Attention, Perception, & Psychophysics
https://doi.org/10.3758/s13414-020-02131-x

The metronome response task for measuring mind wandering: Replication attempt and extension of three studies by Seli et al

Thomas Anderson[1] · Rotem Petranker[2] · Hause Lin[1] · Norman A.S. Farb[3]

Abstract

Science requires replicable tools to measure its intended constructs. Attention research has developed tools that have been used in mind-wandering research, but mind-wandering measures often rely on response-inhibition, which introduces speed-accuracy trade-offs that may conflate errors for mind-wandering. We sought to replicate three studies that used an improved mind-wandering measure: the Metronome Response Task (MRT). In a large (N=300) multisite sample, the primary MRT finding was replicated, showing that continuous rhythmic response time variability reliably predicted self-reported mind-wandering. Our findings also show previously undetected differences between intentional and unintentional mind-wandering. While previously reported mediators (motivation) and moderators (confidence) did not replicate, additional covariates add predictive value and additional constructs (e.g., boredom, effort) demonstrate convergent validity. The MRT is useful for inducing and measuring mind-wandering and provides an especially replicable tool. The MRT's measurement of attention could support future models of the complete cycle of sustained attention.

Keywords Mind-wandering · Attention · Variability · Behavioural · Continuous performance task · Boredom

Introduction

Paying attention is a constant challenge. Mind-wandering (MW), conceived of here as task-unrelated thoughts, has been linked to numerous deleterious outcomes including negative affect (Killingsworth & Gilbert, 2010), decreased reading comprehension (Franklin, Mooneyham, Baird, & Schooler, 2014), diminished driving ability (Yanko & Spalek, 2014), and lower cognitive test scores (Mrazek et al., 2012), though MW can be beneficial in some contexts, for example for creativity (Mooneyham & Schooler, 2013; Schooler et al., 2014). Efficiently detecting MW during tasks could allow for valuable corrective interventions, reducing the impact of attentional lapses.

Several behavioral tasks have been created to assess sustained attention (see Fortenbaugh, DeGutis, & Esterman, 2017, for an extended treatment), but these often emphasise response-inhibition rather than characterizing MW per se (Seli et al. 2013b). Two primary concerns emerge from such an approach: dichotomous categorization and speed-accuracy trade-offs. Consider the Sustained Attention to Response Task (SART; Robertson, Manly, Andrade, Baddeley, & Yiend, 1997) as an example. During the SART, participants are required to respond with a key-press to frequent on-screen non-targets (the numbers 1, 2, 4–9) and withhold responses for infrequent NOGO targets (the number 3). In such a paradigm, if a participant presses the response when the non-target is on-screen, this is considered an error that categorically indicates MW (i.e., erroneous dichotomous categorization as MW). It is possible that the participant was not MW, however,

Electronic supplementary material The online version of this article (https://doi.org/10.3758/s13414-020-02131-x) contains supplementary material, which is available to authorized users.

✉ Thomas Anderson
metathomas.anderson@mail.utoronto.ca

Rotem Petranker
rotem@boredomlab.org

Hause Lin
hause.lin@mail.utoronto.ca

Norman A.S. Farb
norman.farb@utoronto.ca

[1] Department of Psychology, University of Toronto, 27 King's College Cir, Toronto, ON M5S 3H7, Canada

[2] Department of Psychology, York University, 4700 Keele St, Toronto, ON M3J 1P3, Canada

[3] Department of Psychology, University of Toronto Mississauga, 3359 Mississauga Road, Mississauga, ON L5L 1C6, Canada

Published online: 30 September 2020

and was instead strategically focusing on responding as quickly as possible rather than as accurately as possible (i.e., a speed-accuracy trade-off). This seems plausible as SART response-times have been shown to mediate SART errors (Seli et al. 2013b). In contrast, during the non-target trials, participants could engage in mild and fleeting MW or deep, intentional bouts of MW that are never identified (i.e., erroneous dichotomous categorization as on-task). Degrees of MW depth cannot be captured by error-rates that cast MW as dichotomously present or absent.

While the correct/incorrect outcomes erroneously cast each trial as either completely on-task or completely MW, attempts have been made to use reaction time variability in response-inhibition tasks to model MW (Bastian & Sackur, 2013; Cheyne, Solman, Carriere, & Smilek, 2009) . Reaction-time variability is then used to model attention continuously rather than dichotomously, but in such cases a second problem emerges: task errors may represent response strategies, not attention. The use of go/no-go or target/non-target trials means participants are likely engaging in strategic speed-accuracy trade-offs that confound the measurement of attention with the relative strategic priority of speed or accuracy, which are at odds in such tasks. Indeed, changes in performance may reflect changes in strategy to adapt to the prevalence of particular target trials (Seli et al. 2013c; Fortenbaugh et al. 2017). As such, response inhibition measures of sustained attention focusing solely on accuracy may be ill-suited to assess MW as they rarely model the continuum of sustained attention and may erroneously associate changes in response strategy with MW.

The Metronome Response Task (MRT) was created to address these issues (Seli et al. 2013a). The MRT uses behavioural response variability in a paradigm in which every trial contains a response target as a measure of sustained attention. Attention is thus modelled on a continuum and response bias cannot be strategically tuned to improve task performance. In the MRT, participants tap along to a steady beat and occasional probes measure attentional states by asking participants if they were on-task or mind-wandering at probe-onset. Mind-wandering during the MRT has been associated with greater variability in tapping (Seli et al. 2013a). The primary outcome measure used in the MRT is the variability in responses relative to the metronome sound, called Rhythmic Response Time variance (RRTv). The RRTv is calculated using the five trials preceding each probe (see Fig. 1 and Supplemental Fig. 1 in the Online Supplemental Material (OSM)); raw variance tends to be positively skewed so a natural logarithm transform is performed (when referring to RRTv, we are always referring to the log-transformed RRTv).

While the same lab has published work using the MRT a number of times (Seli et al. 2013a, 2015a, 2015b, 2017b) this tool had not yet been independently replicated until very recently (Meier, 2018). Replication is the foundation of science, without which the field of psychology found itself in crisis. Here we report a further effort to replicate three MRT research articles across three independent lab sites. We preregistered (Anderson, Lin, & Petranker, 2017) hypotheses that the results of previous studies, summarized below and in the OSM 'Replication Summary', would be successfully replicated, and we included additional measures to extend previous work. The present work makes scientific contributions in three key ways: attempting a conceptual rather than direct replication, investigating findings that failed to directly replicate, and adding novel extensions that help contextualize the MRT as a MW measure.

First, both direct and conceptual replications of research findings are required to control for the numerous contextual influences of research conducted in a given site (Yarkoni, 2019). Replication is one of the central mechanisms by which science self-corrects (Ioannidis, 2012). Direct replications importantly test the replicability of findings using identical measures in different contexts. Conceptual replications, on the other hand, allow researchers to test whether alterations in measurement affect the replicability of the paradigm, which presents theoretical implications for the generality of the paradigm in a wider context. Our study has a number of these alterations, detailed in the following sections.

Second, not all aspects of the original MRT findings were successfully replicated in Meier (2018), particularly concerning the role of participant confidence ratings moderating the relationship between RRTv and self-reported MW. As such, additional well-powered research is required to characterize the presence or absence of these inconsistent findings. Indeed, in cases where an original finding cannot be directly replicated, it is not always apparent whether the replication or the original should be taken as more authoritative; a body of independent evidence is required to evaluate the effects.

Third, controls and extensions to the MRT paradigm are needed to foster consistency and determine the robustness of the paradigm as a dynamic measure of MW. For example, some, but not all, of the previous MRT studies included a control for time spent in the experiment (see OSM 'Replication Summary'). When this predictor was included, it was significant, which would be expected as the performance decrement is well described in the literature (Mackworth, 1964; Warm, Parasuraman, & Matthews, 2008). Other controls, such as expertise in rhythm, could also affect performance on this rhythm-based task, but such controls have not been used in the paradigm. Furthermore, it is relatively unclear what affective states the MRT elicits. We include measures to characterize these states before, during, and after the MRT, as has been done for other tools, for example SART.

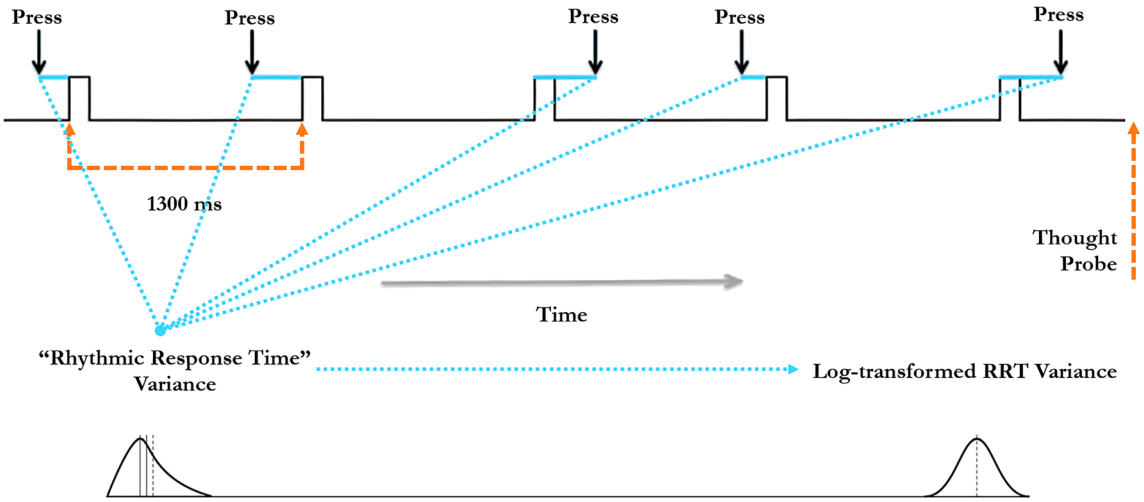

Fig. 1 Rhythmic Response Time variance (RRTv) is calculated as the natural log of variance in response-time in the five trials preceding each thought probe. When referring to RRTv, we are always referring to the log-transformed RRTv. Modified with permission from (Seli et al. 2013a)

Summary of previous findings and replication hypotheses

Report type: Wandering minds and wavering rhythms

Seli et al. (2013a) reported a main effect of report type such that RRTv was lower when participants reported being on-task (ON) and greater when participants were mind-wandering (MW). Mind-wandering was further divided into two types, mind-wandering with awareness ("tuned out"), in which participants knew they were off-task when presented with the probe, and mind-wandering without awareness ("zoned out"), in which participants were surprised to discover that they were off-task (Smallwood, Beach, Schooler, & Handy, 2007). Follow-up analyses found no significant difference between types of MW. Seli et al. (2013a) also reported a main effect of time (i.e., first- or second-half of the experiment) such that the second half of the experiment had higher RRTvs; they found no report type-by-time interaction. Rather than splitting the experiment into halves (i.e., probes 1–9, probes 10–18), the present study indexes time at each individual probe (1–18). Seli et al. (2013a) reported on two samples; the final sample sizes were 41 and 39 respectively.

We performed a conceptual replication, focusing instead on intentionality of MW, as in Seli et al. (2015a). As in their previous work, Seli et al. (2015a) also reported a main effect of report type with lower RRTv when ON compared to MW. MW was further divided into two types: intentionally thinking about task-unrelated thoughts and unintentionally thinking about task-unrelated thoughts. Focusing on MW intentionality allows us to test mediation models presented in Seli et al. (2015a). As such, in the present study, MWg was divided into two types: MW unintentionally (MWu), in which participants reported that they intended to remain on-task but drifted into MW, and MW intentionally (MWi), in which participants intentionally disengaged from the task. We pre-registered the following Report Type hypotheses:

H1a: MW reports will have greater RRTv than ON reports.H1b: No significant differences in RRTv will emerge between MWu and MWi.

Motivation: Motivation, intentionality and mind-wandering

The main focus of Seli et al. (2015a) was motivation, which was measured as a single item following the MRT. They found that motivation significantly predicted mean RRTv such that those with higher motivation performed better on the task (i.e., had lower mean RRTv). Mediation analyses found that the effect of motivation on mean RRTv was mediated by the proportion of probes that indicated MW. Relationships between motivation, mean RRTv and intentionality of MW were also reported and the authors concluded that intentionality of MW was unrelated to task-performance. Seli et al. (2015a) reported on a sample of 166 participants. We performed a conceptual replication using a modified 0–100 slider scale rather than Seli et al.'s (2015a) 1–7 scale. We pre-registered the following Motivation hypotheses:

H2a: Motivation and Mean RRTv will negatively correlate.
H2b: Proportion of MW will mediate the relationship between Motivation and Mean RRTv.

Confidence: Can research participants comment authoritatively on the validity of their self-reports

In Seli et al. (2015b) participants reported mental state (ON, MW) and how confident they felt in the accuracy of their self-reported mental states. Measuring confidence aimed to further validate the MRT as a performance-based index of MW by identifying situations where a lower correlation between behavioural variability and self-report might be expected. Mean RRTv correlated with the proportion of probes reported as MW and, critically, mean confidence moderated this correlation: Participants with high mean confidence showed significant effects of report type, but these effects were non-significant in participants with low mean confidence. Trial-level analyses showed an interaction between report type and confidence. Seli et al. (2015b) reported on a final sample of 100 participants. We performed a conceptual replication using novel probes intended to investigate this effect further. The new probes included "Mostly" and "Completely" variants of the ON and MW responses (see Methods) and were explored as a measure of mind-wandering depth, which may interact with response confidence as proposed in Seli et al. (2015b). We pre-registered the following Confidence hypotheses:

> H3: Confidence will moderate the H1a relation between report type and RRTv. Specifically:
> High-Confidence On-Task will have the lowest RRTv
> High-Confidence Mind-Wandering will have the highest RRTv
> Low-Confidence On-Task will have moderate RRTv
> Low-Confidence Mind-Wandering will have moderate RRTv

Meier (2018) recently attempted to directly replicate Seli et al. (2015b) and this direct replication attempt failed to replicate the moderating effect of confidence. Some of the methods used suggested that probes with the highest confidence may drive some tentative interaction with report-type in predicting RRTv, but, overall, the pattern of results was different to the original and the author "tentatively conclude[s] that participants cannot authoritatively comment on the validity of thought probes by completing retrospective confidence reports."(Meier, 2018, p. 1575) The author further suggests that any effects that may exist "appear relatively weak and noisy". The author suggests instead that other variables, such as working memory capacity and personality variables, may contribute to individual differences in MRT performance, though moderating effects were not ultimately found. Meier's (2018) replication had a final sample of 291.

Control and extension hypotheses

In addition to replicating previous findings, we introduced a number of control variables of interest and also included measurements of mood and mental states thought to be related to MW in order to investigate convergent validity.

Control variables

We included two hypotheses concerning participant-level covariates of interest. The MRT relies on tapping a constant rhythm and it is plausible that experience with rhythm-keeping could result in better performance independent of sustained attention and MW per se. We hypothesized that participants with higher levels of experience with musical instruments (e.g., drumming) and/or modern rhythm-games (e.g., Rock Band, Guitar Hero) would have greater baseline proficiency for tapping a constant rhythm, thus we pre-registered the following hypothesis:

> C1: Music and Rhythm-Game Experience will negatively predict Mean RRTv.

The RRTv is purported to index sustained attention and MW, thus we included a trait-measure of attention-lapses, the Attention Related Cognitive Errors Scale (ARCES; Cheyne, Carriere, & Smilek, 2006), which has been correlated with errors in the SART. This participant-level measure should show convergent validity with the MRT. We pre-registered the following hypothesis:

> C2: Attention Related Cognitive Errors will positively predict Mean RRTv.

Extension hypotheses

The MRT is a low-demand task that might be associated with unique affective and attentional states that differ from those elicited by higher demand or response-inhibition tasks (e.g., Spunt, Lieberman, Cohen, & Eisenberger, 2012). Since the MRT was developed recently and it is relatively unclear what states it elicits, we aimed to characterize these states by measuring task perceptions before and after the task (i.e., prospective and retrospective boredom, frustration, discomfort/distress, fatigue and effort). MW is associated with negative affective states, such as boredom (Eastwood, Frischen, Fenske, & Smilek, 2012), and boredom- and fatigue-related performance decrements increase over time (Davies & Parasuraman, 1982). As such, we sought to verify that the MRT reliably induces such negative affect and performance decrements. We also predicted that participant boredom

would increase throughout the task, leading to higher retrospective than prospective ratings of boredom.

Our previous work suggests that participants tend to overestimate how frustrating future tasks will be, i.e., at the end of an experiment they report tasks were less frustrating than anticipated (Lin & Inzlicht, unpublished data); we thus expected participants to prospectively anticipate more frustration than they would report following the task. Sustained attention tasks induce the aversive state of boredom (Scerbo, 1998), so we investigated whether the MRT might also affect participant discomfort/distress; we predicted that prospective/retrospective discomfort/distress would not differ because the MRT relies on sustained auditory attention, which is less distressing than sustained visual attention (Galinsky, Rosa, Warm, & Dember, 1993; Szalma et al., 2004). Finally, the simplicity of the MRT led us to expect that participants would make more accurate prospective judgements about fatigue and effort; as such, we predicted no discrepancy between prospective and retrospective judgements on fatigue and effort.

E1: Boredom will be positively correlated with Mean RRTv
E2: Fatigue will be positively correlated with Mean RRTv
B1: Retrospective Boredom will be higher than Prospective Boredom
B2: Prospective Frustration will be higher than Retrospective Frustration
B3: Discomfort/Distress, Fatigue, and Effort will not differ Prospective to Retrospective

Deviations from pre-registered hypotheses

The pre-registration used the wording "Overall Performance", which was used in some of the original studies (Anderson et al., 2017). "Overall Performance" was defined as "the participant's average variability throughout the MRT (calculated using five-trial sliding windows)". For the sake of clarity, we rename "Overall Performance" with the term "Mean RRTv" and adjust the direction of effect to reflect this (i.e., lower Overall Performance is higher Mean RRTv and vice-versa). We report acronyms rather than verbatim hypotheses (Anderson et al., 2017). Additionally, the pre-registration erroneously left out the main effect of Confidence in the relevant statistical model: this main effect has been included. Including the main effect makes interpretation of parameter values more straightforward (UCLA Statistical Consulting Group, 2018). Finally, the pre-registration did not account for research assistants restarting the MRT in cases of participants not understanding that they needed to respond to the metronome with a keypress. MRT data files with less than 90 total trials are treated as restarted sessions (n=19: UTM n=1, UTSC n=3, York n=10, unknown n=5) rather than counting them toward participant attrition rates. These files typically contained fewer than five spacebar responses.

Methods

Participants

All procedures were conducted under informed consent in accordance with the Declaration of Helsinki. Undergraduates from three testing sites were recruited: University of Toronto Mississauga campus (UTM), University of Toronto Scarborough campus (UTSC) and York University campus. Participants participated in lab sessions in exchange for course-credit; all participants were provided with over-the-ear headphones. Depending on the available research space, participants were either provided with a private booth or were separated from other participants by room dividers. Each testing site recruited participants until 100 participants not meeting exclusion criteria were obtained (N=300). The power analysis for this study is lengthy and complex, thus it has been reported as OSM "Power Analysis".

In total 31 participants were excluded. As in the original studies, if participants missed more than 10% of trials they were dropped (n=22: UTM n=3, UTSC n=7, York n=10, unknown n=2); in contrast to the original studies, the programming was modified to warn participants when they had missed 8% of trials in an attempt to prevent this problem. Participants were also removed if they indicated that they misunderstood (n=3) or participated in bad faith (n=3: e.g., "I found my absent-mindedness was so great at times that [I] didn't realize that [I] was answering questions with the same answers. Instead, [I] should have been answering using a different one (i.e., intentionally instead of unintentionally)") as per our pre-registered exclusion criteria. Finally, one participant was dropped for mechanical error (faulty keyboard) and two participants requested their responses be removed. For more details, see OSM "Attrition Rate Commentary".

Procedure

Metronome Response Task (MRT)

In the MRT participants tap along to a steady aural beat (Seli et al. 2013a). As in Seli et al. (2013a), MRT trials were as follows: 650 ms of silence, followed by a metronome tone lasting approximately 75 ms, followed by another 575 ms of silence, resulting in a total trial duration of 1,300 ms; trials followed one another immediately. Participants were instructed to use the spacebar to tap along with the tone, which was presented via headphones. Participants completed 900 experimental trials (19.5 min) with 18 probes. Specifically,

participants read on-screen instructions that said: "For this experiment you will hear a metronome sound presented at a constant rate via the headphones. Your task is to press the <spacebar> in synchrony with the onset of the metronome so that you press the space bar exactly when each metronome sound is presented." Full instruction text and Python code to run the MRT are available on the OSF (Anderson et al., 2017).

Report type

Over the 900 trials, 18 thought-probes were randomly presented on the computer screen, with one probe occurring in every block of 50 trials (blocks were not explicitly demarcated for participants). In order to prevent probes from occurring too close together (i.e., at the very end of one block and the very beginning of the next) a minimum of five trials were kept between probes. As in the original study the onset of a probe paused the MRT and participants indicated their attentional state immediately prior to each probe by pressing the appropriate numerical key. The original study used three report types: on-task, tuned out or zoned out. In an attempt to measure subtler distinctions our version used six report types: (1) Completely On-Task, (2) Mostly On-Task, (3) Completely Mind-Wandering Unintentionally, (4) Mostly Mind-Wandering Unintentionally, (5) Completely Mind-Wandering Intentionally, (6) Mostly Mind-Wandering Intentionally. At the beginning of the study all options were defined for participants: "On-task" was defined for participants to mean focusing on completing the task (i.e., performance, responses, or being totally focused on the task). "Mind-Wandering" was defined to mean thinking about something unrelated to the task (i.e., about courses, plans with friends, food, or any other thoughts not related to the experiment). "Mind-Wandering Unintentionally" was defined to mean specifically that thoughts drifted away from the task despite an intention to focus on the task whereas "Mind-Wandering Intentionally" was defined to mean having decided to think about unrelated things.

Motivation

After completing the MRT, participants indicated Motivation ("How motivated were you to do well on the task?") on a 0–100 slider scale.

Confidence

Following each thought-probe, participants also indicated the degree to which they were confident in their response. We used a 1–6 scale with end-point nominal descriptors "Not At All Confident" and "Extremely Confident". After providing responses participants pressed the spacebar to resume the MRT.

Control and exploratory questions

After completing the MRT, participants filled out a short series of questionnaires. After indicating Motivation, participants self-rated Performance ("How well do you think you did on the task compared to other people?" on a 0–100 slider scale under an image of a truncated standard normal distribution). Participants then completed the Attention-Related Cognitive Errors scale (Cheyne et al., 2006) on a 0–100 slider scales with nominal end-point anchors of "Never" and "Very Often"; this measure had excellent internal reliability (Cohen's $\alpha = 0.89$). They subsequently indicated their level of experience with music and rhythm games ("Do you have any experience with drumming or other rhythm-based instruments?" None (0), A Little (1), Some (2), A Lot (3); "Do you have any experience with Rock Band, Guitar Hero, or other rhythm-based games?" None (0), A Little (1), Some (2), A Lot (3)); these values were summed to create a "Timing Experience" score (0–6). Finally, participants reported demographic variables and were allowed to give open-ended feedback about the MRT and experiment as a whole.

Extension questions

Five extension questions were asked before and after completion of the MRT. We asked, prospectively and retrospectively, how much mental effort, frustration, discomfort or distress, boredom and mental fatigue the participant thought the task would make them experience. Participants also indicated the amount of mental effort the task was requiring and the amount of frustration they were feeling twice during the MRT at one-third and two-thirds into the experiment (i.e., on the sixth and 12th probes). These used the same 1–6 scales with nominal descriptors "None" to "A lot" and were reported after the report type and confidence questions in the probe.

Data analysis

The R Language (R Core Team, 2014) was used for statistical analyses using the nlme package (Pinheiro et al., 2018) for modelling and calculating degrees of freedom and p-values (Pinheiro & Bates, 2009), the lavaan package (Rosseel, 2012) for mediation analysis, the hausekeep package (Lin, 2019) for reporting, and the multcomp package (Hothorn et al., 2019) for multiple comparisons in exploratory analyses. Participants were included in all models using random intercepts alongside the fixed-effects of interest, explained below. Probe-level predictors were within-participant standardized and participant-level predictors were standardized across participants. Significant effects for primary analyses survived Holm-Bonferroni corrections for multiple comparisons (see OSM "Holm-Bonferroni correction"). Descriptive statistics

can be seen in Supplemental Table 1 (OSM) and zero-order correlations can be seen in Supplemental Table 2 (OSM).

Report type: ANOVAs

Replicating Seli et al.'s (2013a) analysis we assessed differences in RRTv with a Block × Report Type 2 × 3 within-participant ANOVA. Participant was a random factor and Block (first-half, second-half) and Report Type (On-Task (ON), Mind-Wandering Unintentionally (MWu), Mind-Wandering Intentionally (MWi)) were fixed factors (collapsing over "Completely" and "Mostly"). Follow-up t-tests between ON reports and each type of MW report (MWu, MWi) as well as between the two types of MW report were computed.

Report type: Regression

To test H1a we regressed RRTv onto Probe Number (1–18) and Report Type (ON and MW (collapsing over MWu and MWi)), including the interaction term; the critical test is the main effect of Report Type. For H1b we fitted the same model for RRTv but included the three levels of Report Type (ON, MWu, MWi), including the interaction terms with Probe Number; the critical test is the main effect of Report Type, particularly between MWu and MWi. While the full model was pre-registered, the performance decrement (main effect of probe number) and the interactions between probe number and report type were not explicitly pre-registered as confirmatory and thus may be treated as exploratory.

Motivation: Mediation

Replicating Seli et al.'s (2015a) analysis we tested Motivation as a predictor of Mean RRTv as well as the proportion of MW reports (PropMW) as a mediator of this relationship. Motivation was also tested as a predictor of the proportion of MW reports that were intentional (PropMWi) relative to all MW reports.

Confidence: Regression

To test H3 we added Confidence and its interaction terms with Report Type as additional predictors in the model testing H1b. That is, we regressed RRTv onto Probe Number (1–18), Report Type (ON, MWu, MWi), Probe Number by Report Type interaction, Confidence, and Confidence by Report Type interaction.

Control: Regression

To test C1 and C2 we added Timing Experience and ARCES scores as additional predictors in the model testing H3. That is, we regressed RRTv onto Probe Number (1–18), Report Type (ON, MWu, MWi), Probe Number by Report Type interaction, Confidence, and Confidence by Report Type interaction, Timing Experience, and ARCES.

Extension: Correlations and paired t-tests

Correlations were used to test E1–2 and paired t-tests were used to test B1–3.

Exploration: Report-type extension: Regression

To test our exploratory models, we used the full six-level version of Report Type. That is, rather than modelling Report Type as a three-level categorical variable (ON, MWu, MWi), we modelled Report Type with all six levels (Completely ON, Mostly ON, Mostly MWu, Completely MWu, Mostly MWi, Completely MWi). The resultant linear mixed model included Report Type, Probe Number, Report Type by Probe Number interaction, Timing Experience, and ARCES. Participant was included as a random factor as in all models.

Results

Pre-registered replication hypotheses

Report type

ANOVAs The ANOVA results in Seli et al. (2013a) were fully replicated with a significant main effect of Report Type ($F(2,5386) = 58.092$, $p < .0001$), a significant main effect of Block ($F(1,5386) = 49.79$, $p < .0001$), such that the second half of the experiment had higher RRTvs, and a non-significant Report Type by Block interaction ($F(2, 5386) = 1.13$, $p = .323$). Follow-up t-tests between Report Types also replicated findings from Seli et al. (2015a) with ON reports having significantly lower RRTv than MWu ($t(3531) = 7.65$, $p < .001$, $r = 0.13$) and MWi ($t(2648) = 9.84$, $p < .001$, $r = 0.19$). In contrast to previous research, MWu had significantly lower RRTvs than MWi ($t(2879) = 2.57$, $p = .010$, $r = 0.05$), which may not have been detectable in smaller samples (Seli et al. 2015a).

Regression Supporting H1a, there was a significant main effect of Report Type ($b = 0.29$, $SE = 0.04$, $t(5090) = 6.64$, $p < .001$, $r = 0.09$) and Probe Number ($b = 0.13$, $SE = 0.03$, $t(5090) = 4.60$, $p < .001$, $r = 0.06$) as predictors of trial-by-trial RRTv. In contrast to the original research, the Report Type by Probe Number interaction was significant ($b = 0.09$, $SE = 0.04$, $t(5090) = 2.32$, $p = .020$, $r = 0.03$): RRTv

associated with MW increased at a faster rate over the course of the experiment compared to RRTv associated with ON reports.

When testing H1b, subtler effects were present (Fig. 2, Table 1a). A main effect of Report Type was found such that both MWu and MWi predicted higher RRTv than ON (MWu: $b = 0.25$, $SE = 0.05$, $t(5088) = 5.12$, $p < .001$, $r = 0.07$; MWi: $b = 0.39$, $SE = 0.06$, $t(5088) = 6.77$, $p < .001$, $r = 0.09$). The Report Type by Probe Number interaction was significant for MWu ($b = 0.12$, $SE = 0.04$, $t(5088) = 2.57$, $p = .010$, $r = 0.04$) such that RRTv associated with MWu increased at a faster rate than ON reports (as in H1a). No such interaction existed for MWi ($p = .434$), that is, there was no significant difference between the rate at which ON and MWi RRTv increased over the duration of the experiment. The main effect of Probe Number remained significant ($b = 0.13$, $SE = 0.03$, $t(5088) = 4.56$, $p < .001$, $r = 0.06$).

Motivation

Against H2a no correlation between Motivation and Mean RRTv was found ($r(298) = -0.08$, $p = .187$, 95% CI [-0.19, 0.04]). The pre-registered mediation was still computed to test for an indirect effect of motivation on Mean RRTv through the

Table 1 Stepwise models of RRTv. Panel (a) reports model H1b. Panel (b) reports model H3. Panel (c) reports model C2, the final model including all covariates of interest

	B	SE	t-value	p-value	sig	r-value
(a) Predictors						
(Intercept) (ON)	$b = 8.21$	$SE = 0.05$	$t(5088) = 165.63$	$p < .001$	***	$r = 0.92$
MWu	$b = 0.25$	$SE = 0.05$	$t(5088) = 5.12$	$p < .001$	***	$r = 0.07$
MWi	$b = 0.39$	$SE = 0.06$	$t(5088) = 6.77$	$p < .001$	***	$r = 0.09$
ProbeNumber	$b = 0.13$	$SE = 0.03$	$t(5088) = 4.56$	$p < .001$	***	$r = 0.06$
ProbeNumber × MWu	$b = 0.12$	$SE = 0.04$	$t(5088) = 2.57$	$p = .010$	**	$r = 0.04$
ProbeNumber × MWi	$b = 0.04$	$SE = 0.05$	$t(5088) = 0.78$	$p = .434$	ns	$r = 0.01$
(b) Predictor						
(Intercept) (ON)	$b = 8.24$	$SE = 0.05$	$t(4559) = 155.93$	$p < .001$	***	$r = 0.92$
MWu	$b = 0.23$	$SE = 0.05$	$t(4559) = 4.57$	$p < .001$	***	$r = 0.07$
MWi	$b = 0.36$	$SE = 0.06$	$t(4559) = 5.92$	$p < .001$	***	$r = 0.09$
ProbeNumber	$b = 0.12$	$SE = 0.03$	$t(4559) = 4.10$	$p < .001$	***	$r = 0.06$
Confidence	$b = 0.12$	$SE = 0.03$	$t(4559) = 4.12$	$p < .001$	***	$r = 0.06$
ProbeNumber × MWu	$b = 0.12$	$SE = 0.05$	$t(4559) = 2.65$	$p = .008$	**	$r = 0.04$
ProbeNumber × MWi	$b = 0.04$	$SE = 0.05$	$t(4559) = 0.75$	$p = .451$	ns	$r = 0.01$
Confidence × MWu	$b = -0.08$	$SE = 0.05$	$t(4559) = -1.65$	$p = .098$	ns	$r = 0.02$
Confidence × MWi	$b = -0.03$	$SE = 0.05$	$t(4559) = -0.63$	$p = .531$	ns	$r = 0.009$
(c) Predictors						
(Intercept) (ON)	$b = 8.24$	$SE = 0.05$	$t(4559) = 161.65$	$p < .001$	***	$r = 0.92$
MWu	$b = 0.23$	$SE = 0.05$	$t(4559) = 4.48$	$p < .001$	***	$r = 0.07$
MWi	$b = 0.35$	$SE = 0.06$	$t(4559) = 5.83$	$p < .001$	***	$r = 0.09$
ProbeNumber	$b = 0.12$	$SE = 0.03$	$t(4559) = 4.14$	$p < .001$	***	$r = 0.06$
Confidence	$b = 0.13$	$SE = 0.03$	$t(4559) = 4.17$	$p < .001$	***	$r = 0.06$
Timing Experience	$b = -0.21$	$SE = 0.04$	$t(266) = -4.85$	$p < .001$	***	$r = 0.28$
ARCES	$b = 0.10$	$SE = 0.04$	$t(266) = 2.14$	$p = .033$	*	$r = 0.13$
ProbeNumber × MWu	$b = 0.12$	$SE = 0.05$	$t(4559) = 2.66$	$p = .008$	**	$r = 0.04$
ProbeNumber × MWi	$b = 0.04$	$SE = 0.05$	$t(4559) = 0.70$	$p = .486$	ns	$r = 0.01$
Confidence × MWu	$b = -0.08$	$SE = 0.05$	$t(4559) = -1.66$	$p = .098$	ns	$r = 0.02$
Confidence × MWi	$b = -0.04$	$SE = 0.05$	$t(4559) = -0.74$	$p = .459$	ns	$r = 0.01$

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

Participant was a random-factor in all models

RRTv Rhythmic Response Time variance, *ON* on-task, *MWu* mind-wandering unintentionally, *MWi* mind-wandering intentionally, *ARCES* Attention Related Cognitive Errors Scale

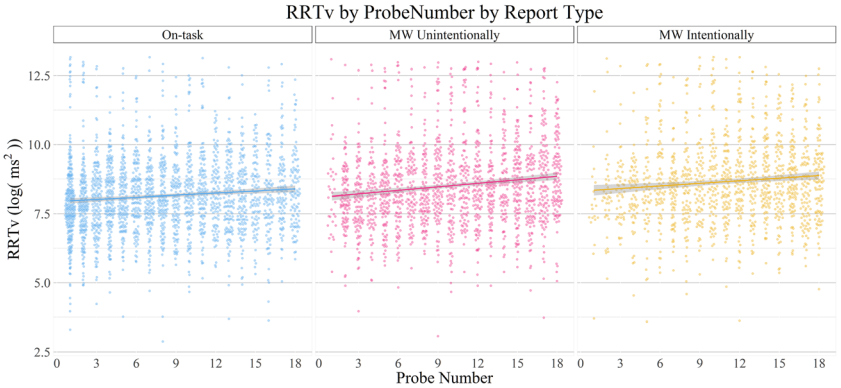

Fig. 2 Linear regressions of H1b showing main effects of ON vs. MWi and interaction of MWu with Probe Number. Shaded area reflects 95% confidence interval around fitted model. Each dot represents one data-point. *RRTv* Rhythmic Response Time variance, *ON* on-task, *MWu* mind-wandering unintentionally, *MWi* mind-wandering intentionally

proportion of MW reported relative to all probe reports (PropMW, Fig. 3a). Motivation negatively predicted PropMW ($r(298) = -0.41$, $p < .001$, 95% CI [-0.50, -0.31]), and PropMW positively predicted Mean RRTv ($r(298) = 0.19$, $p = .001$, 95% CI [0.08, 0.29]). The standardized indirect effect of Motivation on Mean RRTv was significant (indirect: -0.047, $p = .009$, 95% CI [-.087, -.015]).

Motivation also negatively predicted the proportion of MWi reports relative to total MW reports (PropMWi, Fig. 3b, $r(284) = -0.33$, $p < .001$, 95% CI [-43, -0.23]), but PropMWi was not significantly correlated with Mean RRTv ($r(284) = 0.11$, $p = .067$, 95% CI [-0.01, 0.22]) and the standardized indirect effect was non-significant (indirect: -0.022, $p = .079$, 95% CI [-.046, .002]).

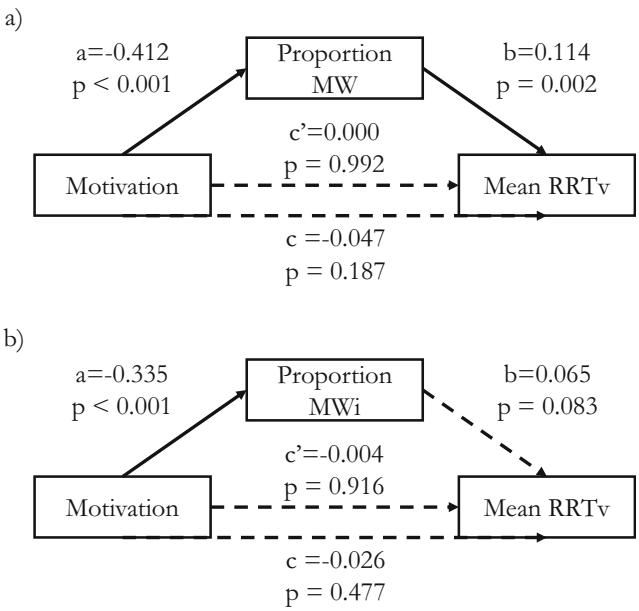

Fig. 3 Mediation analyses (H2b). Note the lack of direct effect to be mediated, counter to H2a. *RRTv* Rhythmic Response Time variance, *ON* on-task, *MW* mind-wandering, *MWi* mind-wandering intentionally

Confidence

Counter to H3, Confidence did not significantly interact with Report Type as a predictor of RRTv (see Table 1b). As only the main effect of Confidence was significant, these results indicate a failure to conceptually replicate the effect of confidence found in the original (Seli et al. 2015b). These findings are in line with another recent failure to replicate the original Confidence findings when using regression (Meier, 2018). Our sample also differed in the range of confidence ratings reported as participants overwhelmingly reported low confidence (1 or 2: 75.1%) and moderate confidence (3 or 4: 20.4%), relatively rarely reporting high confidence (5 or 6: 4.4%) (cf. Seli et al. (2015b), and Meier, 2018, wherein participants seldom reported low confidence).

Pre-registered control and extension hypotheses

Control

As predicted, each control variable was significantly correlated with Mean RRTv (Music Experience: $r(298) = -0.29$, $p < .001$; Rhythm-Game Experience: $r(298) = -0.24$, $p < .001$) such that greater experience resulted in lower Mean RRTv (i.e., better task performance). C1 was supported as Timing Experience was a significant predictor of RRTv ($b = -0.21$, $SE = 0.04$, $t(266) = -4.85$, $p < .001$, $r = 0.28$) with a larger effect size than other predictors. Similarly, the ARCES was significantly correlated with Mean RRTv ($r(298) = 0.18$, $p = .001$) such that higher scale scores predicted higher Mean RRTv. C2 was also supported as ARCES predicted Mean RRTv ($b = 0.10$, $SE = 0.04$, $t(266) = 2.14$, $p = .033$, $r = 0.13$), even among

the other predictors (see Table 1c). The ARCES had excellent internal reliability ($\alpha = 0.89$).

Extension

Boredom and Fatigue were each positively correlated with Mean RRTv (Boredom: $r(298) = 0.20$, $p < .001$; Fatigue: $r(298) = 0.16$, $p = .004$), supporting hypotheses E1 and E2. B1 was supported as retrospective Boredom was higher than prospective ($t(299) = -13.14$, $p < .001$, $r = 0.61$), but B2 was not supported as retrospective Frustration was higher than prospective ($t(299) = -3.65$, $p < .001$, $r = 0.21$). Counter to B3 participants found the MRT more Discomforting/ Distressing than anticipated ($t(299) = -4.99$, $p < .001$, $r = 0.28$), more Fatiguing than anticipated ($t(299) = -5.08$, $p < .001$, $r = 0.28$), and less Effortful than anticipated ($t(299) = 5.66$, $p < .001$, $r = 0.31$).

Pre-registered exploration

Report-type extension: Mostly or completely mind-wandering?

Participants made full use of the extended response options (Fig. 4) though not all participants selected every option. Specifically, 55 participants reported all six attention states, 170 participants reported four to five different attention states; 55 reported three, 15 reported two, and 5 reported only one attention state during the experiment. The most commonly reported attention states were Mostly ON (n = 270) and Mostly MWu (n = 252); the least common attention state was Completely MWi (n = 165). Though relatively rare, some participants reported exclusively MW states (n =14) and others reported exclusively ON states (n = 17).

When split by "Mostly" and "Completely", rather than collapsed, mean RRTvs for depth aligned with the order of effects we originally hypothesized for Confidence (Fig. 5): Completely ON had the lowest RRTv (M=8.07, SD=1.44), followed by moderate variability in Mostly ON (M=8.20, SD=1.41) and Mostly MWu (M=8.44, SD=1.45), with the highest RRTv associated with reports of Completely MWu (M=8.60, SD=1.58), Mostly MWi (M=8.61, SD=1.50), and Completely MWi (M=8.69, SD=1.57). A post hoc test using Tukey's HSD with adjusted p-values (p_{adj}) – calculated as per Hothorn, Bretz, and Westfall (2008) – suggests the significant differences exist primarily between ON and MW reports. Completely ON had lower RRTvs than all other conditions (all $p_{adj} < .001$). Mostly ON was lower than all other conditions (all $p_{adj} <= .001$) other than Mostly MWu ($p_{adj} = .35$). Mostly MWu had lower RRTvs than Completely MWi ($p_{adj} < .05$). All other MW to MW comparisons were non-significant (all $p_{adj} > .24$).

Exploration revealed that Report Type was a predictor of Confidence. Reports endorsing "Mostly" had significantly higher Confidence than reports endorsing "Completely" ($b = 0.33$, $SE = 0.03$, $t(4572) = 11.85$, $p < .001$, $r = 0.17$). ON reports had significantly lower Confidence than MWu ($b = 0.25$, $SE = 0.03$, $t(4571) = 7.81$, $p < .001$, $r = 0.11$) and MWi ($b = 0.21$, $SE = 0.03$, $t(4571) = 5.94$, $p < .001$, $r = 0.09$).

Self-reported performance accuracy

A planned exploratory model showed a significant correlation between self-rated performance and Mean RRTv ($r(298) = -0.25$, $p < .001$), demonstrating that participants had accurate senses of how they performed relative to their peers. Self-rated performance was positively correlated with Motivation ($r(298) = 0.38$, $p < .001$) and inversely correlated with mean Confidence ($r(298) = -0.23$, $p < .001$).

Early and late responders

Only three participants were early responders (i.e., mean response before the metronome), thus the effect of early and late responders was not investigated.

Discussion

This pre-registered replication and extension study (Anderson et al., 2017) investigated the Metronome Response Task (MRT), a continuous performance task wherein response variability indexes sustained attention and mind-wandering (MW). We tested a number of pre-registered hypotheses concerning rhythmic response time variability (RRTv) with large samples from three independent labs. While we successfully replicated the association between MW and greater RRTv, the direct and mediated effects of motivation and the moderating effects of confidence on MRT performance were not replicated. Furthermore, our results add nuance to the relation between MW and RRTv as we found that this association varies according to the intentionality of MW. While MW intentionally (MWi) was consistently associated with elevated RRTv, MW unintentionally (MWu) interacted with time spent in the experiment such that MWu had a greater deleterious impact on RRTv as the experiment progressed (Fig. 2). This significant interaction between report type and probe number suggests that differences between MWu and on-task (ON) reports may be less detectable by statistical models early in the MRT and so we caution against using short versions of the task for monitoring MWu effects, especially considering that MW appears most reported in later trials of the experiment (Fig. 4).

Counter to H1b, there was a main-effect differentiating intentional versus unintentional MW, which has not been

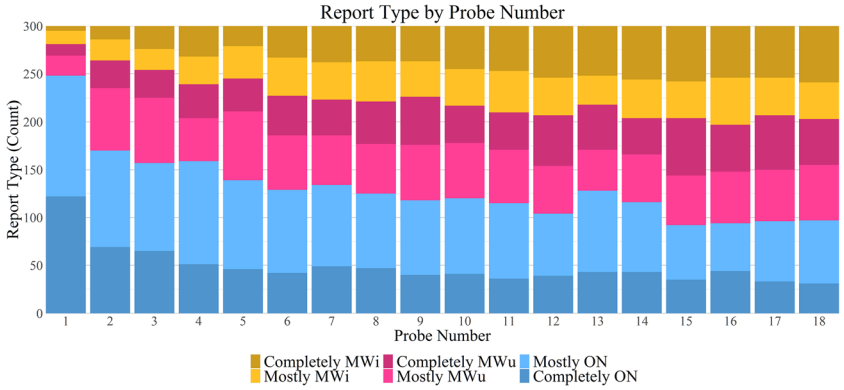

Fig. 4 Proportions of reports of each fine-grained attention state throughout the experiment. Participants made full use of the extended options and the drift toward deeper levels of mind-wandering is apparent. *ON* on-task, *MWu* mind-wandering unintentionally, *MWi* mind-wandering intentionally

consistently found in previous research (Seli et al. 2013a, 2015a, 2015b). The effect size was small and thus may only be detectable in larger samples. In agreement with the wider literature (Seli et al. 2016, 2017a) these results suggest that the difference between intentional versus unintentional MW is still worthy of further consideration; perhaps multi-modal methods (e.g., EEG, eye-tracking) and more complex modelling (e.g., non-linear computational modelling) may be in order to help distinguish intentional from unintentional MW in smaller samples as has been investigated for MW more broadly (Kawashima & Kumano, 2017). We also agree with Seli et al. (2016, 2017a) that the implications for intentional MW extend beyond studies of MW per se: it is important that researchers in any area of psychology that involves interpreting measures of task performance take into account that participants are, in many cases, intentionally disengaged from experiments in which they participate. When interpreting results researchers should reconsider the assumption that participants are fully engaged and attempting to perform optimally by addressing how interpretation of results would differ if this assumption were untenable.

Concerning motivation, findings from Seli et al. (2015a) were not replicated. Finding that motivation did not directly predict performance suggests that performance on the MRT may index qualities of behaviour and attention independent of participants' attempts to engage the task. Indeed, consistent with the expected phenomenology of the MRT it may be that, despite any attempt to engage, participants will inevitably be drawn into MW. Similarly, no matter how disengaged a participant may be, the RRTv can still measure fluctuating attentional states so long as the participant is engaged enough to tap along to the beat. This motivation is not trivial as 22 participants did not attain even minimum tapping consistency over the study period and were excluded for missing 10% of trials; as such, the MRT cannot measure the MW of participants demonstrating extreme disengagement. However, given that most participants were retained, these findings suggest that the MRT could be a useful tool for assessing MW even in minimally motivated samples.

One constraint on the generality of motivation findings is that no high-motivation condition was included. As the present study offered only course credit for participation these results may not generalize to effects found under a performance-based financial or other incentive. Indeed, Seli et al. (2017b) demonstrated that a motivation-condition showed decreased MW compared to a control condition.

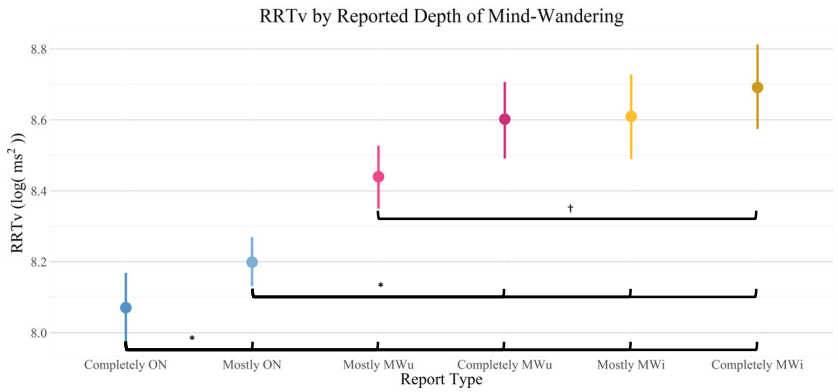

Fig. 5 Mean RRTvs for reports of each fine-grained attention state (error bars indicate bootstrapped SEM). Horizontal ticks indicate post hoc Tukey HSD significant differences. The predicted order based on depth of mind-wandering was found. * post-hoc $p_{adj} <= 0.001$, † post-hoc $p_{adj} < 0.05$. *RRTv* Rhythmic Response Time variance, *ON* on-task, *MWi* mind-wandering intentionally, *MWu* mnd-wandering unintentionally

They also reported a main effect of condition (motivation vs. control) on mean RRTv, whereas our correlational approach showed no relationship between motivation and mean RRTv. Seli et al. (2017b) also showed that this motivation effect was not mediated by MW proportion, whereas here motivation only affected mean RRTv indirectly to the extent that low motivation promoted higher MW proportion. While there was no direct effect of motivation on RRTv, counter to H2a, this indirect effect through the H2b-predicted pathway warrants some discussion. Path A shows that motivation predicts MW proportion, which replicates the titular finding from Seli et al. (2017b) that, at the trait level, more motivated participants mind-wander less. Path B in turn shows that MW proportion predicts mean RRTv such that, at the trait level, participants who mind-wander less have lower mean RRTv (i.e., better performance). While at first this finding appears to support an indirect-only mediation (Zhao, Lynch, & Chen, 2010), we suggest that this relationship has not been clearly established due to the method of measurement and conceptually relevant alternative explanations for these findings. Given that motivation is retrospectively measured only after the MRT, the temporal order does not lend itself to mediation as an interpretation. It is possible that participants are using their MW proportion during the task to inform how motivated they retrospectively believe themselves to have been; this would explain the strong relationship between MW proportion and motivation. Path B, on the other hand, could be explained by the consistently replicable findings that, at the probe-level, MW is related to higher RRTvs. That is, the between-participants relation between MW proportion and mean RRTv constitutes a noisier measure of the within-participants relationship between MW and RRTv.

Previous results taken together with the results of the present work suggest that how motivation affects MRT performance remains unclear. All the results support that motivation, MW proportion, and RRTv as an index of performance are related, but assessing a potential mediation with a between-participants measure of retrospective motivation limits our ability to clarify the relationship between these variables. We speculate that motivation may wax and wane measurably within participants over the course of the experiment. Motivation could be measured at each probe, then modelled as a within-participant mediator with more power and less noise. The present study may have been slightly underpowered to find the mediation (see OSM "Power Analysis") and the sample sizes needed to find such small mediations at the between-participants level further supports the need for within-participants methods. Motivation could still be conflated with an internal sense of performance – i.e., if the participant was MW, they may judge themselves as less motivated. To overcome this conflation, experimenters could manipulate motivation within the MRT, for example by offering performance-based financial or other incentives for certain blocks of the MRT while asking participants about their motivation level once they learn the type of block (rewarded vs. not) they are about to begin. This procedure would extend the innovation of Seli et al. (2017b) to a within-participants measure of condition with more power to detect mediation.

Confidence effects found in Seli et al. (2015b) were not conceptually replicated: confidence did not moderate the relation between Report Type and RRTv. There was a main effect such that higher confidence predicted higher RRTvs, which does not support the hypothesis. This failure to replicate is in line with another recent failure to replicate (Meier, 2018) in which confidence was not a significant moderator of Report Type on RRTv when using regression. While these results do not support the further use of confidence measures in the MRT, our exploratory results suggest an alternative: MW depth. By including options to endorse "Completely" or "Mostly" ON/MW states, of which participants made full use (Fig. 4), we can better investigate the relation between RRTv and the depth of MW. In fact, Seli et al. (2015b) proposed that a "depth" measure of MW could attenuate the predictive power of confidence ratings, which agrees with the present findings, though Meier's (2018) direct replication attempt suggests the original confidence effect may have been spurious, which also aligns with the present findings. While depth and confidence are distinct constructs, they are non-orthogonal as "Mostly" reports were associated with higher confidence than "Completely" reports. Furthermore, when considering these depth options, the RRTv order we originally hypothesized for confidence was followed (Fig. 5): completely ON reports showed the lowest RRTv with RRTvs monotonically increasing as depth of MW increased. These depth findings accord with previous MRT work using a 5-point depth scale (Seli et al., 2014) and recent work on a visual version of the MRT (Laflamme, Seli, & Smilek, 2018). Depth and confidence may also interact with meta-awareness as participants show insight into their performance given that an exploratory correlation revealed that participants were able to accurately estimate their performance relative to their peers. Taken together, these findings support the use of the MRT as a tool for measuring the depth of MW as behavioural variability is reflective of phenomenological report. Still, these behavioural findings on MW depth, while numerically suggestive, are currently not adequate to statistically distinguish adjacent depth reports. As such, they may benefit from research designed to increase model signal-to-noise ratio by adding relevant regressors, physiological predictors (e.g., pupillometry, respiration rate (Melnychuk et al., 2018)), or neuroimaging predictors, which may provide more fine-grained information from which to infer MW depth.

By measuring depth and intentionality simultaneously, the present study unintuitively casts intentionality on a continuum, which deserves further elaboration. For participants, MWu was defined as having thoughts drift away from the task

despite an intention to focus on the task; MWi meant having decided to think about unrelated things. Under these definitions, the continuum of intentionality arguably indicates how much participants presently prioritise the goal of task-performance. Being completely on-task implies prioritizing task-performance and devoting sufficient resources to succeed in that attempt (ON). As the MRT continues, boredom increases and task-performance declines in priority: participants retain enough momentum to remain partially engaged but devote inadequate resources to sustain high performance (MWu). As participants devote fewer and fewer resources to MRT performance and the priority of this goal diminishes, their intention shifts toward other higher-priority goals, for example reducing boredom by increasingly stimulation or mentally searching for more valuable ways to spend their time (MWi). Asking participants to report the degree to which their MW is intentional could amount to asking, "How highly are you prioritising the goal of 'on-task performance' right now?" Performance decrements were replicated in the present study and are well documented in the wider literature (Warm et al., 2008), and a broader construal of intention on a goal prioritization continuum also has a plausible neurological foundation (Pezzulo, Rigoli, & Friston, 2018). Compelling cases arguing for the value of investigating MW intentionality have been put forward (Seli et al. 2016, 2017a), and some evidence suggests these states may be subsumed by distinct EEG patterns (Martel, Arvaneh, Robertson, Smallwood, & Dockree, 2019). We agree that intentionality is worth further study, though we propose that non-dichotomous measures of intentionality – and perhaps also meta-awareness – are needed to more completely understand MW subtleties.

Pre-registered control analyses supported the hypothesis (C1) that musical instrument and rhythm-game experience were important predictors of MRT performance. Indeed, timing experience contributed the largest effect size when modelling RRTv. As the MRT involves tapping along to a steady beat, these other experiences keeping rhythm need to be controlled in future uses of the MRT. C2 was also supported as ARCES scores were significantly predictive of RRTv, showing further convergent validity between self-reported attentional lapses and the MRT as a behavioural measure of sustained attention. Future research could investigate meditation experience as another control variable of interest, which could also plausibly mitigate the emotionally negative response to the MRT.

Our extensions regarding boredom serve as further convergent evidence that the RRTv indexes states associated with MW. Supporting E1 and E2, participants who endorsed greater boredom and fatigue showed greater RRTv. We predicted that participants would find the MRT more boring than expected (B1) but less frustrating than expected (B2) and otherwise not change their views (B3). B1 was supported and, counter to B2/3, participants found every aspect of the MRT worse than anticipated, other than effort. These results suggest the MRT does a potent job of inducing boredom and its attendant negative emotional appraisals in a task where participants exert little effort.

These appraisals provide insights into how the MRT fits among other tasks that have been used to investigate MW (e.g., response inhibition tasks). Of the five extension appraisals (boredom, frustration, distress, fatigue, effort), boredom was the highest at the end of the task and also increased most relative to the beginning of the task (see Supplemental Table 1 (OSM)), suggesting boredom is the primary affective state associated with the MRT. Boredom increases during sustained attention tasks and is related to a decline in performance (Warm et al., 2008). The MRT also increased participant frustration, fatigue, and distress, further substantiating the link between sustained attention and negative affect. The MRT paradigm reveals the same increase in negative affect and decrement in performance shown in other paradigms, which helps indicate that the MRT requires cognitive and affective resources similar to other sustained attention tasks.

Most work on sustained attention has used the SART. As both the SART and MRT induce the same negative affect and performance decrements, the MRT may provide a useful alternative in various conditions, for example when concerned about speed-accuracy trade-offs or in visually impaired participants. Future research could further investigate similarities in emotional response to the SART and MRT by measuring the effect of affective states on performance in the MRT. Experiential avoidance, including attempting to alter or avoid negative thoughts and emotions, significantly predict SART performance and can fully mediate the relationship between trait mindfulness and SART performance (Petranker, 2018). These findings suggest that the cognitive resources required to perform well on the SART interact with affective resources used in emotion regulation, i.e., overcoming the discomfort of boredom and its attendant drop in performance. Future research should examine whether experiential avoidance is predictive of performance on the MRT. Such a result would further support that the MRT could provide a viable alternative to the SART and could provide insight into the affective components of cognitive processes.

Constraints on generality

Part of the utility of science is the ability to generalize beyond the immediate sample to other untested samples. Given Meier's (2018) direct replication, we believe this non-direct conceptual replication provides additional insight into the MRT paradigm. While some facets of a paradigm are expected to be central to their validity and replicability, numerous idiosyncratic facets of particular studies should not theoretically affect the results of those

studies; as such, we have prepared this Constraints on Generality (COG) statement in accordance with Simons, Shoda, and Lindsay (2017). For example, while the original Motivation MRT study took place on Acer desktop computers, there is no theoretical reason why the results should depend on the use of this particular brand of computer, so replications run using other computers (e.g., the Lenovo computers used in the present study) are expected to replicate the results. In contrast, the participants in these studies were undergraduates, but given the extant literature on MW in older adults, different findings may result from running the MRT in elderly populations. Our conceptual replication has a number of alterations that allow us to consider theoretical implications for the generality of the MRT paradigm in a wider context.

First, the original study used a 1- to 7-point Likert scale to measure Motivation following the MRT. Our study failed to replicate this finding using a 1–100 visual-analogue scale. There is no theoretical reason for this difference in measurement scale to reflect fundamental differences in participant motivation or the relation of motivation to performance, thus it seems unlikely that this was the primary reason this finding was not replicated. Indeed, if replication of the original effect hinges on using a 1–7 Likert scale, this would speak to a profound methodological fragility. Instead, we find it more plausible that other factors discussed above (inevitability of MW, moderation by incentives, trial-wise waxing and waning motivation) are more likely to influence generality. Participants were rewarded with course credit, so we predict that other incentives (e.g., financial) could result in direct effects of motivation and performance, especially if incentives are tied to performance (Seli et al. 2017b).

The attrition rate in this study implies a constraint on what can be understood regarding sustained attention using the MRT. Most participants understood the experiment and several provided feedback that the instructions were very clear. Even so, numerous (n = 22: UTM n=3, UTSC n=7, York n=10, unknown n=2) participants failed to maintain adequate performance (see OSM "Attrition Rate Commentary"), even with warnings when omissions accrued, and some (n = 6) participants did not understand or reported that they gave up. Some attrition is to be expected in any study, but care needs to be taken that participants understand the MRT and follow its instructions. To fully model the range of attention, future research should consider more deeply the sample of participants for whom performance is inadequate. Indeed, as William James (1996) said, "No account of the universe in its totality can be final which leaves these other forms of consciousness quite disregarded," and by analogy no total account of attention can be final which leaves participant-attrition quite disregarded.

One notable difference is that this conceptual replication adjusted the wording of probes to include "Completely" and "Mostly". This modification likely interacted with ratings of Confidence as the degree to which participants endorsed the universal (e.g., "Completely On-Task") versus the qualified option (e.g., "Mostly On-Task") differed in mean confidence ratings with higher confidence for qualified options. This modification means our results on confidence are a conceptual replication, not a direct replication. Nevertheless, our findings are supported by the same failure to replicate in a direct replication (Meier, 2018). As participants made full use of the extended options we suggest that there may be many valid ways to measure MW with self-report and that the particular questions used will illuminate specific facets of attention and thus should be chosen with care (Weinstein, 2017).

In our attempt to conceptually replicate three studies with non-identical methods, we necessarily had to make choices about which particular facets to include (see OSM "Replication Summary"). In particular, we chose to divide MW into intentional and unintentional MW, as was done in Seli et al. (2015a). This is in contrast with Seli et al. (2013a), which divided MW into MW with awareness ("tuned out") and MW without awareness ("zoned out"), and Seli et al. (2015b), which did not subdivide MW. By focusing on MWu and MWi, we are not able to comment on what role meta-awareness plays in the MRT. Instead, we were able to add additional insight concerning the performance decrement and how it appears to interact with MWu over time. Block was also a significant predictor in previous MRT studies that included it in analyses and the performance decrement is well described in the literature; we suggest that some metric of time be included in all MRT analyses in the future.

The present study uses the same trial durations as in the original MRT. Adjusting the timing of the metronome may affect overall performance and rates of MW, but we predict that the consistent finding that RRTv is lower in ON and higher in MW should replicate even with small adjustments to the MRT timing; extremely long or short durations would likely result in noisier measurements, however. We have no reason to believe that findings depend on any other characteristics of the participants, materials, or context.

Conclusion

The cycle of sustained attention and its dissolution into MW likely involves the nuanced allocation of cognitive and neural resources (Christoff, Irving, Fox, Spreng, & Andrews-Hanna, 2016; Fortenbaugh et al., 2017; Hasenkamp, Wilson-Mendenhall, Duncan, & Barsalou, 2012). The present study suggests that the MRT can be a useful tool in the study of attention as response-variability was robustly linked with phenomenological reports in a structured and replicable manner. Participant feedback suggests the MRT is easy to understand

and requires minimal effort, but future research should consider more deeply why some participants fail to attain adequate performance. What role motivation plays in MRT performance is still not well understood as present and previous findings do not converge (Seli et al. 2015a, 2017b). Confidence interactions were not replicated, consistent with another failure to replicate (Meier, 2018). Covariates support convergent validity as the MRT does induce and measure its intended constructs. We agree with Laflamme et al. (2018) that the MRT paradigm is an especially direct measure of behavioural variability: as a continuous performance task the MRT may offer insight into MW on a finer time-scale compared to response-inhibition tasks. These results answer a call for better measures of attention and MW (Fortenbaugh et al., 2017), and we suggest that the MRT could be used profitably with neuroimaging methods as every moment in the task returns meaningful data. Replication studies like this one provide the necessary independent verification of previous work, allowing future researchers to use tools with confidence. By building on the ideas and paradigm of the MRT, we will be better able to model the complete cycle of sustained attention.

Acknowledgements We would like to thank Paul Seli and Jérôme Sackur for their valuable comments on a previous draft of the manuscript.

Data availability statement The preregistration is available at https://osf.io/5mbda/ and the materials, data, and analysis scripts are available at https://osf.io/m6gtw/.

References

Anderson, T., Lin, H., & Petranker, R. (2017). *Preregistration of Replication of Seli et al 2013 "Wandering Minds and Wavering Rhythms."* https://osf.io/5mbda

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

Cheyne, J. A., Carriere, J. S. A., & Smilek, D. (2006). Absent-mindedness: Lapses of conscious awareness and everyday cognitive failures. *Consciousness and Cognition*, *15*(3), 578–592. https://doi.org/10.1016/j.concog.2005.11.009

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

Christoff, K., Irving, Z. C., Fox, K. C. R., Spreng, R. N., & Andrews-Hanna, J. R. (2016). Mind-wandering as spontaneous thought: A dynamic framework. *Nature Reviews Neuroscience*, *17*(11), 718–731. https://doi.org/10.1038/nrn.2016.113

Davies, D. R., & Parasuraman, R. (1982). *The psychology of vigilance*. Academic Press.

Eastwood, J. D., Frischen, A., Fenske, M. J., & Smilek, D. (2012). The Unengaged Mind: Defining Boredom in Terms of Attention. *Perspectives on Psychological Science*, *7*(5), 482–495.

Fortenbaugh, F. C., DeGutis, J., & Esterman, M. (2017). Recent theoretical, neural, and clinical advances in sustained attention research. *Annals of the New York Academy of Sciences*, *1396*(1), 70–91. https://doi.org/10.1111/nyas.13318

Franklin, M. S., Mooneyham, BenjaminW., Baird, B., & Schooler, JonathanW. (2014). Thinking one thing, saying another: The behavioral correlates of mind-wandering while reading aloud. *Psychonomic Bulletin & Review*, *21*(1), 205–210. https://doi.org/10.3758/s13423-013-0468-2

Galinsky, T. L., Rosa, R. R., Warm, J. S., & Dember, W. N. (1993). Psychophysical determinants of stress in sustained attention. *Human Factors*, *35*(4), 603–614. https://doi.org/10.1177/001872089303500402

Hasenkamp, W., Wilson-Mendenhall, C. D., Duncan, E., & Barsalou, L. W. (2012). Mind wandering and attention during focused meditation: A fine-grained temporal analysis of fluctuating cognitive states. *NeuroImage*, *59*(1), 750–760. https://doi.org/10.1016/j.neuroimage.2011.07.008

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous Inference in General Parametric Models. *Biometrical Journal*, *50*(3), 346–363. https://doi.org/10.1002/bimj.200810425

Hothorn, T., Bretz, F., Westfall, P., Heiberger, R. M., Schuetzenmeister, A., & Scheibe, S. (2019). *multcomp: Simultaneous Inference in General Parametric Models* (1.4-10) [Computer software]. https://CRAN.R-project.org/package=multcomp

Ioannidis, J. P. A. (2012). Why Science Is Not Necessarily Self-Correcting. *Perspectives on Psychological Science*, *7*(6), 645–654. https://doi.org/10.1177/1745691612464056

James, W. (1996). *The Varieties of Religious Experience: A Study in Human Nature*. https://www.gutenberg.org/ebooks/621

Kawashima, I., & Kumano, H. (2017). Prediction of Mind-Wandering with Electroencephalogram and Non-linear Regression Modeling. *Frontiers in Human Neuroscience*, *11*. https://doi.org/10.3389/fnhum.2017.00365

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

Laflamme, P., Seli, P., & Smilek, D. (2018). Validating a visual version of the metronome response task. *Behavior Research Methods*, 1–12. https://doi.org/10.3758/s13428-018-1020-0

Lin, H. (2019). *hausekeep: Miscellaneous functions for research and housekeeping* (0.0.0.9001) [Computer software]. https://hauselin.github.io/hausekeep/. https://doi.org/10.5281/zendo.2557034

Mackworth, J. F. (1964). Performance decrement in vigilance, threshold, and high-speed perceptual motor tasks. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, *18*(3), 209–223. http://dx.doi.org.myaccess.library.utoronto.ca/10.1037/h0083302

Martel, A., Arvaneh, M., Robertson, I., Smallwood, J., & Dockree, P. (2019). Distinct neural markers for intentional and unintentional task unrelated thought. *BioRxiv*, 705061. https://doi.org/10.1101/705061

Meier, M. E. (2018). Can research participants comment authoritatively on the validity of their self-reports of mind wandering and task engagement? A replication and extension of Seli, Jonker, Cheyne, Cortes, and Smilek (2015). *Journal of Experimental Psychology. Human Perception and Performance* https://doi.org/10.1037/xhp0000556

Melnychuk, M. C., Dockree, P. M., O'Connell, R. G., Murphy, P. R., Balsters, J. H., & Robertson, I. H. (2018). Coupling of respiration and attention via the locus coeruleus: Effects of meditation and pranayama. *Psychophysiology*, *55*(9), e13091. https://doi.org/10.1111/psyp.13091

Mooneyham, B. W., & Schooler, J. W. (2013). The costs and benefits of mind-wandering: A review. *Canadian Journal of Experimental Psychology = Revue Canadienne De Psychologie Experimentale*, *67*(1), 11–18. https://doi.org/10.1037/a0031569

Mrazek, M. D., Smallwood, J., Franklin, M. S., Chin, J. M., Baird, B., & Schooler, J. W. (2012). The role of mind-wandering in measurements of general aptitude. *Journal of Experimental Psychology:*

General, *141*(4), 788–798. Scopus. https://doi.org/10.1037/a0027968

Petranker, R. (2018). *Sitting with It: Examining the Relationship Between Mindfulness, Sustained Attention, and Boredom*. https://yorkspace.library.yorku.ca/xmlui/handle/10315/35545

Pezzulo, G., Rigoli, F., & Friston, K. J. (2018). Hierarchical Active Inference: A Theory of Motivated Control. *Trends in Cognitive Sciences*, *22*(4), 294–306. https://doi.org/10.1016/j.tics.2018.01.009

Pinheiro, J. C., & Bates, D. (2009). *Mixed-Effects Models in S and S-PLUS*. Springer Science & Business Media.

Pinheiro, J. C., Bates, D., DebRoy, S., Sarkar, D., & R Core Team. (2018). *nlme: Linear and Nonlinear Mixed Effects Models* (3.1-137) [Computer software]. https://CRAN.R-project.org/package=nlme

R Core Team. (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. http://www.r-project.org/

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

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1–36.

Scerbo, M. W. (1998). Sources of Stress and Boredom in Vigilance. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *42*(10), 764–768. https://doi.org/10.1177/154193129804201024

Schooler, J. W., Mrazek, M. D., Franklin, M. S., Baird, B., Mooneyham, B. W., Zedelius, C., & Broadway, J. M. (2014). Chapter One - The Middle Way: Finding the Balance between Mindfulness and Mind-Wandering. In B. H. Ross (Ed.), *Psychology of Learning and Motivation* (Vol. 60, pp. 1–33). Academic Press. https://doi.org/10.1016/B978-0-12-800090-8.00001-9

Seli, P., Carriere, J. S. A., Thomson, D. R., Cheyne, J. A., Martens, K. A. E., & Smilek, D. (2014). Restless mind, restless body. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *40*(3), 660–668. https://doi.org/10.1037/a0035260

Seli, P., Cheyne, J. A., & Smilek, D. (2013a). Wandering minds and wavering rhythms: Linking mind wandering and behavioral variability. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(1), 1–5. https://doi.org/10.1037/a0030954

Seli, P., Cheyne, J. A., Xu, M., Purdon, C., & Smilek, D. (2015a). Motivation, intentionality, and mind wandering: Implications for assessments of task-unrelated thought. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(5), 1417–1425. https://doi.org/10.1037/xlm0000116

Seli, P., Jonker, T. R., Cheyne, J. A., Cortes, K., & Smilek, D. (2015b). Can research participants comment authoritatively on the validity of their self-reports of mind wandering and task engagement? *Journal of Experimental Psychology. Human Perception and Performance*, *41*(3), 703–709. https://doi.org/10.1037/xhp0000029

Seli, P., Jonker, T. R., Cheyne, J. A., & Smilek, D. (2013b). Enhancing SART Validity by Statistically Controlling Speed-Accuracy Trade-Offs. *Frontiers in Psychology*, *4*, 265. https://doi.org/10.3389/fpsyg.2013.00265

Seli, P., Jonker, T. R., Solman, G. J. F., Cheyne, J. A., & Smilek, D. (2013c). A methodological note on evaluating performance in a sustained-attention-to-response task. *Behavior Research Methods*, *45*(2), 355–363. https://doi.org/10.3758/s13428-012-0266-1

Seli, P., Ralph, B. C. W., Risko, E. F., Schooler, J. W., Schacter, D. L., & Smilek, D. (2017a). Intentionality and meta-awareness of mind wandering: Are they one and the same, or distinct dimensions? *Psychonomic Bulletin & Review*, *24*(6), 1808–1818. https://doi.org/10.3758/s13423-017-1249-0

Seli, P., Risko, E. F., Smilek, D., & Schacter, D. L. (2016). Mind-Wandering With and Without Intention. *Trends in Cognitive Sciences*, *20*(8), 605–617. https://doi.org/10.1016/j.tics.2016.05.010

Seli, P., Schacter, D. L., Risko, E. F., & Smilek, D. (2017b). Increasing participant motivation reduces rates of intentional and unintentional mind wandering. *Psychological Research*, 1–13. https://doi.org/10.1007/s00426-017-0914-2

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on Generality (COG): A Proposed Addition to All Empirical Papers. *Perspectives on Psychological Science*, *12*(6), 1123–1128. https://doi.org/10.1177/1745691617708630

Smallwood, J., Beach, E., Schooler, J. W., & Handy, T. C. (2007). Going AWOL in the Brain: Mind Wandering Reduces Cortical Analysis of External Events. *Journal of Cognitive Neuroscience*, *20*(3), 458–469. https://doi.org/10.1162/jocn.2008.20037

Spunt, R. P., Lieberman, M. D., Cohen, J. R., & Eisenberger, N. I. (2012). The Phenomenology of Error Processing: The Dorsal ACC Response to Stop-signal Errors Tracks Reports of Negative Affect. *Journal of Cognitive Neuroscience*, *24*(8), 1753–1765. https://doi.org/10.1162/jocn_a_00242

Szalma, J. L., Warm, J. S., Matthews, G., Dember, W. N., Weiler, E. M., Meier, A., & Eggemeier, F. T. (2004). Effects of Sensory Modality and Task Duration on Performance, Workload, and Stress in Sustained Attention. *Human Factors*, *46*(2), 219–233. https://doi.org/10.1518/hfes.46.2.219.37334

UCLA Statistical Consulting Group. (2018). *What happens if you omit the main effect in a regression model with an interaction?* https://stats.idre.ucla.edu/stata/faq/what-happens-if-you-omit-the-main-effect-in-a-regression-model-with-an-interaction/

Warm, J. S., Parasuraman, R., & Matthews, G. (2008). Vigilance Requires Hard Mental Work and Is Stressful. *Human Factors*, *50*(3), 433–441. https://doi.org/10.1518/001872008X312152

Weinstein, Y. (2017). Mind-wandering, how do I measure thee with probes? Let me count the ways. *Behavior Research Methods*, 1–20. https://doi.org/10.3758/s13428-017-0891-9

Yanko, M. R., & Spalek, T. M. (2014). Driving with the wandering mind: The effect that mind-wandering has on driving performance. *Human Factors*, *56*(2), 260–269. Scopus. https://doi.org/10.1177/0018720813495280

Yarkoni, T. (2019). *The Generalizability Crisis*. 10.31234/osf.io/jqw35

Zhao, X., Lynch, J. G., & Chen, Q. (2010). Reconsidering Baron and Kenny: Myths and Truths about Mediation Analysis. *Journal of Consumer Research*, *37*(2), 197–206. https://doi.org/10.1086/651257