excellently on the precision score but poorly on the randomness. In that case, they can observe this in the performance feedback and change their tactics to receive a higher total score.

Figure 2

Examples of the different feedback scales.



Note. A) The Approximate Entropy (AE) feedback scale indicates the randomness within the latest sequence of taps. B) The Behavioural Variability (BV) feedback scale indicates how precise the Inter Tap Interval (ITI) matches the Inter Stimuli Interval (ISI) calculated from the tapping sequence from the previous block. C) Displays a total score of the two main variables, randomness and precision.

1.2.5 Experience Sampling Thought Probes

We intermittently delivered thought probes during the task to probe the participants' meta-awareness about their task focus. The experimental task was automatically paused and

the question "where was your attention focused just before this question?" appeared. Next, the participant answered on a four-point Likert scale from 1, completely focused on the task (on-task) to 4, completely un-focused on the task (off-task).

Importantly, before this rating of one's thought was performed, participants went through detailed instructions and a thorough comprehension quiz to ensure they understood what they were asked about. An example given in the instructions of what is task-relevant is counting the keypresses. Consequently, this should be rated 1, completely focused on the task (on-task). In contrast, thinking of, for example, dinner, self-appearances, or progression through the experiment (e.g., "When will I finally be done?") should result in a rating of 4, completely un-focused (off-task).

1.2.6 Comprehension Quiz

During the instructional phase, participants went through a comprehension quiz to increase the likelihood of participants answering the experimental thought probes correctly. Specifically, the comprehension quiz consisted of three questions exemplifying what is categorized as on-task thoughts and off-task thoughts. Two questions tested understanding of randomness, and lastly, one question checking if participants understood that they should push either "f" or "j" together (synchronously) with the beep. If participants answered any of these questions wrong, a hint appeared, and they had to try again. Participants were only allowed to progress if all questions were answered correctly.

1.2.7 Server

We programmed our experiment using JavaScript based on the jsPsych library (de Leeuw, 2015). For hosting our experimental script, we used "Just Another Tool for Online Studies" (JATOS) installed on a university server (https://uit-jatos-test.azurewebsites.net). Primarily, JATOS makes it easier to test the experiments in different environments and web

browsers. Running studies from JATOS, the researcher can choose between different links to be shared with participants, links allowing for multiple participation, and links allowing for only one-time participation (Lange et al., 2015). This feature allowed us to integrate JATOS with our platform for recruiting participants, "Prolific".

1.3 Procedure

Participants were given the opportunity to adjust their PC volume on the experiment site while the experiment played a song. Interestingly, this use of music is thought to increase commitment by boosting the participant's mood (Shevchenko & Broder, 2019). With a comfortable volume, participants were sent onwards to an audio test controlling if participants indeed received audio from the experiment. The audio test consisted of five different animal sounds and pictures. One animal sound played, e.g., a dog barking, and the participant had to click the corresponding picture, a dog. To continue into the instructional session participants had to answer all five of these questions correctly, ensuring that the participants could hear the auditory stimuli. The audio test automatically restarted if they answered wrong, either by mistake or by trying to rush through the experiment.

Before starting the FT-RSG task, participants went through detailed instructions and training. The instructions explained randomness as with tossing a coin described earlier and displayed a demonstration video of how the button tapping would look in practice. To increase motivation participants were reminded of the bonus payment of 2 £ if they performed well and a warning if they did not comply. The warning stated that payment might be withheld if; (1) they stopped pushing buttons, (2) did not make an effort to produce a random sequence, (3) switched between windows, or (4) they interrupted the task. The warning ended with a statement describing that if they made an honest effort, this should be good enough to receive both the base rate fee and the bonus.

We included the same open-ended question in all studies asking participants to recall the most predominant thought occurring during the task for exploratory purposes. In addition, studies using specific feedback ended with a question asking whether they felt the feedback increased focus, disturbed them, or did not affect them. All participants who completed the study as instructed received the baseline fee of 3 £ plus the bonus of 2 £ resulting in a 5 £ payment.

1.4 Statistical Methods

We used the Bayesian hierarchical ordered-probit regression model to test the within-study hypotheses (i.e., the relationship between behaviour and MW). This method circumvents problems usually encountered when analyzing ordinal data with a metric model (Kruschke & Liddell, 2018; Liddell & Kruschke, 2018). The model includes BV, AE, their interaction, and probe number as predictors and MW as the (ordinal) outcome variable. Converting the raw AE values into a normal distribution with mean zero and standard deviation of one. As noted earlier, lower values of AE indicate less randomness, while a low value of BV indicates that the ITI is highly regular. Thus, we expected high AE values and low BV values during on-task focus. Both AE and BV were z-transformed using the overall mean and standard deviation (across subjects) to preserve between-subject variability. Analyzing the data, we used the "brms" package in R (Bürkner, 2021). Based on the model, we calculated the posterior mean and Highest-Density Intervals (HDI) of the regression coefficients, which can be interpreted as intervals with which the true coefficient falls with 95% probability, given the correctness of the model.

To compare the results between studies, we used the Bayesian t-test implemented in the BayesFactor package in R (Morey & Rouder, 2021). In all analyses, we set the prior standard deviation of the effect-size d to $\sqrt{2}/2$ before calculating Bayes Factors. This non-informative prior assumes that effect sizes are distributed according to a Cauchy distribution

with a scale of 0.707. Thus 50% of the prior mass falls between -0.707 and 0.707. Accordingly, this distribution is suited to discover small effect sizes near zero. However, large effect sizes are also possible as extensive data can "overwrite" information within the prior. Consequently, this prior incorporates the uncertainty within the data, unlike a point estimate of an effect of interest.

We expected AE to increase and BV to decrease during the experiment including performance feedback compared to the baseline. Furthermore, we expected the MW probe to be equal between the studies in part 1 and 2 and therefore use a BF_{01} that quantifies the evidence for the absence of an effect. We performed the Bayesian t-test between all studies expecting a difference in the behavioural measures in study 3 and a difference in the MW probe in study 5 and 6. Testing for a difference, we used BF_{10}, which quantifies the evidence for an effect. Describing the probability for H_1 to be true or alternatively H_0 to be true, we used the classification scheme proposed by Jeffreys (1961) as presented in Wagenmakers and colleagues (2011), see Table 1.

Table 1

Classification Scheme for the Bayes Factor, proposed by Jeffrey (1961) as presented in Wagenmakers et al. (2011).

Bayes factor, BF_{01}	Interpretation
>100	Extreme evidence for H_0
30-100	Very strong evidence for H_0
10-30	Strong evidence for H_0
3-10	Substantial evidence for H_0
1-3	Anecdotal evidence for H_0
1	No evidence
1/3-1	Anecdotal evidence for H_1
1/10-1/3	Substantial evidence for H_1
1/30-1/10	Strong evidence for $H1$
1/100-1/30	Very strong evidence for H_1
<1/100	Extreme evidence for H_1

Each individual study with datasets, experimental scripts, demographics data, and pre-registrations are available on OSF. The following is a list of repositories for all six studies:

- Study 1 (https://osf.io/fju92)
- Study 2 (osf.io/3j7v6)
- Study 3, performance feedback (osf.io/9dprs).
- Study 4, non-specific feedback (osf.io/tf5zw)
- Study 5, camera monitoring feedback (osf.io/r6j5z)
- Study 6, progress bar feedback (osf.io/p9h5y)

2 Study 1

In study 1, we used the previously developed FT-RSG task (Boayue et al., 2021) that required in-person attendance in the laboratory and converted it to work in the online setting (see General Methods). We expected to replicate Boayue et al. (2021)'s results that BV would

increase together with MW and that AE would decrease with MW. Replicating these in-lab results online would provide a "proof of concept". Accordingly, our focus in this study was to make an identical online version of the in-lab version and to determine whether our implemented control measures and comprehension questions ensured that participants paid attention and understood our instructions.

2.1 Participants

In our diagnostic, we measured total time in the experiment window, time used during instructions, number of taps in total, number of "f" taps, number of "j" taps, if they stopped responding, number of probes answered, number of times participants switched between windows (blurs), the blur duration and full-screen exit.

Out of the 40 participants recruited, one received a warning for stopping responding three times, and additionally, we measured 16 blurs. We see from the diagnostic that these three warnings were provided in the middle of the experimental task. As the excluding criteria states, this is not allowed. Hence, the participant was excluded. Following we opened the experiment for one additional participant. The final sample was N = 40, aged from 20 to 43 (M = 28), with a gender ratio of 22 males and 18 females. Detailed demographic data displaying acceptance/rejection, sex, nationality, the inclusion criteria, and more is available in the study repository on OSF (https://osf.io/fju92).

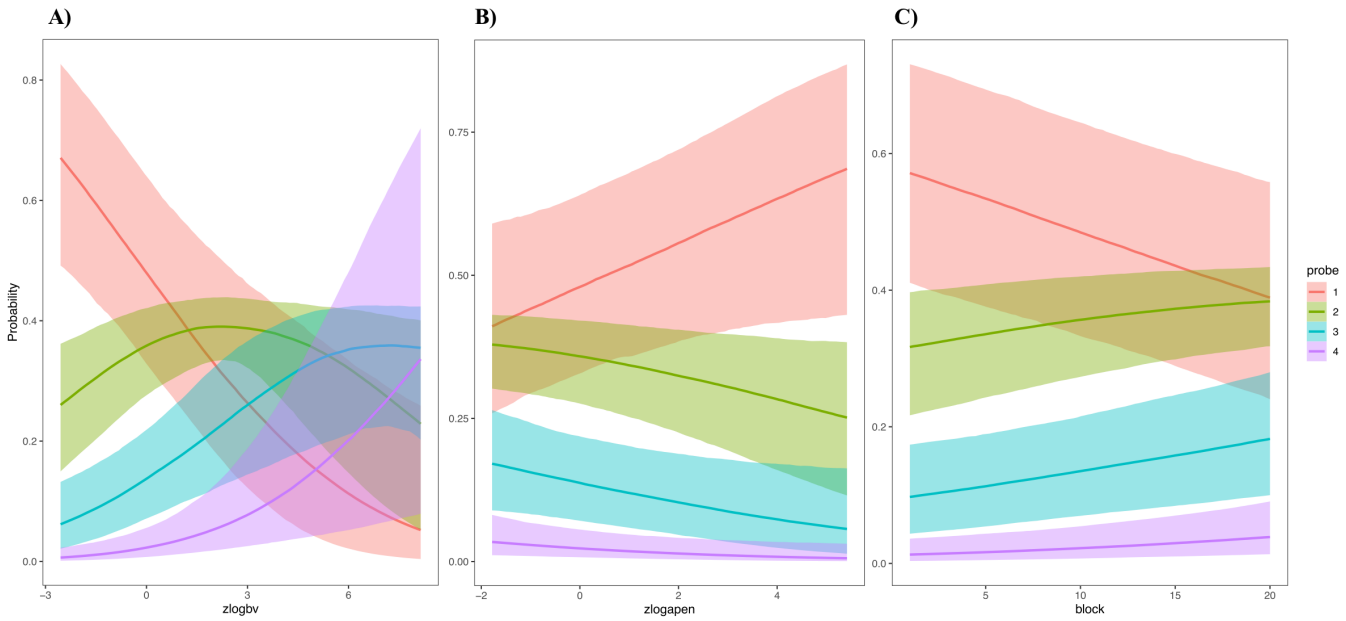
2.2 Results

As expected, according to our hypothesis (H1), we found a clear indication for the self-reported MW probe to decrease with the randomness score (AE), β = -0.09, [-.17, 0.00], ER_{-} = 22.53, see Figure 3B. Less randomness (AE) predicted more MW, while high randomness score increases the probability to answer probe 1 (on task focus). In addition, higher precision accuracy (i.e., lower BV values) predicted less MW (H2), β = 0.20, [0.11,

0.29], $ER_+ = 1999$, see Figure 3A. Finally, we confirmed the time-on-task effect that over time, the probability to self-report MW increases (H3), $\beta = 0.02$ [0.01, 0.04], $ER_+ = 3999$, see Figure 3C.

Figure 3

Within study 1, "proof of concept" effects.



Note. Posterior prediction plots for study 1, "Proof of Concept". A) Displays the z transformed logarithmic Behavioural Variability (BV) against the probability of answering probe 1 (on-task) – 4 (off-task). As BV increases, participants are less likely to rate ones thought to be on-task. B) Displays the z transformed logarithmic Approximate Entropy (AE) against the probability to answer probe 1 (on-task) – 4 (off-task). Increased AE, i.e., randomness within the sequence, makes the participant more likely to rate one's thoughts to be on task. C) Displays the time on task effect. The more time used on the task, the more likely the participants are to rate one's thought to be off task.

2.3 Discussion and Conclusion

We replicated the in-lab results (Boayue et al., 2021), thus demonstrating the robustness of the task. The connection between AE – MW and BV – MW was significant in our online condition where participants performed the experiment from home. Because we recruited participants online from around the world, we extended the representation to a broader sample, overcoming the limitation of most lab-based studies that focus on a sample of university students. Improving the experimental paradigm and because the participants have limitations in asking clarifying questions about the task online, we next, in study 2, include performance feedback in the training session, making sure and allowing the participants to clarify their task-understanding before the experimental session.

3 Study 2

Dandurand and colleagues (2008) found that participants online were less accurate than participants in the laboratory setting when performing a problem-solving task. We implement performance feedback during a training session to limit this tendency, increasing commitment to our sustained attention task. More than only indicating correct or wrong answers (Crump et al., 2013), we deliver the accuracy on an intuitively understandable zero to hundred scale.

Getting through the augmented training phase requires effort from the participants, and they have the opportunity to learn how their tapping strategy influences their score. Thus, this can also be implemented in the lab to improve the experimental paradigm. Subsequently, using this training, they can test their understanding of the task by checking their score after every training block. Moreover, this training can be perceived as a hurdle (Reips, 2002). A hurdle is a high-effort task at the beginning of an online experiment that serves as a barrier to removing participants that are not fully committed early on. This has advantages for both the

researcher, who avoids collecting data from unmotivated participants and the participants who can find out early that they are unwilling to complete the task. Moreover, participants who do not complete the experiment or are timed out because of inactivity are automatically excluded. This hurdle and excluding procedure should make the data more robust as the participants now are more likely to understand the task and are more committed to completing it (Reips, 2002).

Additionally, the provided feedback can be seen as an online equivalent to the informal clarifications given by the in-person experiment present in the laboratory condition (Feenstra et al., 2017). An added effect of this feedback protocol is a standardization of the experimental procedure. Hence, it reduces experimenter bias (Stickland & Suben, 2012). Thus, participants hopefully do not need to ask clarifying questions when using performance feedback training. Instead, they can try out different tactics and observe how that influences the performance score. Moreover, suppose participants experience distractions in their current environment affecting their performance. In that case, they have the opportunity to move to a more suitable environment with fewer distractions before the onset of the experimental session (Chandler et al., 2014).

We published this baseline study with the only difference from "proof of concept" being the performance training session because all subsequent studies use this same training. Hence, this study was published for comparability reasons, i.e., we wanted to isolate each manipulation (different feedback) in the following studies and needed a study where all else is constant.

3.1 Participants

Out of the 40 participants recruited, one consistently tapped the buttons more than once at every beep, sometimes even as many as six times at the sound of one beep. Thus, he did not follow our instructions and was excluded. Following we opened the experiment for

one additional participant. The final sample was N = 40, aged from 18 to 50 (M = 28), with a gender ratio of 24 males and 16 females. Detailed demographics data displaying acceptance/rejection, sex, nationality, the inclusion criteria, and more is available in the study repository on OSF (osf.io/3j7v6).

3.2 Results

As before, we replicated the hypothesized within-subject effects. We found self-reported MW to decrease with AE, β = -0.11 [-.20, -0.2], ER_{-} = 49. Participants were more likely to answer probe 1 (on-task) when randomness increased. We found variability in responding (BV) to increase together with probe responses, β = 0.19 [0.11, 0.28], ER_{+} = 3999. Hence, inconsistent tapping intervals predicted more off-task focus. Lastly, the time on task effect was positive, β = 0.05 [0.03, 0.06], $ER_{+} = \infty$. The longer a participant stays in the task, the more likely (s)he is to mind wander.

As expected and pre-registered, we found no significant difference between study 1 (proof of concept) and study 2 (baseline). We found substantial evidence for H_0 that there would be no difference between the BV measure in both studies, BF_{01} = 4.32. There was only anecdotal evidence for H_0, no difference between studies for AE, BF_{01} = 2.37. Finally, we also found substantial evidence that there was no difference in the subjective self-reported MW probe response, BF_{01} = 4.17.

3.3 Discussion and Conclusion

This study was conducted to improve the task paradigm, mainly when applied online. However, it can also be used in the lab as a general clarification task. This version allows the participants to try out the task, hence delivering specific feedback on the behavioural measures indicating performance in a training session. From this training, participants quickly receive information on whether they have understood the task instructions before

continuing to the experimental session. Next, we implement this performance feedback intermittently throughout the experimental session to improve task focus as measured by our behavioural indices.

4 Study 3

There are no simple solutions to how to keep participants attentiveness on a high level throughout an online experiment (Saravanos et al., 2021). To increase our participants' attentiveness throughout the task, in study 3, we implemented performance feedback intermittently throughout the entire experiment. Thus, the performance training from study 2 was displayed after every experimental block. Because one block consists of 40 – 80 beeps occurring every 750ms, the performance feedback was displayed every 30 – 60 minutes.

We hypothesized that delivering performance feedback intermittently after every experimental block would improve participants' behavioural indices of MW. This ongoing feedback indicating participants' performance could activate a competitive edge, motivating them to perform better than the last received performance feedback. In this way, we gamify the experiment making it more engaging (Marczewski, 2013). Additionally, Sailer et al. (2017) found that implementing performance graphs positively affects the perceived meaningfulness of the task. We did not expect any differences in the self-reported MW because of participants' novelty with the rating of their thoughts.

4.1 Participants

Out of the 40 participants recruited, three were excluded: Two because they tapped the buttons more than twice the amount required, often four times more than instructed. One because (s)he spent more than 60 min performing the task, went into a blur 36 times, and the longest blur was recorded to be 25 min. Following we opened the experiment for three additional participants. The final sample was N = 40, aged from 19 to 50 (M = 29), with a

gender ratio of 28 males and 12 females. Detailed demographics data displaying acceptance/rejection, sex, nationality, the inclusion criteria, and more is available in the study repository on OSF (osf.io/9dprs).

4.2 Results

Again, the within-subject effects were replicated as expected and described in the pre-registration. The effect of AE on MW propensity was negative, β = -0.09, [-.18, 0.00], $ER_- =$ 19.41. The effect of BV on MW was positive, β = 0.19 [0.10, 0.28], $ER_+ = 3999$. And lastly, the time on task effect was positive, β = 0.07 [0.05, 0.08], $ER_+ = \infty$.

As explained in the pre-registration, we expected a difference between the behavioural measures in this current study compared to the previous study 2. We found substantial evidence for BV to decrease, $BF_{10} = 5.57$. Surprisingly, we found no evidence for AE to increase, $BF_{10} = 0.96$.[1] Lastly, we found substantial evidence for the absence of a difference in the subjectively reported MW probe response, $BF_{01} = 4.28$.

4.3 Discussion and Conclusion

Implementing performance feedback seems to improve task focus as measured by BV relative to study 2 when such feedback was missing from the experiment. Surprisingly, the randomness score was not significantly different between this study and the previous one. Given that the feedback was supposed to increase our participant's understanding of the rather complex randomization task, it is surprising that it only affected BV, the easier of the two performance measures. Next, we control whether unspecific feedback intermittently delivered throughout the experiment in similar ways as the performance feedback can

[1] Differences between the other studies are investigated under section 9, "Joint Analysis".

improve task-focus similarly as observed in study 3 or whether that effect was specific to the nature of the feedback used there.

5 Study 4

Even though we found in study 3 that performance training increases performance relative to a baseline without feedback (study 2), we cannot rule out that any feedback delivered throughout the experiment could encourage participants to focus more. Therefore, we designed study 4 to include non-specific, positive feedback indicating that participants are doing well and should continue in the same way. We used a visual depiction of approval ("thumbs up") combined with a verbal statement, "Nice work!", see Figure 4. Because the feedback does not reflect performance, we do not expect any significant difference between this and the studies not using performance feedback throughout the experiment. Hence, the results from this study would reveal whether the feedback needs to be related to the performance or if any encouraging feedback will improve task focus.

Figure 4

Non-Specific Feedback.

Note. Displayed encouragement feedback not reflecting participants' performance.

5.1 Participants

The final sample was N = 40, aged from 20 to 47 (M = 28), with a gender ratio of 26 males and 14 females. Out of the 40 participants recruited, non were excluded. Detailed demographics data displaying acceptance/rejection, sex, nationality, the inclusion criteria, and more is available in the study repository on OSF (osf.io/tf5zw).

5.2 Results

Again, within-subject effects were intact and replicated. The effect of AE on MW propensity was negative, β = -0.07, [-0.15, 0.02], ER_{-} = 9.1. The effect of BV on MW was positive, β = 0.20 [0.10, 0.29], ER_{+} = 3999. And lastly, the time on task effect was positive, β = 0.04 [0.03, 0.06], ER_{+} = ∞.

Between study 3 (Performance) and study 4 (Non-specific), we found anecdotal evidence for BV to decrease in study 3 compared to study 4, BF_{10} = 1.50, and anecdotal evidence for AE to increase in study 3 compared to study 4, BF_{10} = 2.03. Lastly, we found substantial evidence for no difference in the subjective reported probe response (MW), BF_{01} = 3.78.

5.3 Discussion and Conclusion

Using non-specific feedback resulted in poorer task performance on both BV and AE measures compared to study 3 implementing performance feedback. In other words, providing general positive feedback not related to performance does not improve task performance, as reflected in increased BV and decreased AE. Again, there was a dissociation between objective performance and self-reported MW where only the former was affected by the feedback. Hence, to improve performance in the task, the feedback needs to be specific.

6 Study 5

Recent developments in web-based eye-tracking software enable using participants' integrated web camera to gain control during online experimentation (Papoutsaki et al., 2018). The software even makes it possible to monitor the gaze of the participant. Using camera monitoring in their online experiment Buso and colleagues (2021) made participants aware that they were being recorded. Fundamentally, their experiment compared results over three different conditions, participants in (1) the physical lab, (2) in the online condition with camera monitoring, and (3) in the online condition without camera monitoring. They only found small differences between conditions and concluded that data is valid across conditions. We focus on MW and specifically investigate whether camera monitoring online could increase task focus or if it is simply a distraction for the participants.

We hypothesized that camera monitoring would increase participants' self-reported MW as they would more often think about their performance or the purpose of being monitored. Thus, in study 5, we implemented "sham" camera monitoring. This should simulate the feeling of being monitored in similar ways as in the lab situation. Importantly, each participant continued into the experiment after centering their face on a displayed camera monitor box on their screen, purportedly to calibrate the eye-tracking equipment. No data from the camera were recorded either during or after their camera adjustments. However, the integrated camera light continued to be on during the whole experiment, thus, indicating surveillance.

6.1 Participants

The final sample was N = 40, aged from 18 to 47 (M = 28), with a gender ratio of 20 males and 20 females. Out of the 40 participants recruited, non were excluded. Detailed

demographics data displaying acceptance/rejection, sex, nationality, the inclusion criteria, and more is available in the study repository on OSF (osf.io/r6j5z).

6.2 Results

When applying the camera monitor feedback, the effect of AE on MW propensity did not reveal any significant relationship between AE and MW, β = -0.02 [-0.10, 0.07], ER_{-}= 1.55. However, the BV and block effect were replicated. Thus, both the effect of BV on MW, β = 0.12 [0.02, 0.23], ER_{+} = 30.75 and the block effect, β = 0.03 [0.02, 0.04], ER_{+} = ∞ were positive.

Between study 4 (Non-Specific) and study 5 (Camera), we found substantial evidence for no difference between BV measure, BF_{01} = 3.56, and substantial evidence for no difference between AE measurement, BF_{01} = 3.60. Finally, we found substantial evidence that MW probe response increased in study 5 compared to study 4, BF_{10} = 4.71.

6.3 Discussion and Conclusion

As hypothesized and specified in our pre-registration, we found substantial evidence that the self-reported MW probe responses increased when the camera feedback was included. However, there were no differences in the behavioural measures as we expected. It seems like the awareness of being monitored through the web camera makes the participant distracted and possibly more self-aware (Davies, 2005). As Hróbjartsson et al. (2012) report, the awareness of being monitored might make people exaggerate their answers. In this case, make them exaggerate their MW answers. Consequently, even if participants self-rate more MW it does not influence the task performance. It confounds the behaviour indices-MW effect, deteriorating the AE-MW effect while keeping the BV-MW effect constant. Hence, including control measures like camera recording affect subjective answers without affecting the task performance.

7 Study 6

In our final study, we wanted to uncover whether letting participants be aware of their progression through the experiment, thereby giving them more control over the experimental situation, would increase or decrease perceived task focus. Based on our results, this would serve as guidelines for how future online experiments should be designed. I.e., should the participants receive progression information or not (Villar et al., 2013). We modified the FT-RSGT from the baseline experiment (study 2) to include a continuously auto-updating progress bar at the top of the screen. The experiment ended with an opportunity for the participants to rate how the progress bar affected their focus. We hypothesized that the available progression feedback influenced the MW probe without affecting the behavioural measures, similar to study 5.

7.1 Participants

Out of the 40 participants recruited, none were excluded. The final sample was N = 40, aged from 18 to 49 (M = 27), with a gender ratio of 31 males and nine females. Detailed demographics data displaying acceptance/rejection, sex, nationality, the inclusion criteria, and more is available in the study repository on OSF (osf.io/p9h5y).

7.2 Results

When applying the progression feedback, the effect of AE on MW propensity did not reveal any significant relationship between AE and MW, β = -0.02 [-0.10, 0.06], ER_{-} = 1.55. However, the BV and block effect were replicated. Thus, both the effect of BV on MW, β = 0.16 [0.08, 0.24], ER_{+} = 3999 and the block effect, β = 0.03 [0.02, 0.05], ER_{+} = ∞ were positive.

We found no evidence for similar BV measures between study 5 (Camera) and study 6 (Progression), $BF_{01} = 0.54$. This is anecdotal evidence for the alternative hypothesis, see Table 1. BV seems to decrease in study 6 compared to study 5. Anecdotal evidence for no difference in AE measurement, $BF_{01} = 1.17$. Lastly, substantial evidence for MW probe response to be similar in study 5 and study 6, $BF_{01} = 3.83$.

7.3 Discussion and Conclusion

In similar ways to the camera monitoring feedback, progression feedback seems to distract participants resulting in more self-reporting of MW. Influencing the self-report in this way seems to confound the behaviour indices-MW effect as the AE-MW score did not replicate. We found that the BV score was lower, indicating more precise performance in study 6 than in study 5. It seems like progression is less distracting on performance, at least the precision, than when using camera surveillance. Still, it confounds the behaviour indices-MW effect. Finally, we report a joint analysis including all studies to investigate how AE, BV, and MW change depending on feedback.

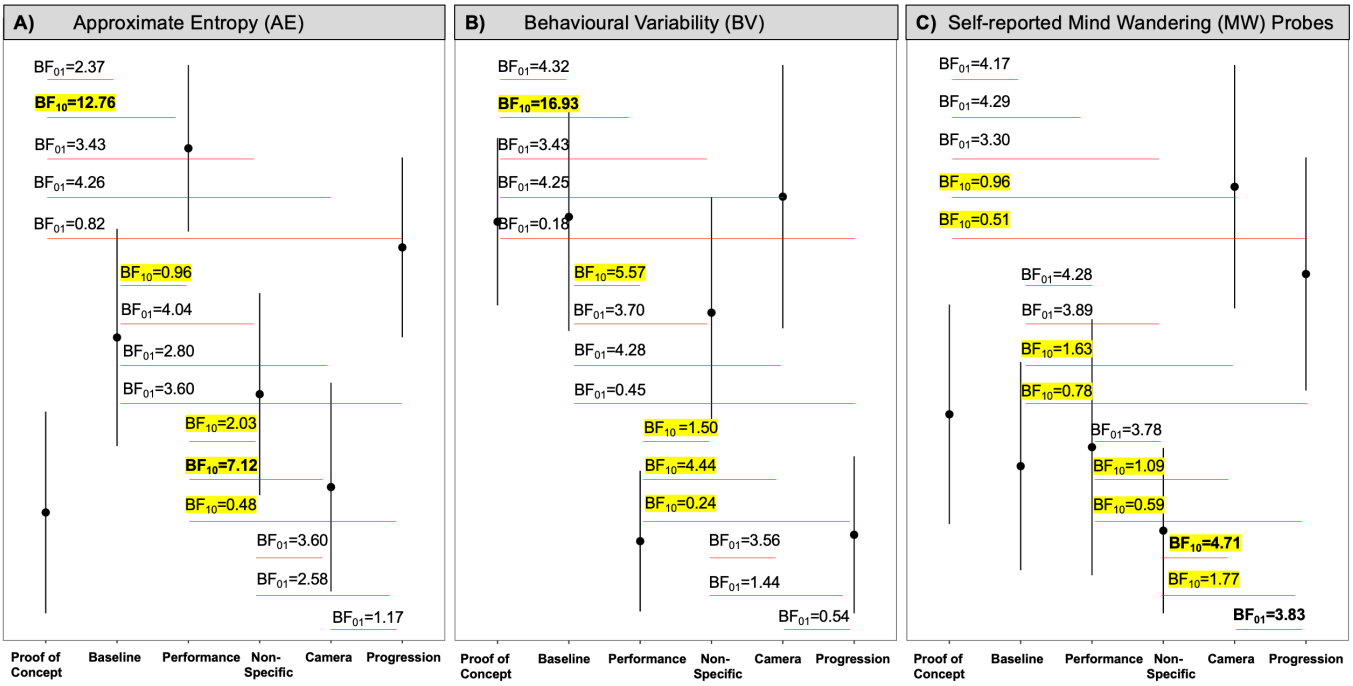
8 Joint Analysis

Figure 5 plots all studies on a graph comparing the AE, BV, and MW values across studies. We attach the t-test Bayes Factor indicating evidence for a difference (BF_{10}) or a Bayes Factor indicating evidence for no difference (BF_{01}). Studying Figure 5, we observe that participants who received performance feedback overall scored highest on AE measures (Figure 5A) and lowest on the BV measures (Figure 5B). We also see that BV improves in both the performance and the progression study, indicating that delivering performance feedback and displaying a progress bar help the participant be more precise in their performance. Lastly, studying Figure 5C, we see an indication for higher probe responses

when the camera monitor feedback is provided. Also, in the progression study, this seems to be the case.

Figure 5

Comparisons of the AE, BV, and MW between all studies.



Note. A study by study comparisons. A) Compares the Approximate Entropy (AE), the randomness score, across studies. B) Compares the Behavioural Variability (BV), the precision score, across studies. C) Compares the MW probe response across studies. BF_{01} = Bayes Factor testing for similar distributions. BF_{10} = Bayes Factor testing for different distributions. These are marked in yellow for clarification purposes. The error bars indicate the 95% highest density interval (HDI).

We observed strong evidence that performance feedback improved both AE and BV as expected. For example, in performance feedback study 3, the AE was approximately 13 times more likely to be increased, and BV was approximately 17 times more likely to be decreased than in the proof-of-concept study (study 1), see Table 2. This could indicate a better focus, or at least a better understanding of the task in study 3. Furthermore, we show

substantial evidence for the probe responses to be similar across study 1 – 4, BF_{01} ranging from 3.30 to 4.29, indicating that MW reports were similar on average.

Table 2

Performance Feedback study 3 compared with the previously conducted studies

Performance (Study 3) VS	Proof-of-concept (Study 1)	Baseline (Study 2)	Non-specific (Study 4)
AE	**BF_{10} = 12.28**	BF_{10} = 0.96	BF_{10} = 2.03
BV	**BF_{10} = 16.93**	BF_{10} = 5.57	BF_{10} = 1.50
MW	BF_{01} = 4.24	BF_{01} = 4.28	BF_{01} = 3.78

Note. AE = Approximate Entropy, a randomness measure. BV = Behavioural Variability, a precision measure. MW = Mind Wandering, self-reported through inter-spread thought probes. Meaningful comparisons referred to in the text are made bold. We used the t-test Bayes Factor for the comparison analysis and sat the prior distribution $R_{scale} = \sqrt{2}/2$. BF_{01} = Bayes Factor, testing H_0 against H_1. BF_{10} = Bayes Factor, testing H_1 against H_0.

We observed tendencies for the participants to report more MW in the camera monitoring study than the others. Hence, the activated camera makes participants perceive more MW. For example, we measured substantial evidence BF_{10} = 4.71 for an increase in reported MW compared to the non-specific study. On the other hand, AE and BV were equal in all studies except study 3. Study 3 showing substantial evidence for a difference against this tendency, AE BF_{10} = 7.12 and BV BF_{10} = 4.44, see Table 3.

Table 3

Camera Feedback study 5 compared to previously conducted studies.

Camera (Study 5) VS	Proof-of-concept (Study 1)	Baseline (Study 2)	Performance (Study 3)	Non-specific (Study 4)
AE	BF_{01} = 4.26	BF_{01} = 2.80	**BF_{10} = 7.12**	BF_{01} = 3.60
BV	BF_{01} = 4.25	BF_{01} = 4.28	**BF_{10} = 4.44**	BF_{01} = 3.56

MW	$BF_{10} = 0.96$	$BF_{10} = 1.63$	$BF_{10} = 1.09$	**$BF_{10} = 4.71$**

Note. AE = Approximate Entropy, a randomness measure. BV = Behavioural Variability, a precision measure. MW = Mind Wandering, self-reported through inter-spread thought probes. Meaningful comparisons referred to in the text are made bold. We used the t-test Bayes Factor for the comparison analysis and sat the prior distribution $r_{scale} = \sqrt{2}/2$. BF_{01} = Bayes Factor, testing H_0 against H_1. BF_{10} = Bayes Factor, testing H_1 against H_0.

As expected, we observed substantial evidence for similar results between the subjective probe answers across the progression and camera study, $BF_{01} = 3.83$. Compared to study 1-4 there was only week evidence for a difference, results only showing anecdotal evidence for a difference between progression and non-specific study, $BF_{10} = 1.77$, see Table 4.

Table 4

Progression Feedback study 6 compared with the other studies in the experimental series.

Progression (Study 6) VS	Proof-of-concept (Study 1)	Baseline (Study 2)	Performance (Study 3)	Non-specific (Study 4)	Camera (Study 5)
AE	$BF_{01} = 0.82$	$BF_{01} = 3.60$	$BF_{10} = 0.48$	$BF_{01} = 2.58$	$BF_{01} = 0.54$
BV	$BF_{01} = 0.18$	$BF_{01} = 0.45$	$BF_{10} = 0.24$	$BF_{01} = 1.44$	$BF_{01} = 1.17$
MW	$BF_{10} = 0.51$	$BF_{10} = 0.78$	$BF_{10} = 0.59$	**$BF_{10} = 1.77$**	**$BF_{01} = 3.83$**

Note. AE = Approximate Entropy, a randomness measure. BV = Behavioural Variability, a precision measure. MW = Mind Wandering, self-reported through inter-spread thought probes. Meaningful comparisons referred to in the text are made bold. We used the t-test Bayes Factor for the comparison analysis and set the prior distribution $R_{scale} = \sqrt{2}/2$. BF_{01} = Bayes Factor, testing H_0 against H_1. BF_{10} = Bayes Factor, testing H_1 against H_0.

9 Summary and Discussion

The purpose of this series of experiments was threefold: First, we wished to establish and confirm that FT-RSGT can detect MW in an online setting similarly to in the lab. Our second goal was to develop the task further by increasing task focus as measured by increased performance indicated by behavioural indices. Finally, we wanted to investigate additional feedback interventions and how they would influence both the perceived task focus and task performance.

Overall, our studies reveal that our in-lab sustained attention task can be reliably replicated in an online condition over multiple replications. The predicted within-subject effects reflecting the behavioural signature of MW in this task was observed in most of the studies (a notable exception was the absence of the AE-MW effect in study 5 and 6). Thus, online experimentation makes it possible to perform replication studies at a high pace. This efficiency and transparency reduce the concerns raised by the replication crisis (Simmons et al., 2011). Secondly, we found that intermittently delivering performance feedback improved the behavioural MW indices. Thirdly, displaying information not useful for the ongoing task increased the self-reported MW without affecting the task performance.

This dissociation between behaviour and subjectively reported MW in the performance feedback study could be because the participants are novel to the task and rate the task similarly when no other distractions are displayed. Hence, the similarity in subjective reported MW in study 1 to study 4. Furthermore, when displaying camera or progression, this could be perceived as distractions. They do not provide information on how well participants are performing the task. Interestingly, the display of camera or progression possibly triggers self-reflection (Davies, 2005) or reflection on time without deteriorating the performance.

However, there might be other explanations for the dissociation. For example, Hróbjartsson et al. (2012) found that participants exaggerated their answers when they were

aware of being surveilled. Perhaps this explains the increase in self-reported MW without affecting the behavioural indices. Moreover, O'Donnel, Ryan & Jettan (2013) found that surveillance led to higher productivity, but at the cost of the quality. In this case, we can consider the AE (which requires most executive recourses) as the quality measure. Hence, this could explain why the AE-MW effect disappeared in the camera study.

Camera data was not used or saved. There is a possibility that the participants did not sit in front of the computer during the whole experiment. They could, for example, collaborate with a friend, taking turns in performing the study. However, this collaboration is likely not true as it would be desirable for the participant not to get caught cheating. Besides, prolific participants are aware that their prolific score is reduced if their study attendance is rejected. This, in turn, reduces their eligibility for future studies. However, using control measures like camera recording or progression displayed to the participant confounds the link between behaviour and self-reports. Nevertheless, camera recording online continues to improve and is helpful if scientists want to investigate things like gaze direction and, possibly further in the future, pupillometry (Papoutsaki et al., 2018).

Improving focus under sustained attention tasks is relevant in many aspects of life: working from home on the computer, reading articles, watching lectures, i.e., activities requiring sustained attention on the task at hand. Adding informative feedback reflecting performance in such "boring" tasks is known as gamification (Marczewski, 2013). Including performance feedback should make online experiments more motivating, make participants receive the task better and result in more reliable data (Sailer et al., 2017). Interestingly, we received feedback from participants telling us it was fun and engaging. Intuitively, making a sustained task more fun and engaging reduces MW. Accordingly, the performance feedback group stood out as the best performing group with the best behavioural scores indicating the highest task focus compared to the other studies.

Even though recruiting participants can be done using a low payment rate ethically, researchers should consider the payment in the physical lab and adjust thoughtfully (Mason & Suri, 2012). Nonetheless, Crump and colleagues (2013) found that lower pay did not affect the data quality. Higher pay still ensures faster recruitment by making the study more financially attractive. In addition, higher payment ensures lower dropout rates and increases commitment and engagement to the task (Sarkar & Cooper, 2018). However, a financially attractive study increases the competition among participants to get into the study.

Scientists should not identify which participant produced which dataset to protect participants' identity. When testing in the lab, the scientists can perform preliminary analysis on the first participants. Often the early participants are someone the experimenter, assistant or recruiter knows. Consequently, it is easy for the one collecting the data to identify the participant if the preliminary analysis is conducted on the first few datasets. Contrasting, in the online condition identifying the person behind the dataset is impossible. Using prolific, we firstly use arbitrary identification numbers. All communications with the participants are under their non-identifiable number. Secondly, participants join from around the world decreasing the likelihood to casually know people who participate.

To increase control online, these experiments need additional automated procedures like audio tests, comprehension quizzes, and training sessions. Contrary, in-lab experiments usually perform similar procedures administered by a research assistant. Thus, online experiments should not require more time than in-lab settings. Additionally, the online procedures should be less prone to experimenter bias (Stickland & Suben, 2012). Furthermore, lab experiments can be more time-consuming as the experimenters want to properly use the participants once they have gotten them to participate in the experiment.

Lastly, future experiments should examine the relationship between behavioural indices indicating MW and self-reported MW as we influenced each of these separately with

different types of feedback. Evaluating this dissociation between self-reported MW and performance measures, it would be interesting to include other factors known to improve task focus. For example, categorize participants into different groups of familiarity with mindfulness practices. E.g., engages with mindfulness practices regularly, familiar with it, do not know/do not believe in it, or use mindfulness-based interventions (MBIs; Baer et al., 2019). As people who practice mindfulness are known to be better at being in the here and know and less prone to be distracted by environmental distractions (Schooler et al., 2014).

10 Conclusion

Online experiments investigating MW using sustained attention tasks are possible and replicable. Participants receiving performance feedback improve their task performance reflected in behavioural indices. Though, participants do not self-report being more focused when asked about task focus. This indicates that there is a behavioural and self-reflection dissociation. During sustained attention tasks, we can improve the behavioural indices without influencing the self-reports and, opposite, influence the self-reports without affecting the behavioural indices of MW. Similarly, we find camera surveillance and progression information to increase the self-report of MW without affecting task performance. These insights may also apply to other situations where increased task performance is desired.

References

Baddeley, A., Emslie, H., Kolodny, J. & Duncan, J. (1998). Random generation and the executive control of working memory. *The Quarterly Journal of Experimental Psychology Section A, 51*(4), 819–852. https://doi.org/10.1080/713755788

Baer, R., Gu, J., Cavanagh, K., & Strauss, C. (2019). Differential sensitivity of mindfulness questionnaires to change with treatment: A systematic review and meta-analysis. *Psychological Assessment, 31*(10), 1247-1263. http://dx.doi.org/10.1037/pas0000744

Boayue, N. M., Csifcsák, G., Kreis, I. V., Schmidt, C., Finn, I., Hovde Vollsund, A. E. & Mittner, M. (2021). The interplay between executive control, behavioural variability and mind wandering: Insights from a high-definition transcranial direct-current stimulation study. *The European Journal of Neuroscience, 53*(5), 1498–1516. https://doi.org/10.1111/ejn.15049

Bürkner, P.-C. (2021). Bayesian Item Response Modeling in R with brms and Stan. *Journal of Statistical Software*, *100*(5), 1–54. https://doi.org/10.18637/jss.v100.i05

Buso, I. M., Di Cagno, D., Ferrari, L., Larocca, V., Lorè, L., Marazzi, F., Panaccione, L. & Spadoni, L. (2021). Lab-like findings from online experiments. *Journal of the Economic Science Association, 7*(2), 184–193. https://doi.org/10.1007/s40881-021-00114-8

Chandler, J., Mueller, P., & Paolacci, G. (2013). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods, 46*(1), 112–130. https://doi.org/10.3758/s13428-013-0365-7

Cheyne, J. A., Solman, G. J., Carriere, J. S., & Smilek, D. (2009). Anatomy of an error: a bidirectional state model of task engagement/disengagement and attention-related errors. *Cognition*, *111*(1), 98–113. https://doi.org/10.1016/j.cognition.2008.12.009

Csikszentmihalyi, M. (2008). *Flow: The psychology of optimal experience.* New York: Harper Perennial Modern Classics.

Csikszentmihalyi, M., Larson, R., & Prescott, S. (1977). The ecology of adolescent activity and experience. *Journal of youth and adolescence*, *6*(3), 281–294. https://doi.org/10.1007/BF02138940

Dandurand, F., Shultz, T. R., & Onishi, K. H. (2008). Comparing online and lab methods in a problem-solving experiment. *Behavior research methods, 40*(2), 428–434. https://doi.org/10.3758/brm.40.2.428

Davies M. F. (2005). Mirror and camera self-focusing effects on complexity of private and public aspects of identity. *Perceptual and motor skills*, *100*(3 Pt 1), 895–898. https://doi.org/10.2466/pms.100.3.895-898

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods, 47*(1), 1-12. https://doi.org/10.3758/s13428-014-0458-y

Dixon, P., & Bortolussi, M. (2013). Construction, integration, and mind wandering in reading. *Canadian journal of experimental psychology*, *67*(1), 1–10. https://doi.org/10.1037/a0031234

Farley, J., Risko, E. F., & Kingstone, A. (2013). Everyday attention and lecture retention: the effects of time, fidgeting, and mind wandering. *Frontiers in psychology*, *4*, 619. https://doi.org/10.3389/fpsyg.2013.00619

Feenstra, H. E., Vermeulen, I. E., Murre, J. M., & Schagen, S. B. (2017). Online cognition: factors facilitating reliable online neuropsychological test results. *The Clinical neuropsychologist, 31*(1), 59–84. https://doi.org/10.1080/13854046.2016.1190405

Hróbjartsson, A., Thomsen, A. S., Emanuelsson, F., Tendal, B., Hilden, J., Boutron, I., Ravaud, P., & Brorson, S. (2012). Observer bias in randomised clinical trials with binary outcomes: systematic review of trials with both blinded and non-blinded outcome assessors. *BMJ, 344,* e1119. https://doi.org/10.1136/bmj.e1119

Jackson, J. D., & Balota, D. A. (2012). Mind-wandering in younger and older adults: converging evidence from the Sustained Attention to Response Task and reading for comprehension. *Psychology and aging*, *27*(1), 106–119. https://doi.org/10.1037/a0023933

Jeffreys, H. (1961). *Theory of probability.* Oxford, England: Oxford University Press.

Killingsworth, M. A., & Gilbert. D. T. (2010). A Wandering Mind Is an Unhappy Mind. Science. *American Association for the Advancement of Science, 330*(6006), 932–932. https://doi.org/10.1126/science.1192439

Kucyi, A., Hove, M. J., Esterman, M., Hutchison, R. M., & Valera, E. M. (2017). Dynamic brain network correlates of spontaneous fluctuations in attention. *Cerebral Cortex*, *27*(3), 1831–1840. https://doi.org/10.1093/cercor/bhw029

Kruschke, J. K., & Liddell, T. M. (2018). Bayesian data analysis for newcomers. *Psychonomic bulletin & review*, *25*(1), 155–177. https://doi.org/10.3758/s13423-017-1272-1

Lange, K., Kühn, S., & Filevich, E. (2015). "Just Another Tool for Online Studies" (JATOS): An Easy Solution for Setup and Management of Web Servers Supporting Online Studies. *PloS One, 10*(6), e0130834–e0130834. https://doi.org/10.1371/journal.pone.0130834

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, *79*, 328–348. https://doi.org/10.1016/j.jesp.2018.08.009

Marczewski, A. (2013). *Gamification: a simple introduction*. Andrzej Marczewski.

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior research methods*, *44*(1), 1–23. https://doi.org/10.3758/s13428-011-0124-6

Morey, R. D., and Rouder, J. N. (2021). BayesFactor: Computation of Bayes Factors for Common Designs. R package version 0.9.12-4.3. https://CRAN.R-project.org/package=BayesFactor

Mortenson, W. B., Sixsmith, A., & Woolrych, R. (2015). The power(s) of observation: theoretical perspectives on surveillance technologies and older people. *Ageing and Society, 35*(3), 512–530. https://doi.org/10.1017/S0144686X13000846

Mooneyham, B. W., & Schooler, J. W. (2013). The Costs and Benefits of Mind-Wandering. *Canadian Journal of Experimental Psychology, 67*(1), 11–18. https://doi.org/10.1037/a0031569

Nussenbaum, K., Scheuplein, M., Phaneuf, C. V., Evans, M. D., & Hartley, C. A. (2020). Moving Developmental Research Online: Comparing In-Lab and Web-Based Studies of Model-Based Reinforcement Learning. *Collabra: Psychology, 6*(1). https://doi.org/10.1525/collabra.17213

O'Donnell, A. T., Ryan, M. K., & Jetten, J. (2013). The hidden costs of surveillance for performance and helping behaviour. *Group Processes & Intergroup Relations, 16*(2), 246–256. https://doi.org/10.1177/1368430212453629

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance, 17*, 22–27. https://doi.org/10.1016/j.jbef.2017.12.004

Papoutsaki, A., Gokaslan, A., Tompkin, J., He, Y., & Huang, J. (2018). The eye of the typer. *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, 1–9. https://doi.org/10.1145/3204493.3204552

Pincus, S. (1991). Approximate entropy as a measure of system complexity. *Proceedings of the National Academy of Sciences*, *88*(6), 2297–2301. https://doi.org/10.1073/pnas.88.6.2297

Pincus, S., & Singer, B. H. (1996). Randomness and degrees of irregularity. *Proceedings of the National Academy of Sciences*, *93*(5), 2083–2088. https://doi.org/10.1073/pnas.93.5.2083

Pincus, S., & Kalman, R. E. (1997). Not all (possibly) "random" sequences are created equal. *Proceedings of the National Academy of Sciences*, *94*(8), 3513–3518. https://doi.org/10.1073/pnas.94.8.3513

Ratcliff, R., & Henickson, A. T. (2021). Do data from mechanical Turk subjects replicate accuracy, response time, and diffusion modeling results? *Behavior Research Methods, 53*(6), 2302–2325. https://doi.org/10.3758/s13428-021-01573-x

Reips, U.-D. (2000). *The web experiment method: advantages, disadvantages and solutions*. In M. H. Birnbaum (ed.), Psychological experiments on the internet (pp. 89–114). San Diego, CA: Academic Press.

Sailer, M., Hense, J. U., Mayr, S. K., & Mandl, H. (2017). How gamification motivates: An experimental study of the effects of specific game design elements on psychological need satisfaction. *Computers in Human Behavior*, *69*, 371-380. https://doi.org/10.1016/j.chb.2016.12.033

Sarkar, A., & Cooper, S. (2018). Comparing paid and volunteer recruitment in human computation games. In *Proceedings of the 13th International Conference on the Foundations of Digital Games* (pp. 1-9). https://doi.org/10.1145/3235765.3235796

Schooler, J. W. (2002). Re-representing consciousness: dissociations between experience and meta-consciousness. *Trends in Cognitive Sciences, 6*(8), 339–344. https://doi.org/10.1016/S1364-6613(02)01949-6

Schooler, J. W., Mrazek, M. D., Franklin, M. S., Baird, B., Mooneyham, B. W., Zedelius, C., & Broadway, J. M. (2014). *The middle way: finding the balance between mindfulness and mind-wandering.* In Psychology of Learning and Motivation (Vol. 60, pp. 1–33).

Elsevier Science & Technology. https://doi.org/10.1016/B978-0-12-800090-8.00001-9

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. http://dx.doi.org.mime.uit.no/10.1037/a0030954

Seli, P., Kane, M. J., Smallwood, J., Schacter, D. L., Maillet, D., Schooler, J. W., & Smilek, D. (2018). Mind-Wandering as a Natural Kind: A Family-Resemblances View. *Trends in cognitive sciences*, *22*(6), 479–490. https://doi.org/10.1016/j.tics.2018.03.010

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Smallwood, J., & Andrews-Hanna, J. (2013). Not all minds that wander are lost: the importance of a balanced perspective on the mind-wandering state. *Frontiers in Psychology, 4*, 441–441. https://doi.org/10.3389/fpsyg.2013.00441

Smallwood, J., & Schooler, J. W. (2015). The Science of Mind Wandering: Empirically Navigating the Stream of Consciousness. *Annual Review of Psychology*, 66(1), 487-518. https://doi.org/10.1146/annurev-psych-010814-015331

Strickland, B., & Suben, A. (2012). Experimenter Philosophy: the Problem of Experimenter Bias in Experimental Philosophy. *Review of Philosophy and Psychology, 3*(3), 457–467. https://doi.org/10.1007/s13164-012-0100-9

Szpunar, K. K., Khan, N. Y., & Schacter, D. L. (2013). Interpolated memory tests reduce mind wandering and improve learning of online lectures. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(16), 6313–6317. https://doi.org/10.1073/pnas.1221764110

Teasdale, J., Dritschel, B., Taylor, M., Proctor, L., Lloyd, C., Nimmosmith, & Baddeley, A. (1995). Stimulus-independent thought depends on central executive resources. *Memory & Cognition, 23*(5), 551–559. https://doi.org/10.3758/BF03197257

Towse, J. N. (1998). On random generation and the central executive of working memory. *British Journal of Psychology*, *89*(1), 77–101. https://doi.org/10.1111/j.2044-8295.1998.tb02674.x

Robertson, I. H., Manly, T., Andrade, J., Baddeley, B. T., & Yiend, J. (1997). Oops!': Performance correlates of everyday attentional failures in traumatic brain injured and normal subjects. *Neuropsychologia*, *35*(6), 747–758. https://doi.org/10.1016/S0028-3932(97)00015-8

Unsworth, N., & McMillan, B. D. (2013). Mind wandering and reading comprehension: examining the roles of working memory capacity, interest, motivation, and topic experience. *Journal of experimental psychology. Learning, memory, and cognition*, *39*(3), 832–842. https://doi.org/10.1037/a0029669

Villar, A., Callegaro, M., & Yang, Y. (2013). Where Am I? A Meta-Analysis of Experiments on the Effects of Progress Indicators for Web Surveys. *Social Science Computer Review, 31*(6), 744–762. https://doi.org/10.1177/0894439313497468

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., Gronau, Q. F., Šmíra, M., & Epskamp, S. & Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*(1), 35–57. https://doi.org/10.3758/s13423-017-1343-3

Wilson, T. D., Reinhard, D. A., Westgate, E. C., Gilbert, D. T., Ellerbeck, N., Hahn, C., Brown, C. L., & Shaked, A. (2014). Just think: The challenges of the disengaged mind. *Science (American Association for the Advancement of Science), 345*(6192), 75–77. https://doi.org/10.1126/science.1250830

World Health Organization. (2020). *Critical preparedness, readiness and response actions for COVID-19: interim guidance, 4 November 2020* (No. WHO/COVID-19/Community Actions/2020.5). World Health Organization.