Increasing propensity to mind-wander by transcranial direct current stimulation: A registered report

Nya Mehnwolo Boayue[1], Gábor Csifcsák[1], Per Aslaksen[1], Zsolt Turi[2], Andrea Antal[2], Josephine Groot[1,3], Guy E. Hawkins[4], Birte Forstmann[3], Alexander Opitz[5], Axel Thielscher[6,7], Matthias Mittner[1*]

[1] *Department of Psychology, University of Tromsø*

[2] *Department of Clinical Neurophysiology, University Medical Center Göttingen*

[3] *Integrative Model-based Cognitive Neuroscience Research Unit, University of Amsterdam*

[4] *School of Psychology, University of Newcastle*

[5] *Department of Biomedical Engineering, University of Minnesota*

[6] *Danish Research Centre for Magnetic Resonance, Copenhagen University Hospital Hvidovre*

[7] *Department of Electrical Engineering, Technical University of Denmark*

Abstract

Transcranial direct current stimulation (tDCS) has been proposed to be able to modulate different cognitive functions. However, recent meta-analyses conclude that its efficacy is still in question. Recently, an increase in subjects' propensity to mind-wander has been reported as a consequence of anodal stimulation of the left dorsolateral prefrontal cortex (Axelrod et al., 2015). In addition, an independent group found a decrease in mind wandering after cathodal stimulation of the same region. These findings seem to indicate that high-level cognitive processes such as mind wandering can reliably be influenced by non-invasive brain stimulation. However, these previous studies used low sample sizes and are as such subject to concerns regarding the replicability of their findings. In this registered report, we implement a high-powered replication of Axelrod et al. (2015)'s finding that mind-wandering propensity can be increased by anodal tDCS. We used Bayesian statistics and a pre-registered sequential-sampling design resulting in a total sample size of

*Corresponding author: matthias.mittner@uit.no

N=192 participants collected across three different labs. Our findings show support against a stimulation effect on self-reported mind-wandering scores. The effect was small, in the opposite direction as predicted and not reliably different from zero. Using a Bayes Factor specifically designed to test for replication success, we found strong evidence against a successful replication of the original study. Finally, even when combining data from both the original and replication studies, we could not find evidence for an effect of anodal stimulation. Our results underline the importance of designing studies with sufficient power to detect evidence for or against behavioral effects of non-invasive brain stimulation techniques, preferentially using robust Bayesian statistics in pre-registered reports.

Keywords: mind wandering, tDCS, DLPFC, non-invasive brain stimulation

1. Introduction

Mind wandering can be tentatively defined as a shifting of the attentional focus from external task demands to internal thoughts (Smallwood & Schooler, 2006). Episodes of mind wandering are very common during activities of daily life (Killingsworth & Gilbert, 2010) and during experimental tasks. Depending on various factors such as task difficulty (Feng et al., 2013) and mood (Smallwood et al., 2009), the percentage of time we spend mind wandering is estimated to be between 30% and 50%. In recent years, much interest has focused on the neural basis of mind wandering (Mason et al., 2007; Christoff et al., 2009; Mittner et al., 2014). One consistent finding is that mind wandering involves the default-mode network (DMN; Raichle et al., 2001), a network of brain areas that are activated during internal mentation (Buckner et al., 2008; Andrews-Hanna et al., 2010; Andrews-Hanna, 2012). The finding that activity in these areas is increased has been replicated in several independent studies employing different tasks and methodologies (Weissman et al., 2006; Christoff et al., 2009; Mittner et al., 2014).

Less well understood is the role of the frontoparietal control network (FPN; Vincent et al., 2008; Spreng et al., 2010) which also seems to be involved in the initiation and sustenance of mind wandering (Smallwood et al., 2012). Several studies have linked perceptual awareness to the propagation of stimulus-induced neural activity to the FPN, representing a "global workspace" that provides conscious access to cognitive representations (for reviews see: Baars et al., 2013; Dehaene et al., 2006; Dehaene & Changeux,

2011). During mind wandering, Smallwood et al. (2012) argue that the FPN might determine the contents of consciousness and serve as a common workspace for both internally focused trains of thoughts (associated with the DMN) and externally-guided cognition (operated by the dorsal attention network; DAN). In this view, the FPN is a flexible network that contributes to switches between different modes of the brain: An internally directed, decoupled mode (DMN) and an externally-focused mode during which activity in the DAN is increased. The dorsolateral prefrontal cortex (DLPFC) is a key region of the FPN and has been hypothesized to be essential in initiating and sustaining internal trains of thoughts, consequently leading to attenuated processing of external stimuli (perceptual decoupling; Smallwood et al., 2012). Based on this theory, it can be hypothesized that modulating the excitability of the DLPFC could affect the frequency and/or length of mind-wandering episodes. However, because the FPN is supposedly crucial both for the maintenance of an externally-focused and an internally-focused state, it is theoretically unclear whether mind wandering would be facilitated or inhibited using neuromodulation.

Recently, three interesting studies (Axelrod et al., 2015; Kajimura & Nomura, 2015; Kajimura et al., 2016) investigated this question empirically using transcranial direct current stimulation (tDCS). This non-invasive brain stimulation technique is thought to be capable of inducing robust excitability changes in the stimulated neural tissue (Stagg & Nitsche, 2011) by modulating synaptic efficacy and inducing synaptic plasticity. Intriguingly, Axelrod et al. (2015) could show an increase in the propensity to mind wander (as measured by self-reports) during a sustained attention task when anodal tDCS was applied above the DLPFC relative to two control conditions, a sham (inactive) stimulation and stimulation of the occipital cortex. This finding would seem to support the theory reviewed above: Higher excitability of the DLPFC (induced by anodal tDCS) in this framework could lead to a better ability of the FPN to suppress distracting perceptual stimuli and/or to maintain the ongoing train of internal thoughts. Furthermore, Kajimura & Nomura (2015) and Kajimura et al. (2016) investigated similar questions in a different experimental setup and found a pattern of results that is complementary in the sense that they observed reduced frequency of task-unrelated thoughts after applying cathodal tDCS above the left DLPFC relative to anodal stimulation. Together these findings appear to provide evidence for Smallwood et al. (2012)'s theory and can be seen as a major advance in the understanding of the neural correlates of mind-wandering episodes.

The result that mind-wandering propensity can be influenced by tDCS has important implications both for basic neuroscience and in more applied settings. In the scientific literature, the finding has attracted the attention of several leading researchers (Fox & Christoff, 2015; Broadway et al., 2015), with 51 independent citations so far. In their commentary on Axelrod et al. (2015), Fox & Christoff (2015) argue that changes in meta-awareness induced by the stimulation of DLPFC might be responsible for the observed changes. Similarly, Broadway et al. (2015) are enthusiastic about Axelrod et al. (2015)'s finding and argue that it "[...] marks a new era for research into mind wandering and previews some of the insights that continued methodological advances will likely make possible". We believe that such strong endorsements from leading researchers in the field are likely to result in a surge of research activity building on Axelrod et al. (2015)'s result. From a more applied perspective, mind wandering has been, e.g., associated with accidents in car driving (Yanko & Spalek, 2014; He et al., 2011) and aviation (Wiegmann et al., 2005) and a technique that consistently and reliably allows to manipulate the propensity to mind-wander has thus great potential to avoid many of these human errors. Furthermore, ruminations, which may be seen as a special case of mind wandering, are core features of clinical conditions such as major depression or obsessive-compulsive disorder. Therefore, a technique to reliably influence such processes could open up exciting avenues towards better treatment alternatives.

However, all of these considerations rest on the validity and most importantly the replicability of the observed effects. Although the findings summarized above have great potential influence, the evidence so far is inconclusive because it is based on clearly underpowered studies. Concretely, the studies used a low sample size (about N=10-20 per group) such that the results could very well be the result of random fluctuations. In addition, even though Axelrod et al. (2015) replicated their main result in a second experiment, Kajimura & Nomura (2015) and Kajimura et al. (2016) failed to replicate Axelrod et al. (2015)'s findings when using anodal stimulation of the DLPFC relative to a sham condition (though the effect was in the expected direction and the replication was not a direct one). Based on these arguments, we believe that a conclusive, high-powered replication of Axelrod et al. (2015)'s finding is essential for establishing a sound basis on which future researchers can advance the understanding and application of tDCS in the setting of mind wandering (or avoid spending unnecessary resources should the effect prove to be unstable).

Pre-registered replications are considered to be the best way to establish a firm basis for the existence of an effect and they provide a rigorous way to avoid the problems underlying the low replicability rate in psychology (Simons et al., 2014; Chambers et al., 2014; Nosek & Lakens, 2014). The need for rigorous replication may be further motivated by the recent meta-analytical findings in the field of tDCS. After an enthusiastic explosion of studies applying tDCS to affect many cognitive functions and psychiatric diseases, recent meta-analytic studies draw much more cautious conclusions (Tremblay et al., 2014; Horvath et al., 2015a,b). In fact, Horvath et al. (2015b) question the very existence of any effect of tDCS on cognition. However, stimulation parameters and tasks are diverse and strong conclusions cannot be made at this point in time and Horvath et al. (2015b) conclude with an urgent call for more direct replications in the field of tDCS. Finally, a review focusing exclusively on stimulation of the DLPFC (the target region of Axelrod et al., 2015) found very variable effects and "[..] sometimes apparent conflicting results" (Tremblay et al., 2014). Clearly, direct, pre-registered replications are necessary to be able to identify findings that are reliable in this important field.

Our project aimed to replicate the finding reported by Axelrod et al. (2015). For this purpose, we conducted a multi-center study (measuring in Tromsø, Amsterdam, and Göttingen) using identical experimental setups following a pre-registered protocol in order to pool an appropriately large sample size. We used Bayesian methods to estimate the effect size of anodal stimulation and to establish success or failure of the replication attempt (Verhagen & Wagenmakers, 2014).

2. Methods

All materials, simulations and analyses are available in a public repository hosted by the Open Science Framework (OSF) at `https://osf.io/dct2r/`. The repository was registered (frozen) before data collection such that none of the materials can be covertly changed after data has been collected. The link to the registered version of the project is `https://osf.io/bv32d/`.

2.1. Participants

Participants were collected from the respective subject-recruitment facilities of three universities, the university of Tromsø (UiT), the university of

Amsterdam (UvA) and the university of Göttingen (UniGö). Ethical approval for the study was granted at all three universities. Based on our design analysis (see below), we applied a sequential data collection protocol (Schönbrodt & Wagenmakers, 2018; Schönbrodt et al., 2017) and set out to collect between at least 120 and maximum 192 participants (a minimum of 20 and maximum of 32 participants per stimulation condition and study site). Subjects who failed to provide a complete dataset for technical (e.g., failure of the equipment) or other reasons (e.g., experiment not completed) were excluded from the analysis and replaced by new subjects. Specifically, in order to be included in the experiment, all of the following conditions needed to be satisfied for a participant:

- the participant did not have any neurological/psychiatric diseases (based on self-report)
- participants did not have previous experience with tDCS (to increase the efficacy of blinding)
- the participant was between 18 and 40 years old
- the participant completed the experimental session
- the stimulation equipment was functional across the complete session
- the data collected by the experimental computer was complete
- the participant complied with the instructions

After recruitment, participants were randomly allocated to either a sham or an anodal DLPFC stimulation condition according to a randomization list.

2.2. Apparatus

As the experiment was conducted across three separate locations, we enforced similar conditions in the three labs by fixing specifications for the apparatus and environment (see `experimental_setup.pdf`). These were set up in collaboration with the authors of the original study to be as close to the original experiment as possible. First, we required a quiet room free from distracting elements. No one besides experimenter and participant was allowed to enter the room during the study. In addition, optimal lighting conditions was ensured (avoid, e.g., frontal lighting that may be disturbing). Standard 19" flat-screen monitors were used in the study and the size of the stimuli was adjusted by the experimental program to ensure that the stimuli were presented in equal size on the retina. The experimental computer

ran identical versions of PsychoPy (release 1.83.04; Peirce, 2007) and the experimental software and experimenters were encouraged to make sure that the computer did not run any unnecessary background processes. Finally, all participants wore earplugs to minimize the influence of environmental noise, which they inserted once they read the instructions and possibly asked questions.

We also provided comprehensive, standardized instructions for the experimenters (see `experimenter_instructions.pdf`) for running the experiments. All experimenters were required to read the instructions and practice testing on at least two pilot subjects before acquiring real data. Experimenter interaction were kept at a minimum and instructions were delivered electronically to ensure a standardized procedure. There were, however, opportunities for the participant to receive clarification and ask questions (prompted by the experimental computer). A list of possible questions and standardized answers that were given by the experimenters is available at `q_and_a.pdf`.

The study used the Sustained Attention to Response Task (SART) which is a variant of the Go/Nogo task that is very commonly used in mind wandering research (Smallwood & Schooler, 2006). In this task, numbers between 0 and 9 were presented in the center of the screen in quick succession. The participant was required to respond to each stimulus by pressing a button (Go-trials) except when the target number "3" was displayed. In this case, the response was to be withheld completely (Nogo-trials). No feedback about the correctness of a response was given and the stimuli stayed on screen for a fixed period of time, irrespective of the users' response. In the context of mind-wandering studies, brief self-reports ("thought-probes") were presented occasionally during the course of the experiment. These probes consisted of a single question, "To what extent have you experienced task-unrelated thoughts prior to the thought-probe?" and were answered on a scale from "1" (minimal) to "4" (maximal).

In accordance with Axelrod et al. (2015), stimuli were presented in black (RGB: [0,0,0]) on a gray background (RGB: [104,104,104]). The stimuli were presented in the center of the screen and covered 3 degrees of visual angle. The subject's distance to the monitor was fixed at 60 cm and the maximum length of the stimuli was readily determined to be 3.14 cm so as not to exceed 3 degrees. Stimulus duration was set to 1 s and an inter-stimulus interval of 1.2 seconds was used. We provided scripts that testet timing and size of stimuli (`teststimsize.py`) and required the experimenters in each lab to run these scripts before data acquisition to ensure comparability.

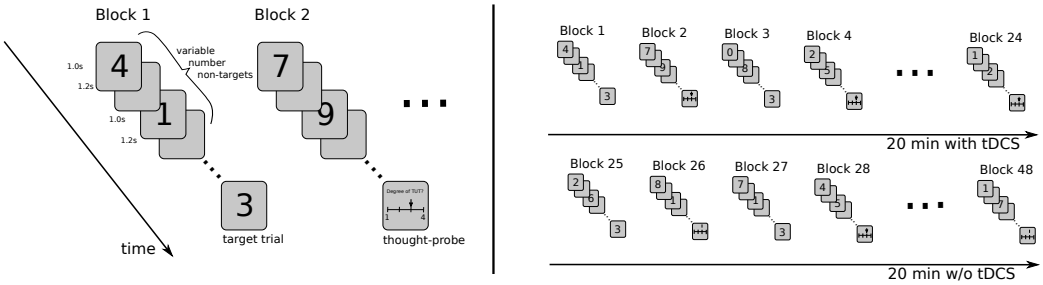

Figure 1: Sustained Attention to Response Task used in this study. The experiment consisted of two halves where tDCS stimulation was online in the first half and turned off in the second. Each half consisted of 24 blocks of trials ending in either a target or a thought-probe. The number of non-target trials was variable in each block. For details see text.

Participants were required to put both hands on the space-key and respond to the stimuli by pressing it (using whatever hand they prefered). They were asked to balance their performance between response speed (Go-trials) and accuracy (omissions in Go- and false alarms in Nogo-trials). At regular intervals during the course of the experiment, thought-probes consisting of a question and a visual scale from 1-4 (see Fig. 1) were presented. When a thought-probe appeared, participants were asked to press a number between 1 and 4 (on the keyboard) to indicate their level of task-unrelated thoughts. Self-report questions were presented for 6 s during which subjects could adjust their response (by pressing one of the keys corresponding to numbers 1-4). After each key press, an arrow appeared above the pressed number to indicate the currently chosen response. After 6 s, the screen was cleared if there was a response and the experiment continues. If no key was pressed for 6 s, the thought-probe remained on screen until a key was pressed.

The total duration of the experiment was around 40 minutes. During the first 20 minutes, participants received tDCS, the second half of the experiment was without stimulation. The original study (Axelrod et al., 2015) used a marked underrepresentation of target-stimuli. In their experiment, they presented a total of 24 targets while approximately 1000 non-targets were presented. We used the same procedure and to ensure that both halves contain an equal number of trials of each type, the following trial-randomization procedure was employed:

- the number of thought-probes was fixed at 24, 12 per 20 min period

- the number of target trials (Nogo-trials) was fixed at 24, 12 per 20 min period
- given these constraints and a total duration of 40 minutes, 1000 non-target trials were presented: 24 thought-probes × 6 s + 24 targets × (1.0s+1.2s) + 1000 non-targets × (1.0s+1.2s)=39 min, 57 s
- trial-presentation was divided into 48 blocks (not known to the participants) of unequal length
  - each block consisted of a variable number of non-target trials (mean 20, sd 5.69, min 12, max 29)
  - non-target stimuli were independently drawn from the set {0, 1, 2, 4, 5, 6, 7, 8, 9} with equal probability
  - each block ended either in a target-trial (stimulus "3") or a thought-probe
  - target-blocks and thought-probe blocks were presented in a pseudo-random manner so that 3 blocks with target stimuli and 3 blocks with thought-probes were appearing randomly in a set of 6 blocks ensuring that thought-probes were not presented exclusively at the beginning/end of the experiment, typically associated with reduced/increased frequency of mind wandering, respectively
- the number of non-targets across blocks was in addition constrained such that a total of 500 non-target trials were used across 24 blocks (such that the durations of the two halves of the experiment were identical)
  - this was achieved by repeatedly drawing 24 samples from a truncated normal-distribution (truncated to lie between 12 and 29) until the sum of their rounded values equaled 500
  - this procedure was repeated for each half of the experiment

Before the start of the experiment proper, there was a short training session of four blocks containing 2 targets and 2 probes (84 trials in total).

A Python-script using the PsychoPy library (Peirce, 2007) implementing this procedure is available at `sart.py`. Instructions were translated into Dutch, German and Norwegian by native speakers (complete instructions and the English template used to derive the local instructions can be found in `instructions_en.py`).

2.3. Additional measures

After completing the experimental procedure, participants were required to complete three questionnaires: One measuring the mood of the participants, a state-mindfulness questionnaire and an own questionnaire referring to the content of the mind-wandering episodes that the participants experienced. The analyses (e.g., correlations between questionnaire scores and thought-probe responses or parameters of task performance) carried out on these additional measures were not pre-registered and are reported as exploratory.

Similar to the study by Kajimura & Nomura (2015), The Positive and Negative Affect Schedule (PANAS; Watson et al., 1988) was used for measuring the mood of our subjects. We used this scale, because of the link between prefrontal activity, task-unrelated thoughts and emotion regulation: First, there seems to be a bidirectional causal link between mind wandering and negative mood states (Killingsworth & Gilbert, 2010; Smallwood et al., 2009). Second, there is converging evidence that the DLPFC plays a critical role in the top-down control of emotion (Okon-Singer et al., 2015), which is in accordance with the fact that symptom severity in major depression was quite consistently reduced by anodal tDCS applied over the left DLPFC (for reviews and controversies see: Brunoni et al., 2012; Berlim et al., 2013; Shiozawa et al., 2014). Finally, two recent study results showed that tDCS applied over the DLPFC can influence the frequency of ruminative thoughts of negative emotional content in healthy volunteers (Kelley et al., 2013; Vanderhasselt et al., 2013). In this regard, monitoring mood changes in studies investigating the effects of non-invasive brain stimulation on mind-wandering propensity seems to be inevitable.

The PANAS scale consists of 20 items (10-10 describing positive or negative emotional states), which are to be rated from 1 (very slightly or not at all) to 5 (extremely). Positive and negative mood scores are calculated separately, and these values are used to assess the current or past mood states of the participants. We hypothesized that increasing intensity of negative feelings during the experiment would be associated with an increase in mind-wandering propensity in the anodal tDCS condition. Therefore, we asked our subjects to complete the PANAS twice: First for measuring their current (post-SART) mood ("how do you feel right now"), and second, to retrospectively measure their baseline (pre-SART) mood ("how did you feel at the beginning of the experiment"). Given that the completion of the PANAS in itself might induce subtle mood changes, we decided not to use

it before the main experiment in order to avoid interference with the replication attempt. The PANAS scale is availabe in the Dutch (Engelen et al., 2006), German (Janke & Glöckner-Rist, 2014) and Norwegian (Gullhaugen & Nøttestad, 2012) languages and the translated versions were used at each of the three locations.

We also asked the participants to complete the Mindful Attention and Awareness Scale (MAAS; Brown & Ryan, 2003), which is a 15-item scale designed to measure an individual's disposition to attend to the present experience and overcome disrupting stimuli or internal states. It has previously been shown that MAAS scores negatively correlate with both the frequency of self-reported mind wandering and behavioral measures (e.g. response time variability, SART errors) of mind wandering (Mrazek et al., 2012). Because low MAAS scores are considered to be indicative of an increased mind-wandering trait that is stable over time (Brown & Ryan, 2003), MAAS scores are expected to correlate with mind-wandering frequency in the sham tDCS condition only. Moreover, the absence of correlations between the MAAS and self-reported mind-wandering propensity in the anodal tDCS condition would indicate that the effect of tDCS is independent of trait-like inter-individual differences. The MAAS is available in Dutch (Schroevers et al., 2008), German (Michalak et al., 2008) and Norwegian (Verplanken et al., 2007).

Finally, because periods of mind wandering are not uniform in nature and distraction from the task can be induced by disturbing external stimuli (Stawarczyk et al., 2011) such as tDCS electrodes placed on the forehead, we also asked the participants to freely report the content of their mind wandering during the task. We also used 4 additonal questions with 7-item Likert scales (1: not at all, 4: to a medium degree, 7: extremely) to estimate the degree to which participants were (1) thinking about task context (e.g., task difficulty, reflections on task performance, etc.), (2) distracted by tDCS (e.g., skin itching, tingling, skin wetness, etc.), (3) distracted by other stimuli (e.g., noises, visual stimuli, body sensations such as thirst or back pain, etc.) and (4) thinking about personal issues (e.g., past memories, future plans, etc.). Also, we asked the participants to guess whether they received real or sham stimulation using a 7-item Likert scale (1: sham, 4: don't know, 7: real). With these questions we aimed to exclude the possibility that the effect of tDCS on mind-wandering propensity was in fact related to the unpleasant sensations caused by the stimulation or by the participants' expectations about stimulation-related effects (Turi et al., 2014). This questionnaire and

a translation into the three local languages can be found at additional_questions_English.pdf.

2.4. Stimulation protocol

The stimulation protocol adhered to the one reported in Axelrod et al. (2015), with only minor modifications. All three labs used an identical model of the NeuroConn DC stimulator (https://osf.io/n4pbd/). To deliver the current, we used rubber electrodes (cathode: 7×5 cm; anode: 4×4 cm) with conductive paste (Ten20; Weaver and Company, USA). One of the electrodes was placed above position F3 (according to the International 10-20 system used in electroencephalography, EEG), the other above the right supraorbital area. The position of the stimulation electrode positioned at F3 was measured by applying the adequately sized EEG cap (circumference 56, 58 or 60 cm) on the participant's head. The EEG cap was chosen based on measuring the circumference of each participant's head. After marking the F3 position, the EEG cap was removed and the center of the stimulating electrode corresponded to the F3 position. In addition, the edges of both electrodes were precisely measured and marked which served as the landmark points for preparing the electrode-skin interface. The skin in the predefined surface regions were gently cleaned by using alcohol and cotton swab without overabrading the skin. A small amount of conductive paste was homogeneously distributed over the previously cleaned skin surface and the rubber electrode surface to ensure good contact between them. The electrodes were pressed firmly with medium pressure to the head in order to adhere the electrodes to the skin. To ensure that the conductive paste was distributed only over the predetermined regions, the extra conductive paste was wiped-off. Connector position was from anterior to posterior direction for the F3 electrode, and from right supraorbital to right temporal lobe direction for the return electrode. Impedance values were kept below 10 kΩ, subjects exceeding this threshold were not included in the study.

In the anodal stimulation condition, participants received 20 minute-long continuous stimulation at 1.0 mA intensity with 30 s fade-in and 30 s fade-out periods, whereas the sham protocol applied the fade-in and fade-out periods and the minimum possible stimulation duration of 15 s. As the study uses double-blind design, the stimulators ran in study-mode where each stimulation protocol was arbitrarily linked to a letter and secured with a 5-digit code. The Neuroconn DC stimulator has certain hardware limitations, that did not allow standard blinding using the 5-digit codes if the exact

stimulation parameters described by Axelrod et al. (2015) were to be used. More specifically, the pseudo-stimulation mode accessible by the 5-digit codes produces a sham protocol with a stimulation duration of 40 s in addition to the fade-in and fade-out periods, which was not desirable. Therefore, part of the stimulator's display was covered with non-transparent tape to avoid the experimenter getting feedback about which condition was currently been run. Details about preparing and using the stimulator are available at `experimental_setup.pdf` and `experimenter_instructions.pdf`. The mapping between stimulator code and stimulation mode were only accessible to a single researcher from each lab that was also responsible for programming the device but not involved in data-acquisition.

2.5. Statistical Methods

We used exclusively Bayesian statistics because of their many advantages compared to the more commonly used null-hypothesis testing (NHST) approach (see e.g., Gelman et al., 2013; Kruschke, 2014). In addition, we report standard frequentist statistics for comparability with the original study.

All pre-registered analyses discussed in the following were implemented as scripts in the R programming language (R Core Team, 2015) using the BayesFactor package (Morey & Rouder, 2015) and Stan (Carpenter et al., 2017) as the modeling backend and R-packages `rstan` (Stan Development Team, 2016) and `brms` (Bürkner et al., 2017) for interfacing Stan from R. The replication and meta-analytic Bayes factors were calculated using code provided by Verhagen & Wagenmakers (2014) on their webpage (`http://www.josineverhagen.com/?page_id=76`). A listing of the exact version of R and all packages used are provided in the file `versions_used.txt` as generated by script `print_versions.R`. The analysis scripts were developed using data generated by pilot subjects using the final experimental software. After the data was collected, these scripts were supposed to be executed without changes (only the pilot data-files exchanged with the real ones) and the results reported. However, several minor adjustments to the analysis scripts were necessary because of coding errors and changes in the analysis-packages used. All such changes are summarised in the Appendix and details are available in the form of difference files in our OSF repository. Both the raw data and all output of the analysis scripts were stored and uploaded to OSF and the quantities described in the following sections reported in the results section of this paper.

2.5.1. Effect of anodal stimulation on self-reported mind wandering

The main result of this study concerns the comparison of the groups receiving sham and anodal stimulation of the left prefrontal cortex in terms of their mean self-reported thought-probe scores. The original study (Axelrod et al., 2015) found that propensity to mind-wander (as measured by the mean of a subjects' responses to all thought-probes presented during the experiment) was increased for subjects receiving anodal stimulation. We tested this prediction using a directed Jeffreys-Zellner-Siow (JZS) Bayes Factor (Rouder et al., 2009) that tests the hypotheses that (1) the effect is in the expected (positive) direction against the hypothesis that (2) the effect is either zero or in the unexpected (negative) direction. We supplemented the analysis with BFs quantifying the evidence in support of the hypothesis that the effect is positive or negative compared to exactly zero and an interval estimate for the effect size.

In particular, we first calculated a directed Bayes Factor, $BF_{directed}$, testing the hypothesis that the result of subtracting the mean thought-probe responses of the anodal group from that of the sham group is larger than zero against the hypothesis that it is less or equal to zero (Morey & Rouder, 2015). We used a prior with an r-scale parameter of $\sqrt{2}/2 = 0.707$ that assumes that effect sizes are distributed according to a Cauchy-distribution with scale 0.707. This choice of prior was motivated by the fact that observed effect-sizes in tDCS studies are mostly small or medium (e.g., the absolute value of effect-sizes for cognitive effects of DLPFC stimulation reported by Horvath et al. (2015b) were on average 0.4). In case this BF is larger than 1, we found evidence for a positive effect of anodal stimulation. Values smaller than 1 quantify evidence for a negative effect. In case the real underlying effect-size is zero, the $BF_{directed}$ is likely to be inconclusive because there is similar amount of evidence for a positive or a negative effect, respectively.

Therefore, to better evaluate evidence for zero effect of stimulation, we calculated two BFs testing the hypotheses that the effect is zero, against the existence of a positive (BF_{null+}) or negative effect (BF_{null-}). We used the same prior distribution as before. BFs larger than one quantify evidence for the hypothesis that the effect is zero while a BF lower than one indicates evidence for a positive (BF_{null+}) or negative effect (BF_{null-}). Thus, while the previous $BF_{directed}$ directly tests the hypothesis predicted by the original study, this BF tests for the absence of any effect.

In addition, we used a final, undirected model (comparing any effect

against a null-effect) to extract an estimate for the posterior distribution of effect sizes which we quantified by its mean and highest-density interval (HDI). This estimate produced a range of values that contains the real effect size with 95% probability given that the model is correct and assigns probabilities to each of those values. Therefore, we can exclude values falling outside of the 95% HDI with high probability.

The four measures described so far are quantifying slightly different aspects of the data but are, of course, not independent. If the directional $BF_{directed}$ is large, we expect the posterior HDI to be mostly or completely positive, the BF_{null+} to be well below one and BF_{null-} to be inconclusive. Conversely, in case of high BFs in favor of the null-hypothesis, we expect a lower BF in favor of a positive effect and a posterior distribution (HDI) that includes zero.

In addition to these analysis, we calculated the replication Bayes Factor developed in Verhagen & Wagenmakers (2014). This Bayes Factor, $BF_{replication}$, pitches two competing theories against one another: A theory that a proponent of the original study might hold (i.e., that the replication effect size will be in line with the distribution of effect sizes implied by the original study) and a skeptic's null-hypothesis that the effect size does only deviate randomly from zero. The advantage of this BF is that it directly tests the question whether or not the results of the original study have been replicated or are more likely the result of random fluctuations. However, the test is likely to be inconclusive when the effect size observed in the replication is much lower than that from the original study (which is often likely, given the "significance filter" ensuring that published effect sizes that are based on low sample size are large; Gelman & Carlin, 2014). This is in line with the finding that underpowered studies might be unfalsifiable per se (Morey & Lakens, 2016). For this reason, we calculated this $BF_{replication}$ only as a secondary measure of replication success as it was likely to be inconclusive. Only when the difference between the original effect size and the obtained one is large enough compared to that between zero and the replication effect size, the replication BF favors the null-hypothesis instead of the presence of an effect.

Finally, we were interested in the total amount of evidence for the presence of an effect when pooling both the original study and the replication attempt (because the two studies are very similar, data can be assumed to be exchangeable). For this purpose, the fixed-effect meta-analytic Bayes factor BF_{meta} (Rouder & Morey, 2012) has been developed which merges the

original and the new data. The original study showed strong support for the presence of an effect, possibly because of the significance filter that ensures large effect-sizes of significant findings (Gelman & Carlin, 2014). Therefore, we expected the BF_{meta} to be biased in favor of a positive effect (Nuijten et al., 2015) and the results from the BF_{meta} received less weight when drawing conclusions from our analyses.

The script for the analyses described here is available at `anodal_mw.R`.

2.5.2. Design Analysis

The previous section described our main analyses that determine success or failure of this replication attempt. Based on these primary analyses, we conducted a design analysis based on simulations to find a sampling plan that would allow to find conclusive evidence for these measures.

In order to determine an appropriate sample size that allows to find an effect with high probability, we are required to specify a realistic effect size estimate. It is a well-known fact that published effect sizes that are based on small sample sizes and the criterion of statistical significance are inflated because of the "significance filter" (Gelman & Carlin, 2014): For an effect to become significant at low sample-sizes the effect must be large. We therefore thought it likely that the very strong effect of $d = 1.59$ reported by Axelrod et al. (2015) was an overestimate and that the real effect-size would be much lower. We note here, that the effect size reported in Axelrod et al. (2015) used a non-standard estimate of the pooled variance that accounts for differences in means and therefore results in the lower (though still huge) estimate of $d = 1.24$ that was reported in their study. In the field of tDCS, observed effect sizes are usually of small or medium size. The absolute value of effect-sizes for cognitive effects of DLPFC stimulation reported by Horvath et al. (2015b) were on average 0.4 (SD=0.59; median=0.29, meta-analytic mean=0.31, SD=0.41) and a recent preregistered tDCS study (which does not suffer from the significance filter) found an effect-size of $d = 0.45$ (Minarik et al., 2016).

We therefore designed our study to be able to detect effects in this range with appropriate probability and report a design analysis for a wide range of effect sizes. It has recently been proposed that underpowered studies are unfalsifiable (Morey & Lakens, 2016). These authors convincingly argue that even large discrepancies between an original, underpowered study and a (direct) replication study cannot be detected with high probability even if the replication study has infinite sample size. Accordingly, we choose to

base our power calculations not on the goal to replicate (or not-replicate) the original study but rather focus on estimating the real effect and of excluding the possibility of a zero effect while also analysing the expected distributions of the BFs.

Following (Kruschke, 2014), we ran a Bayesian power analysis where our primary goal was to exclude the null-hypothesis of an effect-size of $d = 0$ from the posterior 95% highest-density interval in the positive direction. Practical reasons did not allow us to exceed a sample size of N=192, such that each lab committed to collecting a maximum of N=64 subjects (32 per condition). In addition, we did not want to collect more data than necessary for ethical reasons. Therefore, we chose to apply a sequential design with a specified maximum sample size of N=192 (Schönbrodt & Wagenmakers, 2018; Schönbrodt et al., 2017). In order to avoid spurious rejections of the existence of an effect, we chose to first collect a minimum sample size of N=120 (20 per lab and condition). If the 95% posterior highest density interval (HDI) did not exclude zero at this point, we continued sampling until a maximum of N=192 had been reached. Once the initial 120 subjects were collected, we stopped after each batch of 18 subjects (3 per lab and condition) and evaluated whether the lower bound of the 95% HDI was larger than zero. If that would have been the case, we would have stopped data-collection; otherwise we would continue until the designated maximum (this was the case in our study, see Results). Note, that this was a directional stopping rule: We would only stop collecting data in case the HDI was fully positive. If it would have been fully negative, we would have continued sampling up to the full sample-size. The reason for this asymmetry was that a negative effect would have been surprising (given that we expected a positive effect) and we would have wanted to collect as much evidence for that as possible. The final posterior HDI was not biased in either direction, though.

In Figure 2, we provide a simulation-based analysis of this design. The simulation underlying this analysis proceeded as follows:

1. Pick an effect-size estimate d (we ran this simulation for effect sizes ranging between 0 and 1 in steps of 0.05)
2. For each d, run $n_{\text{rep}} = 10000$ simulations as follows:
    - generate a random dataset with an effect-size of d
    - following the sampling plan described above, calculate

a) the posterior HDI from the (undirected) Bayesian t-test described by Rouder et al. (2009) and implemented in Morey & Rouder (2015)
b) the Bayes Factors discussed above, $BF_{directed}$, BF_{null+} and BF_{null-}

and return the first N for which the lower bound of the HDI is above zero (or N_{max} if this did not happen), the associated BFs, the associated width of the HDI and whether or not the HDI excluded zero

3. Summarize/visualize the results for each effect-size estimate

The code for running this analysis and to produce Figure 2 is available at `power_sequential_hdi.R`.

Given this sampling plan, the probability of obtaining a false-positive, concluding that the HDI excludes zero even if $d = 0$, is 4.02%. The probability to find a conclusive HDI that excludes zero (power) is a function of the underlying real effect size (Fig. 2 a). For realistic estimates of the effect-size around $d = 0.4$, we have a power between 0.8 ($d = 0.39$) and 0.9 ($d = 0.46$). We could also determine the expected size of our sample (Fig. 2 b): With a real effect-size of 0.4, we had a probability to stop after the initial sample of N=60 per group of 0.54 and the probability to go to the maximum was 0.18. This illustrates the efficiency of this sampling plan as we had a good chance of being able to stop data-collection at an earlier stage. Figure 2 c) and d) show the distribution of the expected $BF_{directed}$, BF_{null+}, BF_{null-} and the expected width of the posterior HDI. At $d = 0.4$, the expected directional BF is around 86 and the expected width of the HDI around 0.7 (see Table 1). In case of a zero underlying effect size, the design is less efficient: the BFs in favor of the null-hypothesis were only expected to be of moderate size (around 6).

The analyses described so far used a Cauchy-distribution with scale parameter $r = \sqrt{2}/2$ as the prior distribution on the effect-size. The expected results for both the HDI and the BFs are not sensitive to the choice of this prior parameter. We re-ran the simulation described above for two other common choices of the scale-parameter, $r = 1$ and $r = \sqrt{2}$ and the effect on the outcome variables was minimal. This is due to the rather large sample even with the lowest possible sample size allowed by our sampling plan because the likelihood eventually overwhelms any reasonable choice of prior.

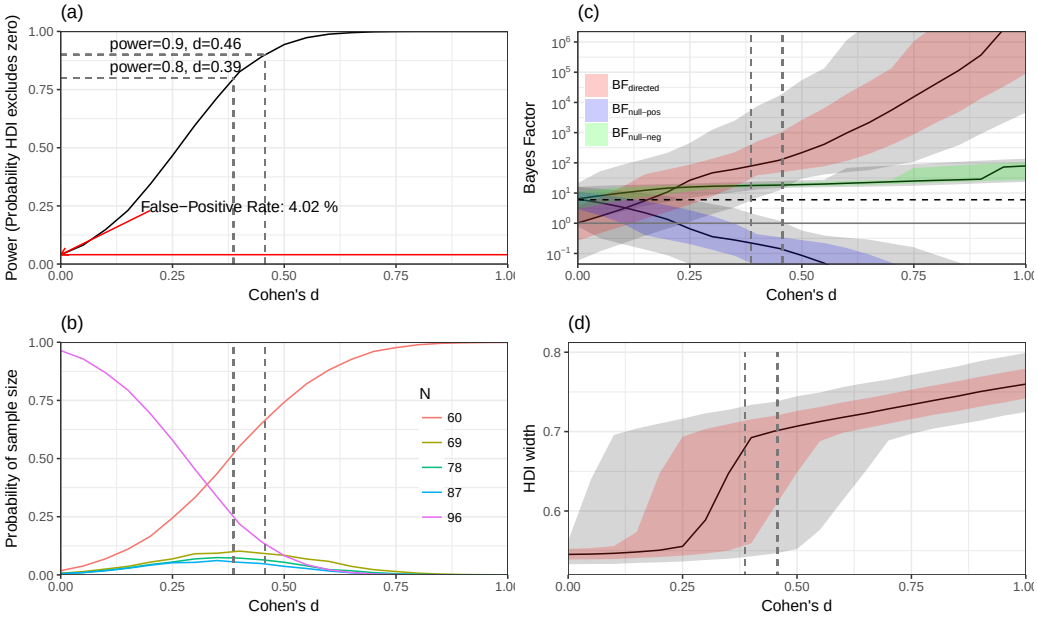

Figure 2: Design analysis for a sequential design with a maximum N of 192, an initial N of 120 and optional stopping after batches of 18 subjects in case the 95% HDI excluded zero. (a) Probability that the HDI excludes zero as a function of the real underlying effect-size. Dashed lines show the effect-size for which our sampling plan has 80% and 90% power, respectively. (b) Probability to collect samples of different sizes as a function of real effect-size. In case of a low real effect size, collection of the full sample of N=96 per group is highly likely while only the minimal N=60 per group will likely be collected if the effect size is large. (c) Distribution of BFs (both $BF_{directed}$ and BF_{null}) we are likely to find given the underlying effect size. Horizontal dashed line indicates BF=6. (d) The expected width of the posterior HDI given the underlying effect-size. Because needed sample size decreases with increasing effect-size, the width of the HDI increases as well. Colored and grey ribbons show 80% and 95% HDI for the respective parameter.

Table 1: Summary of the sampling plan in case of two hypothetical scenarios: The null-hypothesis is true ($d = 0$, left) and the real effect has an effect-size of $d = 0.4$ (right). If the null-hypothesis is correct, the directional BF, $BF_{directed}$, will be inconclusive as there is about the same amount of evidence for the effect being negative or positive, while both BF_{null+} and BF_{null-} are likely to be of moderate size. In the case of a small-to-medium effect size of $d = 0.4$, the $BF_{directed}$ results in compelling evidence while the BF_{null+} is less compelling (median $1/BF_{null+}$ only moderately in support of positive effect). The BF_{null-} shows compelling evidence for the null and is not easy to interpret when the real underlying effect is positive as it only compares evidence for negative and zero effect-sizes. The expected width of the HDI is about 0.55 in case of $d = 0$ but only 0.69 for the case of $d = 0.4$. This effect exists because sample size is maximal when $d = 0$.

	$d = 0$			$d = 0.4$		
	median	P(BF>6)	quantiles	median	P(BF>6)	quantiles
$BF_{directed}$	1.02	0.13	$[0.06, 21.4]$	86.2	0.96	$[6.97, 7473.6]$
BF_{null+}	6.3	0.52	$[0.78, 16.11]$	0.20	0.003	$[0.003, 1.88]$
$1/BF_{null+}$	0.16	0.01	$[0.06, 1.28]$	4.89	0.44	$[0.53, 310.5]$
BF_{null-}	6.45	0.53	$[0.93, 16.0]$	17.9	0.99	$[13.11, 24.1]$
$1/BF_{null-}$	0.16	0.006	$[0.06, 1.07]$	0.06	0	$[0.04, 0.08]$
HDI width	0.55		$[0.53, 0.56]$	0.69		$[0.54, 0.73]$
$P(\text{HDI} > 0)$	0.043			0.81		

2.5.3. Hierarchical ordered probit model

In addition to the aforementioned analysis, we analyzed the data using a novel analysis method that has not been used previously to analyze thought probe data. We used a hierarchical Bayesian model developed for analyzing rank-ordered data. In the previous analyses and in most if not all of the literature, mind-wandering thought-probes are first averaged within-subject before this average is submitted to the final between-subject analysis. This kind of analysis is problematic in at least three ways: First, it constitutes a "waste" of data because information about within-subject variability of responses to thought-probes is lost. Second, treating thought-probe responses as a metric variable is problematic because assumptions underlying the employed methods are likely not to be met. Finally, interesting and known effects on responding are ignored. Most prominently, an effect that is visible in all mind-wandering studies we have seen so far, is the time-on-task effect that is well-known to affect how likely subjects are to respond positively to mind-wandering probes (Thomson et al., 2014).

These points can be improved upon by using an appropriate model.

The first point, modeling within- and between-subject variability, can be accounted for by a hierarchical modeling approach where subject-level parameters are separately estimated while constraining these estimates by a group-level distribution. The second point (treating ordered variables as metric) can be improved upon by using an ordered probit model. A Bayesian implementation of such a model is described in Kruschke (2014) (Ch. 23). Basically, the assumption of an underlying metric (normal) variable is made which is thresholded by the participant into discrete response bins. In this setting, both the threshold and the parameters of the underlying distribution are estimated separately. Finally, covariates (e.g., time-on-task) can be easily integrated using this method.

To justify the need for these advanced analysis methods, we compared models of different complexity on a thought-probe dataset. Because we did not have access to Axelrod et al. (2015)'s original data, we used data from an unpublished study collected in our lab. In this study, we also used the SART paradigm (though using slightly different parameters, such as number of trials and targets). We also employed the same 4-point scale as used in the current study and 20 thought-probes spread out across the experiment were collected from each of 19 participants. A detailed description of this study can be found in `bsc_christian_fossheim.pdf`. We believe that this data, while not identical to the current study, could give an indication of the magnitude of within-/between-subject variation in responding to thought-probes.

In preparation of the analysis, we analyzed these data using a range of models of increasing complexity (code for fitting and diagnosing these models is available at `analysis/ordered_probit`). We compare the models based on their predictive performance using leave-one-out cross-validation (LOOC) and Watanabes information criterion (WAIC) implemented in the `loo` package (Vehtari et al., 2015) which are the state-of-the-art model-selection criteria for hierarchical Bayesian models (Gelman et al., 2014). These critera are reported on the deviance scale and differences of about 10 units are considered strong (Spiegelhalter et al., 2002). In general, LOOC is the preferred criterion, while WAIC can be a viable and computationally easier approximation to LOOC (Gelman et al., 2014) when calculation of the LOOC is not possible. For all reported models, LOOC and WAIC produced identical results and we therefore only report the former.

The first model uses a basic analysis strategy as a baseline, treating MW probes as metric and interchangeable across trials and subjects. Next, we im-

plemented an ordered-probit model where individual responses were treated independently. The comparison of these two models determined whether treating the data as metric was justified. The third and fourth model implement a hierarchical version of the first two models, where subject-level means are constrained by a group-level distribution. Comparing these two models to the first two can help to determine whether the explicit modeling of within- and between-subject variation is necessary. Finally, we added time-on-task as a covariate to the hierarchical ordered probit model. Table 2 lists the LOOC criterion (standard error in parentheses) for each of the models. It is clear that the ordered probit model more appropriately models the data than a model treating the data as metric both in the basic (ΔLOOC=34.1, SE=6.0) and the hierarchical case (ΔLOOC=31.9, SE=5.9). Finally, adding the covariate time-on-task strongly improves predictive accuracy, ΔLOOC=12.5, SE=5.0.

Table 2: Model selection criteria for models of increasing complexity. The hierarchical ordered probit-model including a time-on-task covariate is the most appropriate of the models. weights=posterior probability that each model has the best expected out-of-sample predictive accuracy; LOOC=leave-one-out cross-validation criterion. The model with the lowest LOOC is preferred.

Model	Description	LOOC (SE)	weight
1	metric	1116.8 (17.7)	0.0
2	ordered probit	1048.6 (6.3)	0.0
3	hier. metric	992.8 (22.6)	0.0
4	hier. ordered probit	929.1 (18.3)	0.0
5	hier. ordered probit + time-on-task	904.2 (20.2)	1.0

Based on these considerations, we chose the hierarchical ordered probit model that included a time-on-task covariate as the final analysis model. The model is mathematically fully specified in Appendix 1, including choice of the prior distribution, and implemented in the R-script `hier_ordered_probit.R`. We report and interpret all coefficients in terms of posterior mean and HDI.

2.5.4. Effect of location (lab)

Despite the uniform study design applied at all locations (UiT, UvA, UniGö), unknown contextual factors might cause substantial variability in effect sizes between the three labs. Therefore, we compared the tDCS effects resulting from the data from all three labs independently by calculating

independent estimates per lab for the full hierarchical ordered probit model presented in the previous section. These estimates in terms of posterior mean and HDI are presented side-by-side for comparing the variability in the different variables across labs. We also augmented the model with covariates for study location (UiT, UvA, UniGö). Comparing the posterior means for the location-coefficients and their HDI as well as a model-comparison analysis of the augmented vs. the non-augmented model enabled us to rule out or quantify location-specific effects. For details see Appendix 1. The script implementing these analyses is available at `location_effects.R`.

2.5.5. Frequentist analyses

For comparabililty with the previous literature, we also conducted standard two-sample t-tests on mean thought-probe responses for sham vs. anodal stimulation (both directed and undirected). We also report standardized effect sizes (Cohen's d) for these effects. These analyses are only conducted because they correspond directly to the analytical strategy chosen by the authors of the original study (Axelrod et al., 2015). Unfortunately, our sequential sampling scheme prevents us from calculating these statistics for the final sample as the stopping-rule invalidates the p-values. We therefore use only the guaranteed initial sample size of N=60 per group for this analysis. The script implementing these analyses is available at `frequentist.R`.

2.5.6. Exploratory analyses

To further assess whether mind wandering or other task-related measures were influenced by tDCS, we conducted five Bayesian repeated-measures analyses of variance (ANOVA) tests along with their frequentist equivalents with time (2 levels: first vs. second parts of the task, associated with online vs. offline effects, respectively) as within-subject and stimulation (2 levels: anodal vs. sham tDCS) as between-subject factors. This analysis design is identical to that used by the original study (Axelrod et al., 2015), which focused on three measures of interest, each entered as dependent variable in separate ANOVAs: Thought-probe ratings, mean reaction times for Go stimuli (GoRT) and mean error rates for Nogo stimuli (commission errors). We extended this analysis with two additional parameters: Reaction time coefficients of variation (RTCV) and error rates for Go stimuli (omission errors). RTCV was quantified as dividing the standard deviation by mean RT scores, calculated for both parts of the task and for each participant separately. Both RTCV and omission errors were proposed to index lapses of

attention during the SART, and therefore, are regarded as behavioral indices of mind wandering (Cheyne et al., 2009). All analyses within this section were done using JASP 0.9 (JASP Team, 2018). Bayesian tests were run with default prior scales of JASP (r scale fixed effects: 0.5). Interaction terms were assessed by comparing models including the effect to equivalent models without the effect ($BF_{inclusion}$). Based on the recommendation by Jeffreys (1961), we report results with BF values providing moderate evidence for either the alternative ($BF > 3$) or null-hypothesis ($BF < 0.33$). Depending on the type of variable (continuous vs. ordinal), correlations between behavioral measures were assessed by calculating either Pearson's or Kendall's correlation coefficients. To demonstrate effect size for frequentist ANOVAs, we report partial η^2 values. Given the exploratory nature of correlation analyses performed herein, the reported p values are not corrected for multiple comparisons and findings should be treated with caution.

3. Results

3.1. Demographics

Our sample consisted predominantly of females (70%, 134/192) who were young adults (M=22.2 yrs, SD=3.19 yrs, range 18-35 yrs). There were no strong differences in these characteristics between labs, see Table 3. During data acquisition, three subjects in Tromsø had to be excluded due to missing electrode contact after the first half of the experiment (two subjects) and a technical malfunction of the electrode cables (one subject). In Amsterdam, two subjects had to be excluded, one because of an interruption of the experimental session and one that turned out not to fulfill the inclusion criteria after the session. No subjects were excluded in Göttingen.

Table 3: Demographics across the three labs.

Lab	Proportion male	Mean/SD Age	Min/Max Age
AMS	10/64	20.66 (2.35)	[18, 31]
GOE	28/64	23.30 (2.66)	[18, 34]
TRM	20/64	22.75 (3.77)	[19, 35]
all	58/192	22.2 (3.19)	[18, 35]

3.2. Pre-Registered Analyses

In agreement with our sequential-sampling plan, we tested several times during data acquisition whether our stopping criterion was fulfilled. This criterion was that the 95% HDI of the posterior effect-size estimate would exclude zero in the positive direction. This did not turn out to be the case and therefore the maximum sample size was collected resulting in $N = 64$ subjects per lab and a total of 192 participants. In summary, the mean posterior effect size was consistently estimated to be slightly negative and the HDIs all included zero, see Table 4 and Figure 3.

3.2.1. Effect of anodal stimulation on self-reported mind wandering

With our final sample size, the effect-size estimated according to our pre-registered analysis plan was $d = -0.11$, HDI$= [-0.38, 0.17]$. Negative effect-sizes indicate that subjects in the anodal stimulation condition were less likely to respond off-task on the thought-probes than subjects in the sham stimulation condition. Accordingly, the directional Bayes Factor, $\mathrm{BF}_{\mathrm{directed}}$, which compared the hypotheses that the effect was positive to the hypothesis that it was zero or negative was in support of negative effect-sizes ($\mathrm{BF}_{\mathrm{directed}}=$ 0.29) but only slightly so. According to this test, it is about 3.4 times as likely that the effect-size was zero or negative when compared to a strictly positive effect. We also pre-specified several BFs that would test the null-hypothesis of a zero effect against several alternatives (against a positive, $\mathrm{BF}_{\mathrm{null}+}$, a negative, $\mathrm{BF}_{\mathrm{null}-}$, or any effect, $\mathrm{BF}_{\mathrm{null}}$, respectively). All of these Bayes Factors were in support of the null-hypothesis with varying degrees of strength. When comparing the null-hypothesis to the a-priori hypothesized positive effect, the null-hypothesis was about 10.65 times more likely to be true, $\mathrm{BF}_{\mathrm{null}+}= 10.65$. When comparing the null-hypothesis to any non-zero effect-size, the null-hypothesis was less strongly supported, $\mathrm{BF}_{\mathrm{null}}= 4.79$ and even when comparing the null against a negative effect-size (that was unlikely a-priori but seems more plausible given the observed negative effect-size), the null was slightly favored, $\mathrm{BF}_{\mathrm{null}-}= 3.09$.

Finally, we also calculated the replication Bayes Factors, $\mathrm{BF}_{\mathrm{replication}}$, and the meta-analytic BF, $\mathrm{BF}_{\mathrm{meta}}$ (Verhagen & Wagenmakers, 2014). The replication BF tests the hypothesis that the observed data from our replication study is consistent with the originally reported effect-size against the alternative that it is not. We found strong support for the alternative ($\mathrm{BF}_{\mathrm{replication}}= 0.002$) indicating that it is about 500 times as likely that the effect was not consistent with the originally reported effect-size, i.e., that the

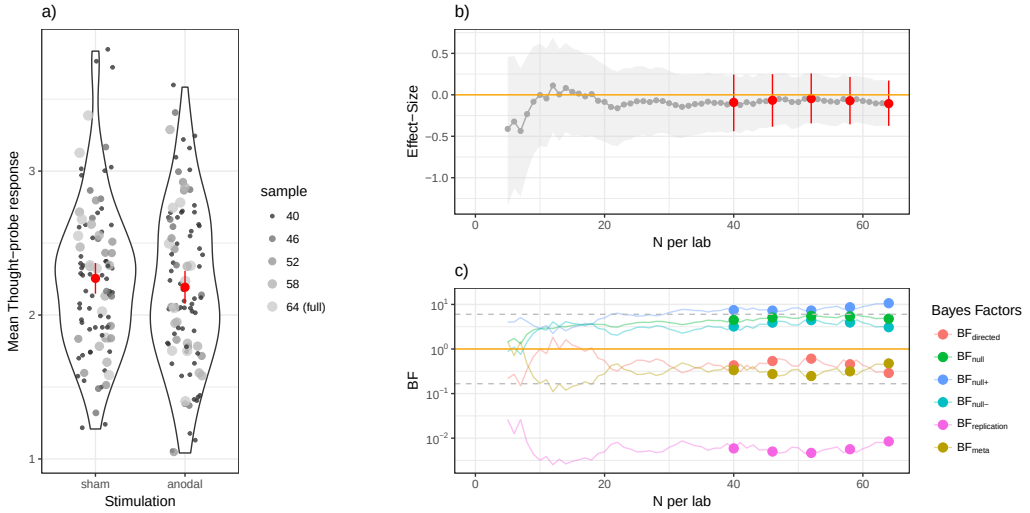

Figure 3: Results of the sequential sampling plan. Target statistics for increasing sample size (per lab) are plotted. Dots represent the pre-registered timepoints at which data-collection could have been stopped should the HDI have excluded zero in the positive direction. (a) Scatter-plot of individual subjects' mean thought-probe responses together with a density estimate and mean and confidence interval (red). (b) Effect-size and 95% HDI for the effect of anodal stimulation on mean thought probes. All HDI's included zero at all times. The final mean effect-size was in the opposite direction than hypothesized. (c) Bayes-factors quantifying evidence in support of various hypotheses (see text for details).

Table 4: Results at the pre-registered stopping points. The criterion for stopping the data-collection was that the 95% HDI around the effect-size would exclude zero in the positive direction. The effect-size was consistently negative and all HDIs included zero and therefore the complete sample was collected.

N	Cohen's d	BF_{null+}	BF_{null-}	BF_{null}	$BF_{directed}$	$BF_{replication}$	BF_{meta}
120	-0.09 [-0.44, 0.24]	7.46	3.21	4.48	0.43	0.002	0.34
138	-0.06 [-0.38, 0.25]	7.27	3.91	5.08	0.54	0.003	0.28
156	-0.05 [-0.35, 0.25]	7.30	4.44	5.52	0.61	0.003	0.25
174	-0.07 [-0.36, 0.22]	8.65	3.93	5.41	0.45	0.003	0.32
192	-0.11 [-0.38, 0.17]	10.65	3.09	4.79	0.29	0.002	0.48

effect did not replicate. The meta-analytic BF was calculated to judge overall support for the presence of any effect of anodal stimulation on thought-probes when pooling both the original and the replication study. Also this BF supported the null but only weakly so (BF_{meta}=0.48) which was expected given that the original study reported a huge, and most likely overestimated, effect-size ($d_{original} = 1.24$) which would bias the result of the meta-analytic BF in favor of a positive effect.

3.2.2. Hierarchical ordered probit model

The pre-registered hierarchical ordered probit model was fit to the final dataset. The posterior mean and HDIs are reported in Table 5. We ran 12 parallel chains for 2000 iterations each, treating the first 1000 samples as warmup resulting in final of 12000 independent samples from the posterior distribution. We used that many samples in order to properly estimate the tails of the distribution which were needed for accurately reporting the 95% HDI. The Gelman-Rubin diagnostic (Gelman & Rubin, 1992) was calculated to ensure that all reported results had an $\hat{R} \leq 1.05$. We also visually inspected the traceplots for all variables and no anomalies were spotted.

In order to show the appropriateness of the model, we conducted posterior predictive checks (Gelman et al., 1996). We generated $n_{rep} = 100$ complete datasets by drawing coefficients randomly from the posterior distribution and simulating datasets according to the model specification. The distribution of summary statistics from these posterior simulations can be compared to the actually observed data to evaluate model fit. Figure 4 shows the result of these checks. Model fit is excellent on the group-level but not all individual differences are picked up by this model.

Table 5: Results of fitting the hierarchical ordered probit model. As expected, there is a positive effect of trial number (time on task). However, contrary to our hypothesis, the coefficient coding for the effect of anodal stimulation is negative (with the HDI including zero).

Variable	Coefficient (Mean and 95% HDI)
Intercept (μ_g)	2.25 [2.14, 2.35]
trial (β_1)	0.20 [0.18, 0.23]
stimulation (β_{anodal})	-0.09 [-0.24, 0.07]
threshold (θ_2)	2.53 [2.51, 2.56]
probe-level variance (σ)	0.78 [0.76, 0.80]
group-level variance (σ_g)	0.62 [0.57, 0.68]

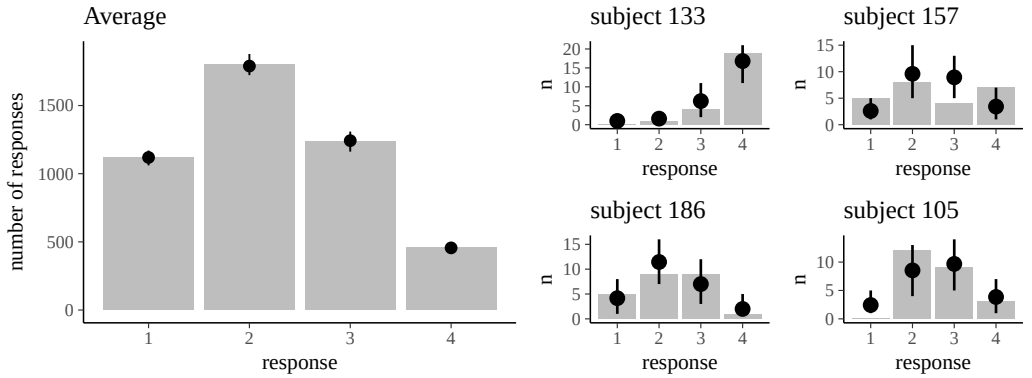

Figure 4: Posterior predictive distribution of average responses to thought-probes (left) and for four randomly selected subjects (right). Grey bars represent data, black dots and error bars represent mean and 95% HDI for simulated data.

The results of this analysis show a clear positive effect of time-on-task as previously reported, $\beta_1 = 0.20$ $[0.18, 0.23]$, indicating that subjects were more likely to report being off-task later in the experiment (about 0.67 units on the 4-point Likert-scale comparing the end to the beginning of the experiment). The results also show that anodal stimulation did not appear to increase the likelihood to answer off-task on the thought-probes, $\beta_{\text{anodal}} = -0.09$ $[-0.24, 0.07]$. While the mean coefficient estimate is negative, its 95% HDI includes zero and therefore does not provide evidence against the null-hypothesis.

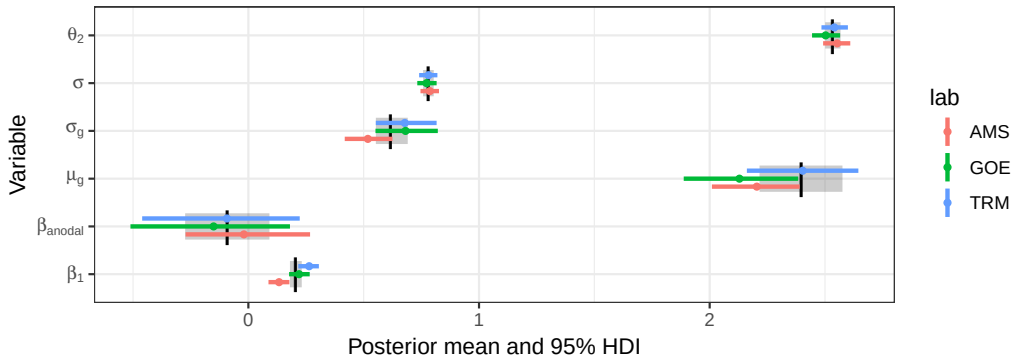

Figure 5: Coefficient estimates independently for each lab and from a combined model. Colored lines are estimates from individual lab data and the black line and grey area corresponds to posterior mean and 95% HDI from the combined model.

3.2.3. Effect of location (lab)

In order to test whether the lab in which each of the three subsets of data were collected would have an impact on the estimation of the effects, we pre-registered to fit the model from the previous section separately to the data from the three locations. In addition, we estimated a pre-registered extended model where lab was entered as a covariate (see Appendix for details). The same model-fitting and -checking procedure as detailed above was used to ensure that the model-fits were reliable.

Results for these analyses are presented in Figure 5. The estimates of the relevant coefficients are in good agreement between labs: Coefficients are estimated to be of a similar magnitude and the HDIs of the separately estimated coefficients overlap in almost all cases. The combined model, treating lab as a fixed-effect covariate seems to provide a good compromise between the independent estimates. The only exception is the coefficient for the time-on-task effect, β_1. The HDIs estimated for the Amsterdam sample $\beta_1 = 0.13$ $[0.088, 0.18]$ does not overlap with those from the Tromsø $\beta_1 = 0.26$ $[0.22, 0.31]$ or the Göttingen $\beta_1 = 0.22$, $[0.18, 0.27]$ samples. This finding indicates that participants in the AMS lab showed a lesser time-on-task effect on thought-probes than those in GOE or TRM.

We hesitate to provide an interpretation of this finding as it is quite possibly a spurious result: Analyzing the result from Figure 5 involves 18 comparisons. Therefore using 95% HDIs and decision by non-overlap of these intervals, we would already expect to see 1 or 2 positive results due to chance

alone (given that the models were fit on independent datasets).

We also pre-registered a model-comparison between the ordinal probit-regression model with and without the lab-covariate based on the LOOIC and the WAIC. This analysis can provide evidence for or against the suitability of including lab as a covariate in the model, i.e., whether a considerable amount of the variation in the data is being explained by this factor or not. The model that does not have any information about which lab the data was collected in resulted in a LOOIC of 10093.2 (SE= 83.1) and a WAIC of 10091.8 (SE= 83.0) while the extended model had a LOOIC of 10092.7 (SE= 83.1) and a WAIC of 10091.6 (SE= 83.0). These are virtually identical (ΔLOOIC= -0.3, SE= 0.8; ΔWAIC= -0.1, SE= 0.8) and therefore these criteria do not prefer any of the two models.

Even though the extended model did not provide a better model fit, we can check the regression coefficients corresponding to the different labs. Analyzing the extended model further, these coefficients were estimated as $\beta_{\text{AMS}} = -0.17, [-0.35, 0.02]$ and $\beta_{\text{GOE}} = -0.29, [-0.47, -0.10]$. According to this model, participants at the university of Göttingen were therefore less likely to respond to be off-task when compared to participants in Tromsø. As before when investigating the data from the labs separately, participants from Amsterdam were slightly less likely to respond with off-task than participants from Tromsø but slightly more likely to response off-task than subjects from Göttingen (though these HDIs did overlap).

We did not expect a priori to find any differences between the estimates from the three different labs. Since there were some indications of possible differences in the data, we chose to run several exploratory analyses to investigate possible reasons for this finding (see section 3.3.2).

3.2.4. Frequentist analyses

In accordance with our pre-registered analysis plan, we performed independent t-tests on individually calculated mean thought-probe scores. Note that only the initial sample of $N = 120$ is used in these tests as the stopping rule would invalidate p-values calculated for the complete sample since these would have to be corrected for the intermediate looks at the data. The two-tailed t-test exploring whether anodal tDCS resulted in altered (i.e., either increased or decreased) mind-wandering propensity relative to sham stimulation was not significant (t(117.68) = -1.01, p = 0.312, Cohen's d = -0.102). Also, the one-tailed t-test assessing directional effects indicated that anodal tDCS was not associated with increased propensity of mind wandering

(t(117.68) = -1.01, p = 0.843).

3.3. Exploratory Analyses

3.3.1. Influence of brain stimulation on other task measures

Table 6: Summary statistics of different outcome variables split by stimulation and online (part 1) and offline (part 2). Mean ± standard deviations are reported.

	1^{st} part Anodal	1^{st} part Sham	2^{nd} part Anodal	2^{nd} part Sham
Thought-probes	2.08 ± 0.56	2.15 ± 0.49	2.30 ± 0.62	2.36 ± 0.63
RT (ms)	393.4 ± 71.6	381.5 ± 61.8	380.6 ± 87.2	368.5 ± 55.6
RTCV	0.29 ± 0.13	0.28 ± 0.08	0.30 ± 0.12	0.29 ± 0.11
Commission errors (%)	35.7 ± 19.8	38.4 ± 18.8	43.1 ± 23.6	42.9 ± 20.6

In accordance with the well-known time-on-task effect on mind wandering (i.e., more attentional lapses in later parts of the task) that we already reported in our pre-registered analyses, we found compelling evidence for the effect of time (BF_{10}= 7.03 × 10^8; F(1,190) = 52.421; p < 0.001; η^2 = 0.216), although this effect was numerically rather small (first part: M = 2.12; SD = 0.52; second part: M = 2.33; SD = 0.62). In addition, participants became faster (BF_{10}= 106.46; GoRT: F(1,190) = 14.714; p < 0.001; η^2 = 0.072) and made more key presses on Nogo trials (commission errors: BF_{10}= 1958.5; F(1,190) = 21.409; p < 0.001; η^2 = 0.101) in the second part of the experiment. This finding indicates a change in the speed-accuracy tradeoff with task progress (Pearson's correlation between GoRT and commission errors for the whole task: BF_{10}= 4.07; r(190) = -0.199; p = 0.006), and might be related to more mind wandering during the second part of the task (Kendall's correlation between thought-probe ratings and GoRT for the whole task: BF_{10}= 3.55; τ(190) = 0.131; p = 0.008; between thought-probe ratings and commission errors: BF_{10}= 554.09; τ(190) = 0.203; p < 0.001). Finally, response times were more variable in the second part of the SART (RTCV: BF_{10}= 5.83; F(1,190) = 8.352; p = 0.004; η^2 = 0.042), an effect that can also be attributed to increasing mind wandering propensity with time spent on the task (Kendall's correlation between thought-probe ratings and RTCV: BF_{10}= 3639.73; τ(190) = 0.224; p < 0.001; Pearson's correlation between GoRT and RTCV: BF_{10}= 1411.99; r(190) = 0.312; p < 0.001; between commission errors and RTCV: BF_{10}= 1.08 × 10^8; r(190) = 0.446; p

< 0.001). Although omission errors on Go trials were not affected by time-on-task ($BF_{10}= 0.11$), they correlated positively both with mind wandering ($BF_{10}= 10.99$; $\tau(190) = 0.150$; $p = 0.004$) and with other task measures (GoRT: $BF_{10}= 101.1$; $r(190) = 0.268$; $p < 0.001$; RTCV: $BF_{10}= 5.42 \times 10^{27}$; $r(190) = 0.711$; $p < 0.001$).

With respect to the effect of tDCS on mind wandering or task performance, neither the main effect of stimulation (BF_{10} between 0.23 and 0.53; $F < 1.59$, $p > 0.208$), nor its interaction with time ($BF_{\text{inclusion}}$ between 0.15 and 0.28; $F < 1.241$, $p > 0.265$) were significant for either of the five measures of interest.

3.3.2. Exploratory analysis of location effects

In order to further investigate the effects of lab in which each of the three datasets was collected on thought-probe responses reported earlier, we extended the hierarchical probit regression model described in Appendix 1 by introducing interaction effects for lab $\times$ stimulation and lab $\times$ trial treating Tromsø as the baseline. The resulting model produced a better fit in terms of model-selection criteria (LOOIC= 10077.2, SE= 83.4) than the model with only lab as a main effect (ΔLOOIC= 7.3, SE= 4.3). Using this model, the HDIs for the main effect of lab no longer exclude zero, $\beta_{\text{AMS}} = -0.19, [-0.45, 0.07]$, $\beta_{\text{GOE}} = -0.24, [-0.50, 0.02]$ even though they are still indicating reduced off-task reports in both Amsterdam and Göttingen when compared to Tromsø. There is no evidence that the brain stimulation affected the thought-probe reports differentially in the three labs, $\beta_{\text{GOE}\times\text{stimulation}} = -0.09, [-0.45, 0.27]$, $\beta_{\text{AMS}\times\text{stimulation}} = -0.06, [-0.29, 0.42]$. Finally, the time-on-task effect seems to be reduced in subjects from Amsterdam as compared to Tromsø, $\beta_{\text{AMS}\times\text{trial}} = -0.13, [-0.18, -0.08]$ but not in Göttingen, $\beta_{\text{GOE}\times\text{trial}} = -0.04, [-0.09, 0.01]$. This finding agrees with the results from the pre-registered analysis which found that the time-on-task effect was reduced in Amsterdam in independent analyses for each lab.

Furthermore, we were interested in whether the apparent effect of lab might not actually be due to a gender effect. Previous research has reported gender differences in mind-wandering propensity (Bertossi et al., 2017) and given that we sampled a slightly higher proportion of females in Amsterdam than in the other labs (see Table 3), the observed lab-effect might actually be due to differences in mind-wandering in males and females. We investigated this possibility by augmenting the probit-regression model that includes lab as covariate with an additional covariate coding for the gender of the partic-

ipant. Assuming that any differences between the labs were due to gender effects, we would therefore expect the lab-coefficients to be estimated near zero and the coefficient coding for gender to show an effect. This augmentation of the model did not improve the model-fit (LOOIC= 10091.8, SE= 83.1; ΔLOOIC= -0.4, SE= 0.2). The coefficients for the lab-variables were similar to the ones estimated from the model not including gender as a covariate, $\beta_{\text{AMS}} = -0.16, [-0.35, 0.01]$ and $\beta_{\text{GOE}} = -0.27, [-0.45, -0.08]$ and the coefficient for gender was spread wide around zero, $\beta_{\text{male}} = -0.06, [-0.22, 0.11]$ indicating that gender was not likely to be responsible for the aforementioned lab effect.

3.3.3. Questionnaires

When analyzing changes in self-reported mood states during the task, both Bayesian and frequentist repeated-measures ANOVA revealed a main effect of time for positive, but not negative mood scores (PANAS-positive: BF_{10}= 8.37×10^{14}; F(1,190) = 92.480; p < 0.001; η^2 = 0.327; PANAS-negative: BF_{10}= 0.32; F(1,190) = 2.236; p = 0.136; η^2 = 0.012), indicating a significant reduction in positive mood by the end of the task (pre-task rating: M = 29.35; SD = 6.26; post-task rating: M = 25.09; SD = 7.22). Neither the main effect of stimulation nor its interaction with time was significant for the PANAS scores. Furthermore, since mind wandering has been associated with negative mood states (Killingsworth & Gilbert, 2010; Smallwood et al., 2009), we hypothesized a correlation between mind-wandering propensity (subjective thought-probe reports) and changes in mood scores measured by the PANAS. Despite our expectations, thought-probe responses did not correlate with pre- vs. post-SART difference scores for PANAS-negative (anodal tDCS group: BF_{10}= 0.36; $\tau(94)$ = 0.099; p = 0.179; sham tDCS group: BF_{10}= 0.13; $\tau(94)$ = 0.009; p = 0.908) or PANAS-positive items (anodal tDCS group: BF_{10}= 0.36; $\tau(94)$ = 0.98; p = 0.052; sham tDCS group: BF_{10}= 0.15; $\tau(94)$ = 0.035; p = 0.622).

Using the MAAS questionnaire, we have also collected self-reported scores on the individual's inherent ability to attend to the present experience and remain undistracted. Higher MAAS scores indicate higher level of concentration, and therefore, we anticipated that MAAS scores would negatively correlate with thought-probe scores. However, in contrast to our hypothesis, neither group showed a relationship between MAAS scores and mind wandering, albeit the correlations were in the expected direction (anodal tDCS group: BF_{10}= 0.36; $\tau(94)$ = -0.098; p = 0.166; sham tDCS group: BF_{10}=

0.29; $\tau(94)$ = -0.088; p = 0.214).

4. Discussion

The aim of the study was to replicate the findings reported by Axelrod et al. (2015) about the potential effect of anodal tDCS on mind-wandering propensity. Mind-wandering propensity was assessed by self-reports (thought-probes) while participants were engaged in a sustained attention task. Building upon the findings of the original publication, we tested the hypothesis that anodal tDCS over the left DLPFC would increase mind-wandering propensity relative to an inactive (sham) stimulation. The present replication study was performed as a fully pre-registered, multi-center study utilizing a sequential sampling plan with equal sample size across laboratories.

Contrary to our hypothesis and the findings from Axelrod et al. (2015), we found that the participants receiving anodal stimulation were numerically less likely to respond being off-task when compared to the group receiving sham stimulation over the left DLPFC. Overall, however, our findings show support in favor of a null-effect of stimulation on self-reported thought-probe scores as shown by an analysis based on Bayes Factors. When comparing a null-effect to an effect in the positive direction as hypothesized *a priori*, there was strong evidence for a null effect (BF_{null+}= 10.65). Also, when testing the hypothesis of the effect being zero against the full range of possible non-zero effects, there was moderate evidence for a null effect (BF_{null}= 4.79) and even when comparing against a purely negative effect, the null was somewhat favored (BF_{null-}= 3.09). In addition, there was extreme evidence ($BF_{replication}$= 0.002) that the original study was not replicated using a special Bayes Factor designed to indicate replication success (Verhagen & Wagenmakers, 2014). When pooling data from both the original and replication study there was strong evidence (BF_{meta}= 0.059) for the absence of an effect of anodal stimulation. We conclude from these results that there is no support for the supposition that bipolar anodal tDCS in the form used in our and the original study (Axelrod et al., 2015) can influence the propensity to mind-wander. On the contrary, we found substantive evidence *against* the existence of such an effect.

Our failure to replicate the original study is perhaps not particular surprising when viewed in the context of previous replication failures in the field of psychology (e.g.; Open Science Collaboration et al., 2015; Klein et al.,

2014; Wagenmakers et al., 2016) in general and brain stimulation in particular (Learmonth et al., 2017; Horvath et al., 2016; Vannorsdall et al., 2016). Typically, a result obtained in an initial, often low-powered study fails to be reproduced in large-sample replication attempts (Boekel et al., 2015). Replications are the cornerstone of empirical research and crucial for scientific progress. Even though this is a well-known fact, replication attempts are still rare (Makel et al., 2012). Several reasons for this problematic state of affairs have been pointed out by many authors (Simmons et al., 2011; Chambers, 2017) which comprise factors on many different levels. We conclude that the original result by Axelrod et al. (2015) was most likely a false positive finding caused by strong variability and low sample size. We believe that it is crucial that future studies aiming to establish a specific experimental effect should be required to (a) employ sample sizes that are adequate to find effects of a reasonable magnitude, and (b) to either pre-register their study from the outset or provide a pre-registered replication of their own result. Such requirements would go a long way to protect the literature from the omnipresent false-positives, even though replication by independent, if possible multiple, labs is the ultimate goal (Simons, 2014).

It is important to point out, however, that our failed replication of the study by Axelrod et al. (2015) does not imply that tDCS is an ineffective tool for modulating mind wandering propensity. On the contrary, we are aware of four other studies that reported evidence for active stimulation either increasing or reducing the mind-wandering propensity during various tasks. In three studies, Kajimura and colleagues showed that anodal stimulation of the right inferior parietal lobule (rIPL) reduces mind wandering propensity (Kajimura & Nomura, 2015; Kajimura et al., 2016, 2018). In their first two reports (Kajimura & Nomura, 2015; Kajimura et al., 2016), the cathode was placed above the left DLPFC, rendering the contribution of left DLPFC vs. rIPL to the observed effect impossible to distinguish. However, in their most recent study, the authors used an extracephalic return electrode, providing evidence for rIPL stimulation being primarily responsible for the mind wandering-reducing effect (Kajimura et al., 2018). Interestingly, analysis of effective connectivity patterns revealed that the behavioral effect of anodal tDCS on decreased mind wandering propensity was mediated by weaker afferent connections from the medial prefrontal cortex (MPFC) to the posterior cingulate cortex, highlighting the MPFC node within the DMN as a key mediator for inducing and/or maintaining task-unrelated thoughts (Kajimura et al., 2016). The role of the MPFC in influencing mind wandering is also

supported by another study showing that cathodal tDCS targeting the left MPFC reduces attentional lapses during a choice reaction time task in males (Bertossi et al., 2017). Given the negative results of the current study, however, it is important to replicate any of these positive effects before accepting them as facts.

As detailed in the introduction, several neuroimaging studies and theoretical accounts attribute an important role to the FPN (and, more specifically, to the DLPFC) in regulating mind wandering episodes under various circumstances (Christoff et al., 2009, 2016; Dumontheil et al., 2010; Smallwood et al., 2012). In this regard, the positive finding by Axelrod et al. (2015) fits well in this framework, seemingly providing direct evidence for the causal (rather than correlational) involvement of the left DLPFC to regulating mind wandering propensity. However, the poor spatial focality of bipolar tDCS montages is well-known (Csifcsák et al., 2018; Laakso et al., 2016; Opitz et al., 2015), with stimulation-induced electric fields (EFs) spreading well beyond the area of scalp electrodes, most probably influencing neural excitability in a wide range of cortical areas (Keeser et al., 2011). Using high-resolution realistic head models of healthy adults, we have recently demonstrated that tDCS protocols targeting the left DLPFC show substantial inter-individual variability in the spatial distribution of tDCS-induced EFs (Boayue et al., 2018). Using our previously described and publicly available pipeline (Boayue et al., 2018), we now present new modelling results to gain insight into the potential underlying neural effects that were induced by our tDCS protocol. We focused on the normal component of the EF, that is, on the component perpendicular to the cortical surface, either entering (positive values) or leaving the cortex (negative values). Previous work identified these currents as being excitatory or inhibitory in nature (Rahman et al., 2013), enabling us to assess the direction of the expected effect. In Figure 6 (left panel), we show that despite targeting the left DLPFC, this montage induces EFs in both the medial and lateral aspects of the two hemispheres. Moreover, the right and left MPFC receives excitatory and inhibitory stimulation, respectively, which is particularly interesting as both the enhancement and reduction of MPFC activity by tDCS was associated with changes in mind wandering propensity (Bertossi et al., 2017; Kajimura et al., 2016). Based on these, we argue that stimulation of the MPFC could just as well be responsible for the effect reported by Axelrod et al. (2015) than that of the left DLPFC. In addition, the variability maps shown in Figure 6 (right panel) clearly indicate that the magnitude of EFs in the bilateral DLPFC is highly variable between

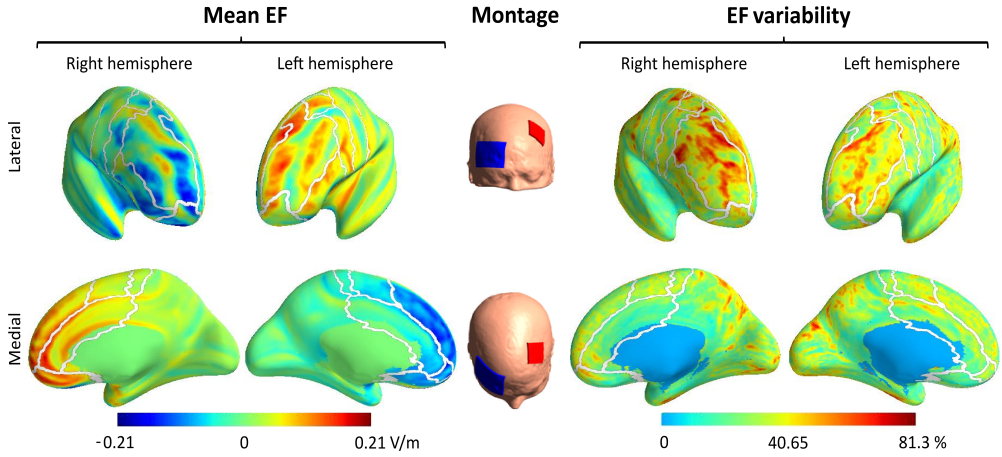

Figure 6: Simulation of transcranial direct current stimulation-induced electric fields (EFs) in the cortex of 18 head models for the montage used in our study and by Axelrod et al. (2015). Group-averaged mean values are presented on the left side, whereas the variability of effects across individuals is presented on the right side. For these simulations, we focused on the normal component of the EF, manifesting in positive (anode-like) and negative (cathode-like) values in the mean maps. Across-subject variability was quantified as the EF coefficient of variation ($\frac{standard\ deviation}{mean} \times 100$). Simulation parameters and methods were as described in Csifcsák et al. (2018).

participants.

The tDCS protocol employed in our and the original study even though standard in the field has some drawbacks: First, the protocol used a weak stimulation intensity (1 mA) resulting in electric-field magnitudes of about 0.1-0.2 V/m in the target area (see Figure 6). These estimates are based on computational models that have also been validated by intracranial measurements (Opitz et al., 2016). It is unclear whether the electric field induced by transcranial electric stimulation is robust and strong enough to cause any physiological effect (Huang et al., 2017), let alone manifest at the behavioural level. Therefore, it is possible that the stimulation intensity of 1 mA with the present bipolar montage is just not potent enough for the tDCS-induced electric field to have an effect on neural excitability (Vöröslakos et al., 2018). Second, the bipolar tDCS protocol produces diffuse electric fields resulting in a lack of specificity and the unintended stimulation of other regions (Csifcsák et al., 2018). The result is a diffuse stimulation of the target region. A better approach might be the use of recently developed high-definition brain stimulation protocols, e.g., 4 × 1 ring protocols, which allows for more targeted

stimulation (Datta et al., 2009). These protocols allow a much more precise targeting of a region of interest while minimizing the electric field in other parts of the brain. However, this increased focality comes at the price of possibly influencing different regions in different subjects because of substantial differences in brain anatomy (Opitz et al., 2015). It is therefore desirable to use individualized montages based on head models from high resolution magnetic resonance (MR) images to guide optimal electrode placement to result in comparable electric field distributions in individual brains. Taken together, routine usage of this approach could in the future help to increase focality of stimulation and to reduce between-subject variance of the results.

As part of our exploratory analysis, we found that anodal tDCS was not associated with either online or offline effects on task performance. Still, we found robust time-on-task effects regarding thought-probes, accuracy and reaction time measures, which are in line with previous findings (Bastian & Sackur, 2013; Cheyne et al., 2009; McVay & Kane, 2012; Smallwood & Schooler, 2006). Interestingly, although the negative correlation between response times and commission error rates are indicative of a speed-accuracy tradeoff, these parameters were inversely influenced by mind wandering propensity on a between-subject level: Participants reporting more mind wandering were characterized by higher error rates, but also by longer (rather than shorter) reaction times. Response time slowing has been associated with task-unrelated thoughts previously, and it was also found to be predictive of omission errors, as in our study (McVay & Kane, 2012; Smallwood & Schooler, 2006). Nevertheless, these data strengthen views that there is a complex relationship between self-reported mind wandering intensity and performance patterns on the SART (McVay & Kane, 2012), since the latter can be influenced by factors other than mind wandering per se (e.g., impulsivity or response strategy; Helton et al., 2010). Finally, it is worth mentioning that RT variability (RTCV) showed the strongest correlation with thought-probes, highlighting this measure as the most promising objectively quantifiable SART performance index for estimating the prevalence of off-task periods (Bastian & Sackur, 2013).

Rather surprisingly, we did not find a relationship between mind wandering propensity and the participants' mood scores. Despite the often described link between negative mood and task-unrelated thoughts (Killingsworth & Gilbert, 2010; Smallwood et al., 2009), the causal relationship between these phenomena might be too subtle to be detected by our relatively simple questionnaires and thought-probes. Moreover, to avoid inducing mood changes

prior to tDCS, we asked our participants to rate their pre-task mood retrospectively, which most probably restricted the reliability of our mood data. The individual's predisposition to mindfully attend to the present has been regarded as a personality attribute that is opposed to the propensity to mind wander (Mrazek et al., 2012). However, in our dataset, we did not observe a negative correlation between thought-probe responses and MAAS scores. Interestingly, recent work pointed out that rather than merely being in contrast, these phenomena can interact in a very complex, and at times synergistic way (Seli et al., 2015; Agnoli et al., 2018). For example, it was suggested that the deliberate vs. spontaneous nature of mind wandering is differently related to certain factors of mindfulness (Seli et al., 2015). Thus, the fact that our thought-probes were not enquiring about this aspect of mind wandering might have rendered our analysis insensitive to unveiling the relationship between these phenomena.

We also found indications for differences in mind-wandering propensity between the labs. Even thought the results were not very strong (0.2-0.3 units on the 4-point Likert scale) and did not increase the model fit in terms of the model-selection criteria, participants from the university of Amsterdam were generally less likely to respond off-task to the thought-probes than participants from Tromsø. This finding may have several possible explanations. For example, subtle differences in how the thought-probes are being expressed in the three languages (German, Dutch and Norwegian) may have caused participants to give slightly different interpretations to the meaning of the scale. This is a common issue when comparing scales across languages and it is often recommended to disregard any cross-language main effects, assuming that the scales still have metric equivalence but may have a shifted origin (van de Vijver, F. J. R. & Leung, K, 2011). Another possibility are national differences in acceptability of deviations from task-conform behaviour. Recently, researchers have begun to look more closey into boundary conditions of the thought-probe technique (Weinstein et al., 2018; Weinstein, 2017). This finding is a first indication that it may be important to consider language- or nationality-specific effects as well.

In summary, in a high-powered, pre-registered multi-center study, we were not only unable to detect an effect of anodal transcranial direct current stimulation on mind wandering propensity, but we actually found evidence for the absence of such an effect. Our findings further emphasize the significance of direct replications for the further advancement of the field of cognitive neuroscience in general and brain-stimulation in particular.

5. Appendix

5.1. Hierarchical ordered probit model

The model is fully specified as follows: Each response to a thought-probe (one of the set {1,...,K}) given by subject j in trial t, is modeled as a categorical variable with probability K-simplex p (a K-simplex is a set of K positive numbers that sum to one)

$$\text{probe}_{j,t} \sim \text{Categorical}(p).$$

The probabilities for each of the responses are calculated by assuming an underlying, continuous, normally-distributed "mind-wandering" variable y with parameters $\mu_{j,t}$ and σ that is thresholded into the discrete responses at thresholds $\theta_1, \ldots, \theta_{K-1}$. The probabilites to give each of the responses is the area under the normal curve of y that falls into the K response-bins $[-\infty, \theta_1], \ldots, [\theta_{K-1}, \infty]$. Therefore, the probabilites are calculated as

$$p_k = \Phi\left(\frac{\theta_k - \mu_{j,t}}{\sigma}\right) - \Phi\left(\frac{\theta_{k-1} - \mu_{j,t}}{\sigma}\right)$$

where Φ is the cumulative standard normal distribution (see Kruschke, 2014, for a comprehensive presentation of this model).

The underlying distribution is modeled with a hierarchical linear model

$$\mu_{j,t} = \beta_{0,j} + \beta_1 z(t) + \beta_{\text{anodal}}\text{anodal}_j \tag{1}$$

where $z(t)$ is the z-transformed trial number and anodal_j is an indicator variable specifying whether a subject was in the control group (0) or in the anodal stimulation group (1). The subject-level intercepts are constrained by a group-level distribution

$$\beta_{0,j} \sim \text{Normal}(\mu_g, \sigma_g).$$

Priors are set to be vague as recommended in Kruschke (2014):

$$\mu_g \sim \text{Normal}\left(\frac{1+K}{2}, K\right),$$

$$\sigma_g \sim \text{Uniform}\left(K/1000, 10K\right),$$

$$\sigma \sim \text{Uniform}\left(K/1000, 10K\right)$$

and

$$\beta_1 \sim \text{Normal}(0, K).$$

The test of the hypothesis that anodal stimulation can increase mind-wandering is whether the distribution for the β_{anodal} coefficient will be larger than zero.

For analyzing the effect of lab where the data for a specific subject was collected, we run three instances of this model with the datasets from the three universities and present the resulting posterior distribution side-by-side. In addition, we augment this model with a covariate for lab, modifying Eq. 1 to read

$$\mu_{j,t} = \beta_{0,j} + \beta_1 z(t) + \beta_{\text{anodal}}\text{anodal}_j + \beta_{\text{labAMS}}\text{AMS}_j + \beta_{\text{labGOE}}\text{GOE}_j$$

where AMS and GOE are indicator variables coding for whether a subject was recorded in Amsterdam or Göttingen, respectively (with Tromsø serving as the baseline). This augmented model will be compared to the model without these covariates using the LOO and WAIC indicators to evaluate whether the inclusion of this information would improve the fit of the model.

5.2. Changes to the original protocol

The changes detailed here are part of our OSF protocol and can also be found under https://osf.io/37kfj/.

5.2.1. Changes made after pre-registering with EJN but before any data was collected

The changes documented here have been made before the first dataset was collected. It is part of a registration at OSF that has been made on November, 2nd 2017, https://osf.io/bv32d/.

Additional instructions for experimenter.

- added three more questions (the last three) to the Q&A sheet with standardized answers to questions that the data-collectors from the three labs are using in case there are questions from the participants; those were added purely for preventive reasons because of experiences during piloting

Adapted translated instructions.

- adapted the German instructions to reflect the English template; this was because of an oversight in which only the English template was adjusted during preparation of the study while the translations were forgotten. This oversight was spotted by our German collaborators and we fixed this before any data-collection

Expanded instructions to avoid accidental unblinding.

- during the course of the pilots at our partnering institutions, we became aware of the fact that our previously detailed protocol could result in accidental unblinding of the experimenter. This is due to the fact that the impedance measurement on the stimulator reflects the ramp-down period which is earlier in the sham as compared to the real stimulation condition. We account for this by requiring the experimenters to cover the stimulation device after recording the initial impedance measurement and to turn it off without lifting the cover before turning it on again for the final post-stimulation measurement of impedance. This is reflected in updated portions of the experimenter instructions.

- we added a note to the datasheet where the experimenter should input the number of times the impedance measurement had to be repeated to come below the required 10 kOhm

Screen size. We became aware of an error in our pre-registration where we specified that we would be using 12" flatscreen monitors. The actual screen size in the three labs was 19". This difference in screen sizes had no impact on the size of the displayed stimuli as those were adjusted to cover 3 degrees of visual angle independently for each lab.

5.2.2. Changes made after starting the data collection but before any analysis was conducted

None.

5.2.3. Changes made after finished data-collection

It was necessary to adapt several of the pre-registered analysis scripts. There were two reasons for these changes:

1. There were updates to some of the used analyses packages which required changes to the code in order to run as intended

2. There were errors in the original analysis-script that were only spotted when confronted with real data.

At our OSF-repository https://osf.io/dct2r/, we store a copy of the updated analysis files and we also keep the output of the `diff` utility that stores any changes made to the original scripts in an easily readable format. These files are called <`scriptname`>`.diff` where <`scriptname`> is replaced with each of the changed script files. The original script files can be retrieved from the pre-registration at https://osf.io/bv32d/.

6. References

Agnoli, S., Vanucci, M., Pelagatti, C. & Corazza, G. E. (2018) Exploring the link between mind wandering, mindfulness, and creativity: A multidimensional approach. *Creativity Research Journal*, **30**, 1, 41–53.

Andrews-Hanna, J. R. (2012) The brain's default network and its adaptive role in internal mentation. *The Neuroscientist*, **18**, 3, 251–270.

Andrews-Hanna, J. R., Reidler, J. S., Sepulcre, J., Poulin, R. & Buckner, R. L. (2010) Functional-anatomic fractionation of the brain's default network. *Neuron*, **65**, 4, 550–562.

Axelrod, V., Rees, G., Lavidor, M. & Bar, M. (2015) Increasing propensity to mind-wander with transcranial direct current stimulation. *Proc. Natl. Acad. Sci. U. S. A.*, p. 201421435.

Baars, B. J., Franklin, S. & Ramsoy, T. Z. (2013) Global workspace dynamics: cortical "binding and propagation" enables conscious contents. *Front. Psychol.*, **4**, 200, 10–3389.

Bastian, M. & Sackur, J. (2013) Mind wandering at the fingertips: automatic parsing of subjective states based on response time variability. *Frontiers in psychology*, **4**, 573.

Berlim, M. T., Van den Eynde, F. & Daskalakis, Z. J. (2013) Clinical utility of transcranial direct current stimulation (tdcs) for treating major depression: a systematic review and meta-analysis of randomized, double-blind and sham-controlled trials. *J. Psychiatr. Res.*, **47**, 1, 1–7.

Bertossi, E., Peccenini, L., Solmi, A., Avenanti, A. & Ciaramelli, E. (2017) Transcranial direct current stimulation of the medial prefrontal cortex dampens mind-wandering in men. *Scientific reports*, **7**, 1, 16962.

Boayue, N. M., Csifcsák, G., Puonti, O., Thielscher, A. & Mittner, M. (2018) Head models of healthy and depressed adults for simulating the effects of non-invasive brain stimulation [version 1; referees: 1 approved]. *F1000Research*, **7**.

Boekel, W., Wagenmakers, E.-J., Belay, L., Verhagen, J., Brown, S. & Forstmann, B. U. (2015) A purely confirmatory replication study of structural brain-behavior correlations. *Cortex*, **66**, 115–133.

Broadway, J. M., Zedelius, C. M., Mooneyham, B. W., Mrazek, M. D. & Schooler, J. W. (2015) Stimulating minds to wander. *Proc. Natl. Acad. Sci. U. S. A.*, **112**, 11, 3182–3183.

Brown, K. W. & Ryan, R. M. (2003) The benefits of being present: mindfulness and its role in psychological well-being. *J. Pers. Soc. Psychol.*, **84**, 4, 822.

Brunoni, A. R., Ferrucci, R., Fregni, F., Boggio, P. S. & Priori, A. (2012) Transcranial direct current stimulation for the treatment of major depressive disorder: a summary of preclinical, clinical and translational findings. *Prog. Neuropsychopharmacol. Biol. Psychiatry*, **39**, 1, 9–16.

Buckner, R. L., Andrews-Hanna, J. R. & Schacter, D. L. (2008) The brain's default network. *Ann. N.Y. Acad. Sci.*, **1124**, 1, 1–38.

Bürkner, P.-C. et al. (2017) brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software*, **80**, 1, 1–28.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P. & Riddell, A. (2017) Stan: A probabilistic programming language. *J. Stat. Softw.*, **76**, 1, 1–32.

Chambers, C. (2017) *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice*. Princeton University Press.

Chambers, C. D., Feredoes, E., Muthukumaraswamy, S. D. & Etchells, P. (2014) Instead of" playing the game" it is time to change the rules: Registered reports at aims neuroscience and beyond. *AIMS Neuroscience*, **1**, 1, 4–17.

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.

Christoff, K., Gordon, A. M., Smallwood, J., Smith, R. & Schooler, J. W. (2009) Experience sampling during fmri reveals default network and executive system contributions to mind wandering. *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 21, 8719–8724.

Christoff, K., Irving, Z. C., Fox, K. C., Spreng, R. N. & Andrews-Hanna, J. R. (2016) Mind-wandering as spontaneous thought: a dynamic framework. *Nature Reviews Neuroscience*, **17**, 11, 718.

Csifcsák, G., Boayue, N. M., Puonti, O., Thielscher, A. & Mittner, M. (2018) Effects of transcranial direct current stimulation for treating depression: A modeling study. *Journal of affective disorders*, **234**, 164–173.

Datta, A., Bansal, V., Diaz, J., Patel, J., Reato, D. & Bikson, M. (2009) Gyri-precise head model of transcranial direct current stimulation: improved spatial focality using a ring electrode versus conventional rectangular pad. *Brain stimulation*, **2**, 4, 201–207.

Dehaene, S. & Changeux, J.-P. (2011) Experimental and theoretical approaches to conscious processing. *Neuron*, **70**, 2, 200–227.

Dehaene, S., Changeux, J.-P., Naccache, L., Sackur, J. & Sergent, C. (2006) Conscious, preconscious, and subliminal processing: a testable taxonomy. *Trends Cognit. Sci.*, **10**, 5, 204–211.

Dumontheil, I., Gilbert, S. J., Frith, C. D. & Burgess, P. W. (2010) Recruitment of lateral rostral prefrontal cortex in spontaneous and task-related thoughts. *Quarterly Journal of Experimental Psychology*, **63**, 9, 1740–1756.

Engelen, U., De Peuter, S., Victoir, A., Van Diest, I. & Van den Bergh, O. (2006) Verdere validering van de positive and negative affect schedule (panas) en vergelijking van twee nederlandstalige versies. *gedrag en gezondheid*, **34**, 2, 61–70.

Feng, S., D'Mello, S. & Graesser, A. C. (2013) Mind wandering while reading easy and difficult texts. *Psychon. Bull. & Rev.*, **20**, 3, 586–592.

Fox, K. C. & Christoff, K. (2015) Transcranial direct current stimulation to lateral prefrontal cortex could increase meta-awareness of mind wandering. *Proc. Natl. Acad. Sci. U. S. A.*, p. 201504686.

Gelman, A. & Carlin, J. (2014) Beyond power calculations assessing type s (sign) and type m (magnitude) errors. *Perspect. Psychol. Sci.*, **9**, 6, 641–651.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A. & Rubin, D. (2013) *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis.

Gelman, A., Hwang, J. & Vehtari, A. (2014) Understanding predictive information criteria for bayesian models. *Stat. Comput.*, **24**, 6, 997–1016.

Gelman, A., Meng, X.-L. & Stern, H. (1996) Posterior predictive assessment of model fitness via realized discrepancies. *Statistica sinica*, pp. 733–760.

Gelman, A. & Rubin, D. B. (1992) Inference from iterative simulation using multiple sequences. *Statistical science*, pp. 457–472.

Gullhaugen, A. S. & Nøttestad, J. A. (2012) Under the surface the dynamic interpersonal and affective world of psychopathic high-security and detention prisoners. *Int. J. Offender Ther. Comp. Criminol.*, **56**, 6, 917–936.

He, J., Becic, E., Lee, Y.-C. & McCarley, J. S. (2011) Mind wandering behind the wheel: performance and oculomotor correlates. *Hum. Factors*, **53**, 1, 13–21.

Helton, W. S., Weil, L., Middlemiss, A. & Sawers, A. (2010) Global interference and spatial uncertainty in the sustained attention to response task (sart). *Consciousness and Cognition*, **19**, 1, 77–85.

Horvath, J. C., Carter, O. & Forte, J. D. (2016) No significant effect of transcranial direct current stimulation (tdcs) found on simple motor reaction time comparing 15 different simulation protocols. *Neuropsychologia*, **91**, 544–552.

Horvath, J. C., Forte, J. D. & Carter, O. (2015a) Evidence that transcranial direct current stimulation (tDCS) generates little-to-no reliable neurophysiologic effect beyond MEP amplitude modulation in healthy human subjects: A systematic review. *Neuropsychologia*, **66**, 213–236.

Horvath, J. C., Forte, J. D. & Carter, O. (2015b) Quantitative review finds no evidence of cognitive effects in healthy populations from single-session transcranial direct current stimulation (tdcs). *Brain Stimul.*

Huang, Y., Liu, A. A., Lafon, B., Friedman, D., Dayan, M., Wang, X., Bikson, M., Doyle, W. K., Devinsky, O. & Parra, L. C. (2017) Measurements and models of electric fields in the in vivo human brain during transcranial electric stimulation. *Elife*, **6**, e18834.

Janke, S. & Glöckner-Rist, A. (2014) Deutsche version der positive and negative affect schedule (panas). In *Zusammenstellung sozialwissenschaftlicher Items und Skalen.*, volume 10, p. 6102.

JASP Team (2018) JASP (Version 0.9)[Computer software]. URL https://jasp-stats.org/.

Kajimura, S., Kochiyama, T., Abe, N. & Nomura, M. (2018) Challenge to unity: Relationship between hemispheric asymmetry of the default mode network and mind wandering. *Cerebral Cortex*, p. bhy086. URL http://dx.doi.org/10.1093/cercor/bhy086.

Kajimura, S., Kochiyama, T., Nakai, R., Abe, N. & Nomura, M. (2016) Causal relationship between effective connectivity within the default mode network and mind-wandering regulation and facilitation. *Neuroimage*, **133**, 21–30.

Kajimura, S. & Nomura, M. (2015) Decreasing propensity to mind-wander with transcranial direct current stimulation. *Neuropsychologia*, **75**, 533–537.

Keeser, D., Meindl, T., Bor, J., Palm, U., Pogarell, O., Mulert, C., Brunelin, J., Möller, H.-J., Reiser, M. & Padberg, F. (2011) Prefrontal transcranial direct current stimulation changes connectivity of resting-state networks during fmri. *Journal of Neuroscience*, **31**, 43, 15284–15293.

Kelley, N. J., Hortensius, R. & Harmon-Jones, E. (2013) When anger leads to rumination induction of relative right frontal cortical activity with transcranial direct current stimulation increases anger-related rumination. *Psychol. Sci.*, **24**, 4, 475–481.

Killingsworth, M. A. & Gilbert, D. T. (2010) A wandering mind is an unhappy mind. *Science*, **330**, 6006, 932–932.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr., R. B., Bahnik, S., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., Hasselman, F., Hicks, J. A., Hovermale, J. F., Hunt, S. J., Huntsinger, J. R., IJzerman, H., John, M.-S., Joy-Gaba, J. A., Barry Kappes, H., Krueger, L. E., Kurtz, J., Levitan, C. A., Mallett, R. K., Morris, W. L., Nelson, A. J., Nier, J. A., Packard, G., Pilati, R., Rutchick, A. M., Schmidt, K., Skorinko, J. L., Smith, R., Steiner, T. G., Storbeck, J., Van Swol, L. M., Thompson, D., van 't Veer, A. E., Vaughn, L. A., Vranka, M., Wichman, A. L., Woodzicka, J. A. & Nosek, B. A. (2014) Investigating variation in replicability: A "many labs" replication project. *Social Psychology*, **45**, 3, 142–152.

Kruschke, J. (2014) *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.

Laakso, I., Tanaka, S., Mikkonen, M., Koyama, S., Sadato, N. & Hirata, A. (2016) Electric fields of motor and frontal tdcs in a standard brain space: a computer simulation study. *Neuroimage*, **137**, 140–151.

Learmonth, G., Felisatti, F., Siriwardena, N., Checketts, M., Benwell, C. S., Märker, G., Thut, G. & Harvey, M. (2017) No interaction between tdcs current strength and baseline performance: A conceptual replication. *Frontiers in neuroscience*, **11**, 664.

Makel, M. C., Plucker, J. A. & Hegarty, B. (2012) Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science*, **7**, 6, 537–542.

Mason, M. F., Norton, M. I., Van Horn, J. D., Wegner, D. M., Grafton, S. T. & Macrae, C. N. (2007) Wandering minds: the default network and stimulus-independent thought. *Science*, **315**, 5810, 393–395.

McVay, J. C. & Kane, M. J. (2012) Drifting from slow to "d'oh!": Working memory capacity and mind wandering predict extreme reaction times and executive control errors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **38**, 3, 525.

Michalak, J., Heidenreich, T., Ströhle, G. & Nachtigall, C. (2008) Die deutsche version der mindful attention and awareness scale (maas) psychometrische befunde zu einem achtsamkeitsfragebogen. *Zeitschrift für klinische Psychologie und Psychotherapie*, **37**, 3, 200–208.

Minarik, T., Berger, B., Althaus, L., Bader, V., Biebl, B., Brotzeller, F., Fusban, T., Hegemann, J., Jesteadt, L., Kalweit, L. et al. (2016) The importance of sample size for reproducibility of tdcs effects. *Front. Hum. Neurosci.*, **10**.

Mittner, M., Boekel, W., Tucker, A. M., Turner, B. M., Heathcote, A. & Forstmann, B. U. (2014) When the brain takes a break: a model-based analysis of mind wandering. *J. Neurosci.*, **34**, 49, 16286–16295.

Morey, R. D. & Lakens, D. (2016) Why most of psychology is statistically unfalsifiable. Preprint available at https://github.com/richarddmorey/psychology_resolution/blob/master/paper/response.pdf.

Morey, R. D. & Rouder, J. N. (2015) *BayesFactor: Computation of Bayes Factors for Common Designs*. URL http://CRAN.R-project.org/package=BayesFactor. R package version 0.9.12-2.

Mrazek, M. D., Smallwood, J. & Schooler, J. W. (2012) Mindfulness and mind-wandering: finding convergence through opposing constructs. *Emotion*, **12**, 3, 442.

Nosek, B. A. & Lakens, D. (2014) Registered reports: A method to increase the credibility of published results. *Soc. Psychol.*

Nuijten, M. B., van Assen, M. A., Veldkamp, C. L. & Wicherts, J. M. (2015) The replication paradox: Combining studies can decrease accuracy of effect size estimates. *Rev. Gen. Psychol.*, **19**, 2, 172.

Okon-Singer, H., Hendler, T., Pessoa, L. & Shackman, A. J. (2015) The neurobiology of emotion–cognition interactions: fundamental questions and strategies for future research. *Front. Hum. Neurosci.*, **9**.

Open Science Collaboration et al. (2015) Estimating the reproducibility of psychological science. *Science*, **349**, 6251, aac4716.

Opitz, A., Falchier, A., Yan, C.-G., Yeagle, E. M., Linn, G. S., Megevand, P., Thielscher, A., Milham, M. P., Mehta, A. D., Schroeder, C. E. et al. (2016) Spatiotemporal structure of intracranial electric fields induced by transcranial electric stimulation in humans and nonhuman primates. *Scientific reports*, **6**, 31236.

Opitz, A., Paulus, W., Will, S., Antunes, A. & Thielscher, A. (2015) Determinants of the electric field during transcranial direct current stimulation. *Neuroimage*, **109**, 140 – 150. URL http://www.sciencedirect.com/science/article/pii/S1053811915000488.

Peirce, J. W. (2007) Psychopy—psychophysics software in python. *J. Neurosci. Methods*, **162**, 1, 8–13.

R Core Team (2015) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rahman, A., Reato, D., Arlotti, M., Gasca, F., Datta, A., Parra, L. C. & Bikson, M. (2013) Cellular effects of acute direct current stimulation: somatic and synaptic terminal effects. *The Journal of physiology*, **591**, 10, 2563–2578.

Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A. & Shulman, G. L. (2001) A default mode of brain function. *Proc. Natl. Acad. Sci. U. S. A.*, **98**, 2, 676–682.

Rouder, J. N. & Morey, R. D. (2012) Default bayes factors for model selection in regression. *Multivariate Behav. Res.*, **47**, 6, 877–903.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D. & Iverson, G. (2009) Bayesian t tests for accepting and rejecting the null hypothesis. *Psychon. Bull. & Rev.*, **16**, 2, 225–237.

Schönbrodt, F. D. & Wagenmakers, E.-J. (2018) Bayes factor design analysis: Planning for compelling evidence. *Psychonomic bulletin & review*, **25**, 1, 128–142.

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M. & Perugini, M. (2017) Sequential hypothesis testing with bayes factors: Efficiently testing mean differences. *Psychological Methods*, **22**, 2, 322.

Schroevers, M., Nykliček, I. & Topman, R. (2008) Validatie van de nederlandstalige versie van de mindful attention awareness scale (maas). *Gedragstherapie*.

Seli, P., Carriere, J. S. & Smilek, D. (2015) Not all mind wandering is created equal: Dissociating deliberate from spontaneous mind wandering. *Psychological Research*, **79**, 5, 750–758.

Shiozawa, P., Fregni, F., Benseñor, I. M., Lotufo, P. A., Berlim, M. T., Daskalakis, J. Z., Cordeiro, Q. & Brunoni, A. R. (2014) Transcranial direct current stimulation for major depression: an updated systematic review and meta-analysis. *Int. J. Neuropsychopharmacol.*, **17**, 09, 1443–1452.

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.

Simons, D. J. (2014) The value of direct replication. *Perspectives on Psychological Science*, **9**, 1, 76–80.

Simons, D. J., Holcombe, A. O. & Spellman, B. A. (2014) An introduction to registered replication reports at perspectives on psychological science. *Perspect. Psychol. Sci.*, **9**, 5, 552–555.

Smallwood, J., Brown, K., Baird, B. & Schooler, J. W. (2012) Cooperation between the default mode network and the frontal–parietal network in the production of an internal train of thought. *Brain Res.*, **1428**, 60–70.

Smallwood, J., Fitzgerald, A., Miles, L. K. & Phillips, L. H. (2009) Shifting moods, wandering minds: negative moods lead the mind to wander. *Emotion*, **9**, 2, 271.

Smallwood, J. & Schooler, J. W. (2006) The restless mind. *Psychol. Bull.*, **132**, 6, 946.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & Van Der Linde, A. (2002) Bayesian measures of model complexity and fit. *J. R. Stat. Soc. B*, **64**, 4, 583–639.

Spreng, R. N., Stevens, W. D., Chamberlain, J. P., Gilmore, A. W. & Schacter, D. L. (2010) Default network activity, coupled with the frontoparietal control network, supports goal-directed cognition. *Neuroimage*, **53**, 1, 303–317.

Stagg, C. J. & Nitsche, M. A. (2011) Physiological basis of transcranial direct current stimulation. *The Neuroscientist*, **17**, 1, 37–53.

Stan Development Team (2016) RStan: the R interface to Stan. URL http://mc-stan.org/. R package version 2.14.1.

Stawarczyk, D., Majerus, S., Maj, M., Van der Linden, M. & D'Argembeau, A. (2011) Mind-wandering: phenomenology and function as assessed with a novel experience sampling method. *Acta Psychol.*, **136**, 3, 370–381.

Thomson, D. R., Seli, P., Besner, D. & Smilek, D. (2014) On the link between mind wandering and task performance over time. *Conscious. Cogn.*, **27**, 14–26.

Tremblay, S., Lepage, J.-F., Latulipe-Loiselle, A., Fregni, F., Pascual-Leone, A. & Théoret, H. (2014) The Uncertain Outcome of Prefrontal tDCS. *Brain Stimulation*, **7**, 6, 773–783.

Turi, Z., Ambrus, G. G., Ho, K.-A., Sengupta, T., Paulus, W. & Antal, A. (2014) When size matters: large electrodes induce greater stimulation-related cutaneous discomfort than smaller electrodes at equivalent current density. *Brain stimul.*, **7**, 3, 460–467.

van de Vijver, F. J. R. & Leung, K (2011) Equivalence and bias: A review of concepts, models, and data analytic procedures. In Matsumoto, D. R. & van de Vijver, F. J. R. (eds.), *Cross-cultural research methods in psychology*, pp. 17–45. New York: Cambridge University Press.

Vanderhasselt, M.-A., Brunoni, A. R., Loeys, T., Boggio, P. S. & De Raedt, R. (2013) Nosce te ipsum–socrates revisited? controlling momentary ruminative self-referent thoughts by neuromodulation of emotional working memory. *Neuropsychologia*, **51**, 13, 2581–2589.

Vannorsdall, T. D., Van Steenburgh, J. J., Schretlen, D. J., Jayatillake, R., Skolasky, R. L. & Gordon, B. (2016) Reproducibility of tdcs results in a

randomized trial: failure to replicate findings of tdcs-induced enhancement of verbal fluency. *Cognitive and Behavioral Neurology*, **29**, 1, 11–17.

Vehtari, A., Gelman, A. & Gabry, J. (2015) *loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models*. URL https://github.com/jgabry/loo. R package version 0.1.3.

Verhagen, J. & Wagenmakers, E.-J. (2014) Bayesian tests to quantify the result of a replication attempt. *J. Exp. Psychol. Gen.*, **143**, 4, 1457.

Verplanken, B., Friborg, O., Wang, C. E., Trafimow, D. & Woolf, K. (2007) Mental habits: metacognitive reflection on negative self-thinking. *J. Pers. Soc. Psychol.*, **92**, 3, 526.

Vincent, J. L., Kahn, I., Snyder, A. Z., Raichle, M. E. & Buckner, R. L. (2008) Evidence for a frontoparietal control system revealed by intrinsic functional connectivity. *J. Neurophysiol.*, **100**, 6, 3328–3342.

Vöröslakos, M., Takeuchi, Y., Brinyiczki, K., Zombori, T., Oliva, A., Fernández-Ruiz, A., Kozák, G., Kincses, Z. T., Iványi, B., Buzsáki, G. et al. (2018) Direct effects of transcranial electric stimulation on brain circuits in rats and humans. *Nature communications*, **9**, 1, 483.

Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams Jr, R., Albohn, D., Allard, E., Benning, S., Blouin-Hudon, E.-M. et al. (2016) Registered replication report: Strack, martin, & stepper (1988). *Perspectives on Psychological Science*, **11**, 6, 917–928.

Watson, D., Clark, L. A. & Tellegen, A. (1988) Development and validation of brief measures of positive and negative affect: the panas scales. *J. Pers. Soc. Psychol.*, **54**, 6, 1063.

Weinstein, Y. (2017) Mind-wandering, how do i measure thee with probes? let me count the ways. *Behavior research methods*, pp. 1–20.

Weinstein, Y., De Lima, H. J. & van der Zee, T. (2018) Are you mind-wandering, or is your mind on task? the effect of probe framing on mind-wandering reports. *Psychonomic bulletin & review*, **25**, 2, 754–760.

Weissman, D., Roberts, K., Visscher, K. & Woldorff, M. (2006) The neural bases of momentary lapses in attention. *Nat. Neurosci.*, **9**, 7, 971–978.

Wiegmann, D., Faaborg, T., Boquet, A., Detwiler, C., Holcomb, K. & Shappell, S. (2005) Human error and general aviation accidents: A comprehensive, fine-grained analysis using hfacs. Technical report, Federal Aviation Administration.

Yanko, M. R. & Spalek, T. M. (2014) Driving with the wandering mind the effect that mind-wandering has on driving performance. *Hum. Factors*, **56**, 2, 260–269.