IN PRESS AT JOURNAL OF EXPERIMENTAL PSYCHOLOGY: HUMAN PERCEPTION AND PERFORMANCE

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)

Matt E. Meier

Western Carolina University

Word count: 12,034

Please address correspondence to:

Matt E. Meier

Western Carolina University

Department of Psychology

Room 302I, Killian Bldg.

Western Carolina University

Cullowhee, North Carolina 28723

mmeier@wcu.edu

Abstract

Seli, Jonker, Cheyne, Cortes, and Smilek (2015) found that, through retrospective confidence reports, subjects can distinguish the validity of their mind wandering reports during a sustained attention (“metronome response”) task. In addition, some subjects were better able to make this distinction than others. Here, I sought to replicate both the within and between-subjects’ effects of confidence judgments on thought probe validity. To this end, I executed a preregistered close replication of Seli et al. (2015) and extended this work by administering the metronome response task twice and by measuring potential individual difference markers for which subjects may be better than others at monitoring their thoughts: working memory capacity, conscientiousness, neuroticism, and dispositional mindfulness. With data from 291 subjects, I found only weak evidence for a within-subjects effect of confidence on thought-report validity in the first administration of the metronome response task and weak to non-existent evidence for individual differences in thought monitoring. No evidence was found for individual differences in the ability to provide valid thought reports.

Public significance statement: This study provides evidence that people are generally able to report when they are on- or off-task, are limited in their ability to provide detailed information about these reports, and that people do not differ from one another in their ability to monitor thoughts.

Keywords: mind wandering, cognitive control, individual differences, working memory capacity, metacognition

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)

Studying thoughts through introspective methods has a long and (sometimes) contentious history in psychology (Buhler, 1907; Nisbett & Wilson, 1977; Watson, 1913; Wundt, 1907). Recently, the study of mind wandering has experienced rapid growth as shown by the increasing number of publications found on Google Scholar (Callard, Smallwood, Golchert, & Margulies, 2013; Weinstein, De Lima, & van der Zee, 2017). A popular method of pursuing questions of mind wandering relies on introspective evidence (Weinstein, 2017), more specifically, collecting subjective reports from subjects while they execute a task. Seli, Jonker, Cheyne, Cortes, and Smilek (2015) asked whether some people are better than others at introspecting (i.e., reporting on their thoughts with clarity and certainty) and whether, at the within-subjects level, subjects have access to metacognitive information that their ability to give accurate subjective reports varies within a task. Seli et al. interpreted their results as providing affirmative evidence to both hypotheses. That is, some subjects are better than others at introspecting and subjects have access to information about which subjective reports they make more accurately than others. Here, I continue this exploration of introspection by seeking to replicate Seli et al.'s results and potentially extend their findings by examining the robustness of these reports across two task administrations and testing whether cognitive and non-cognitive individual difference variables predict differences in the ability to monitor and report on current thoughts.

Seli et al.’s 2015 Study

Seli et al. (2015) had subjects complete a metronome response task while answering thought probes that asked subjects to characterize their “mental state just before this screen appeared” as either on-task or mind wandering. Immediately following this thought probe, subjects were asked to indicate how confident they were in the thought-probe response on a 5-point Likert-type scale (1 = Not at all confident; 5 = Extremely confident). The metronome response task presents subjects with a constant series of tones with which subjects are instructed to respond synchronously with tone onsets by pressing the space bar. Previous work by Seli and colleagues (Seli, Carriere, Levene, & Smilek, 2013, Seli et al., 2014, Seli, Cheyne, & Smilek, 2013; Seli, Cheyne, Xu, Purdon, & Smilek, 2015; Seli, Jonker, Cheyne, & Smilek, 2013) has demonstrated that variability of subjects’ responses (i.e., response times) to the tones predicts the endorsement of mind wandering in response to thought probes. More specifically, when subjects respond that they were mind wandering, their response time variability on the preceding five trials is greater than on five trials preceding reports of on-task. The relation between mind wandering and response time variability has also been found with go/no-go tasks (McVay & Kane, 2009, 2012a) and in latent variables formed from attentional control and memory tasks with mind wandering measures (Kane et al., 2016; Unsworth, 2015). Thus, response time variability has assisted in validating subjective reports and appears to be a robust objective indicator of mind wandering propensity and episodes.

Seli et al. (2015) found that subjects who were overall more confident in their thought probe responses showed a greater distinction in response time variability between reports of on-task and mind wandering. In other words, some subjects’ subjective reports were more accurate than others and the confidence reports provided evidence that subjects had metacognitive awareness (meta-awareness) of their report accuracy. Furthermore, when subjects reported low confidence on individual responses, reports of mind wandering and on-task were not distinguished by response time variability. From these findings, Seli et al. concluded that subjects can provide "*reliable* reports of what one might call 'introspective uncertainty'". Thus, it appears that Seli et al.'s findings take the relation between response time variability and subjective mind wandering reports (i.e., response time variability and reports of mind wandering are associated) further than previous research and suggests that not only are subjective thought reports generally valid, but people (some more than others) have access to fine-grained information about this validity. To explain their individual differences findings, Seli et al. suggested that some subjects may differ in their ability to be meta-aware of their current thoughts, but they did not speculate further about what may drive these individual differences.

Theoretical and Practical Importance of Seli et al. (2015)

I found these results worthy of more scrutiny because individual differences in thought monitoring ability have implications for theories of cognitive control and daily life. Influential theories of cognitive control like Botvinick et al.'s conflict monitoring theory (2001) have proposed that people monitor their environment for conflict, when conflict is detected, cognitive control is engaged. Conflict is usually created by task conditions and is noticed because performance suffers. Here, I consider whether conflict should be thought of more generally. That is, I ask if conflict monitoring happens consciously and whether a conflict signal is generated when one finds their thoughts straying from the (experimenter-assigned) primary task to task-unrelated thoughts.

Practically, a lack of meta-awareness has been suggested to play a causal role in the negative effects of mind wandering on performance with subjects who score higher on Attention Deficit Hyperactivity Disorder (ADHD) measures (Franklin et al., 2017). More specifically, meta-awareness (assessed by asking subjects how aware they were of their mind wandering following thought probes in reading task and in daily life) partially mediated the relation between ADHD and the negative effects of mind wandering in daily life suggesting that the ability to recognize mind wandering is critical for eventual course correction that if not instantiated can lead to the often-found negative relation between mind wandering and task performance (McVay & Kane, 2009, 2012a, 2012b; Mrazek et al., 2013; Seli, 2016; Unsworth & McMillan, 2013; Yanko & Spalek, 2014). The rationale in Franklin et al. (2017) is that greater awareness of mind wandering produces greater cognitive control and thus less adverse effects from mind wandering. Because of the potential theoretical and practical importance of the impact of individual differences in thought monitoring, I thought it prudent to seek out potential markers of this ability.

Individual Difference Markers

Seli et al.'s individual difference finding and Franklin et al.'s ADHD-related result has the potential to explain why subjects who have higher working memory capacity mind wander less in contexts where they are trying to concentrate than subjects with lower working memory capacity (Kane et al., 2007; Kane et al., 2017). That is, the process through which higher working memory capacity subjects achieve their superior performance may be thought monitoring. The control failure × concerns model of mind wandering (McVay & Kane, 2010; Kane & McVay, 2012) maintains that variation in the propensity to mind wander is determined jointly by cognitive abilities and concerns. Currently, the processes enacted by these cognitive abilities are underspecified. Recently, Adam and Vogel (2017) suggested metacognitive monitoring may drive individual differences in visual working memory capacity. They proposed that all people may have the same “true” capacity but because some people are better than others in noticing when their attention wanes from the task at hand and correcting for this, these people exhibit a higher functional visual working memory capacity.

Unsworth and Robinson (2017) supply a schematic of how working memory capacity related differences in meta-awareness may work out at the level of large-scale brain networks. In discussing the nature of working memory capacity and attentional control, they focus on the interplay of three brain networks: the fronto-parietal network, the default mode network, and the salience network. The default mode network is associated with self-generated thoughts (Andrews-Hanna, Smallwood, & Spreng, 2014; Raichle et al., 2001), the salience network with detecting these thoughts (Ham, Leff, de Boissezon, Joffe, & Sharpe, 2013) and the fronto-parietal network with initiating control to dampen down activity in the default mode network (Cai et al., 2015; Zanto & Gazzaley, 2013). Superior monitoring would suggest either a more sensitive salience network or fronto-parietal network. Because working memory capacity is often associated with the fronto-parietal network (Coull, Frith, Frackowiak, & Grasby, 1996; Darki & Klingberg, 2014; Gublinatie, van Rinj, & Cohen, 2014), the latter explanation of a more sensitive fronto-parietal network is suggested if working memory capacity predicts variation in thought monitoring. If thought monitoring is a stable individual difference but working memory capacity is not related to this difference, it would reasonable to assert that the salience network is responsible.

Also, suggestive of a working memory capacity and thought-monitoring relation, is that mindfulness training can increase the functional limit of working memory capacity and reduce mind wandering (Chambers, Lo, & Allen, 2008; Mrazek, Franklin, Phillips, Baird, & Schooler, 2013; Zeidan, Johnson, Diamond, David, & Goolkasian, 2010). Mindfulness, in its own right, is a potential individual difference marker of monitoring ability. Moore and Malinowski (2009) claim that people higher in mindfulness have greater cognitive flexibility than people with lower mindfulness (based on superior performance on a Stroop and visual discrimination tasks). Chambers, Lo, and Allen (2008) present evidence that, in addition to enhancing working memory capacity, increased mindfulness may lead to a greater ability to sustain attention. Perhaps more to the point, is how mindfulness is defined. A popular operationalization of the construct (Bishop et al., 2004; cited 4254 times on google scholar on 3/24/2018), portrays a two-component process with the first component being aware of current experiences (e.g., thoughts, feelings, and sensations) and the second focused on how these thoughts and feelings are regarded. If this is an accurate description of mindfulness, then I may expect people who are higher in mindfulness to be more aware of their thoughts than people who are less mindful. Teper, Segal, and Inzlicht (2013) propose a model for how through increased mindfulness, people can better regulate their emotions. Here, rather than test the effectiveness of mindfulness training, I test if variation in dispositional mindfulness predicts differences in the monitoring and reporting of thoughts.

Besides working memory capacity and mindfulness, I tested if the personality variables of neuroticism and conscientiousness predict thought monitoring abilities. People higher in neuroticism mind wander more than people lower in neuroticism in controlled lab studies (Kane et al., in press; Robinson, Gath, & Unsworth, 2016). Like mind wandering, neuroticism has been associated with response time variability (Robinson & Tamir, 2005). The relationship between neuroticism and response time variability may be driven by mind wandering. A recent proposal has suggested that neuroticism is caused by an overabundance of concerns (Perkins, Arnone, Smallwood, & Mobbs, 2015). While this is within the realm of possibility and fits nicely within the control failure × concerns model of mind wandering, it is premature to attribute neuroticism's relation to mind wandering solely to the amount of concerns. Robinson, Gath, and Unsworth (2016) reported that not only did people who scored higher in neuroticism report more mind wandering, they also had lower working memory capacity and worse attentional control performance indicators than people who had lower neuroticism scores. Similarly, Fox, Dutton, Yates, Georgiou, and Mouchlianitis (2015) found that negative thought intrusions, a hallmark of neuroticism, were associated with worse performance on an attentional control task (i.e., a flanker task). These effects may be the product or antecedents of mind wandering. In the current study, I test if the neuroticism and mind wandering relation is associated with thought monitoring (and if thought monitoring is best thought of as a cognitive ability).

Evidence from work with older adults' and mind wandering has suggested that older adults mind wander less than younger adults because they are more conscientious (Jackson & Balota, 2012). That is, because older adults were more serious about the primary task than younger adults, they reported mind wandering less. Previous work has found a positive association between conscientiousness and the propensity for self-control (Tangney, Baumeister, & Boone, 2004). The few reports of mind wandering relations with conscientiousness when considered together produce inconclusive results with Jackson and Balota (2012) finding a significant negative association, Kane et al. (2017) reporting non-significant negative associations in both lab and daily life contexts, and Jackson, Weinstein, and Balota (2013) simply reporting no association. Here, I investigate if higher-conscientiousness people are better able to control their thoughts than lower-conscientiousness people because they are more invested in the task and thus more accurately monitor their thoughts.

The Current Study

The current study had four main goals: 1) To replicate the between-subjects moderating effect of confidence reports on the relation between subjective reports and response time variability where subjects who report higher confidence than other subjects provide more valid thought reports; 2) To replicate the within-subjects' relations among confidence reports, report type, and response time variability where on instances where a subject reports higher confidence in a subjective report those reports are more valid than instances where that subject reports lower confidence; 3) To test the robustness of these effects across two administrations of the metronome response task; 4) To determine if other individual differences markers are associated with the ability to monitor current thoughts or thought probe validity. To this end, I had a large sample of subjects complete two administrations of the metronome response task, two working memory capacity tasks, and a questionnaire assessing conscientiousness, neuroticism, and dispositional mindfulness.

Method

I report how I determined the sample size, all data exclusions, all manipulations, and all measures in the study (Simmons, Nelson, & Simonsohn, 2011). This study was preregistered on September 1, 2016 (https://osf.io/n9yxk/)[1].

Subjects

Three hundred and four undergraduates from Western Carolina University (*M* first-year student SAT scores from 1019 to 1043 for cohorts entering Fall 2014 through Fall 2016) completed the informed consent for this study. I collected demographic data from 298 of these subjects (data from six subjects was lost because of technical errors). Of these 298 subjects, 63% were female. Subjects had a mean age of 19(SD = 2; one student errantly reported an age of 1 in the demographics; that age was not included in these statistics). Of the subjects who gave ethnicity information (9 subjects declined), 79% identified as white, 8% as black, 4% as multiracial, 4% as Asian, and 2% as other. Subjects received partial credit for a course requirement as compensation for their participation. The stopping rule for data collection was the end of the Spring 2016 semester or at the end of a semester when I had at least 300 subjects. This sample size was chosen on the basis that correlations stabilize when approaching 250 subjects (Schönbrodt & Perugini, 2013) thus allowing precise estimates. I stopped collecting data at the end of the Spring 2016 semester.

General Procedure

Subjects volunteered to complete one two-hour session. I administered the tasks in the following order: operation span, metronome response task (1st administration), symmetry span, metronome response task (2nd administration), personality questionnaire, a memory task (for a project unrelated to the one presented here), and a demographic questionnaire. Experimenters read all on-screen instructions aloud while subjects read along silently. All tasks were programmed and administered with E-prime software (Psychology Software Tools, Pittsburg, PA).

Metronome Response Task

This task was provided by the lead author of Seli et al. (2015). Subjects (wearing headphones) pressed the space bar every time they heard a tone. The experimenter instructed the subjects that the tone was presented at a constant rate (it was presented once a second) and that they should press the space bar in sync with the tone. Subjects completed a total of 900 trials, broken up into 18 blocks of 50 trials each, blocks were not apparent to the subject. Within the metronome response task, subjects responded to 18 embedded thought probes that appeared pseudo-randomly within the middle 40 trials of each 50-trial block. Subjects responded to the probe “Which of the following responses best characterizes your mental state JUST BEFORE this screen appeared” by pressing designated keyboard key that they were either on-task or that they were mind wandering. Immediately following the thought probe, subjects indicated on a 5-point Likert scale how confident they were in the accuracy of their thought report (1 = *Not Confident at All,* 5 = *Extremely Confident*).

From this task, I extracted three variables for analyses. Following from Seli et al., to compute response time variability, I created a five-trial moving window of response times, excluding the first five trials and five trials after each thought probe. The moving window increased incrementally, one-trial at a time. The first moving window would consist of response times from trials 5, 6, 7, 8, and 9 and the second moving window would consist of response times from trials 6, 7, 8, 9, and 10. The first five trials of the task and the five trials following probes were not included in these windows. Finally, I computed the variance for each moving window. In addition to response time variability, I used thought probe responses and the confidence ratings that followed them.

Working Memory Capacity Complex Span Tasks

I assessed working memory capacity with two complex span tasks. In the complex span tasks, subjects memorized short sequences of items while completing an interleaved processing task. Following trial sequences of unpredictable length, subjects recalled the memorial items in serial order.

Before beginning the scored task, subjects practiced memorizing small sets, practiced the processing task alone, and then practiced both task components together. From the processing-only practice a response deadline was determined. If on any processing-task portion of a trial, a response was not made within 2.5 standard deviations of the processing-only practice RT mean, the program skipped the subsequent memory stimulus and the trial was counted as a processing error. I instructed all subjects that I could not use their data if they did not achieve 85% accuracy on the processing portion of the task (see Meier et al. [2018] for a rationale for this criterion).

Operation Span. Subjects memorized sequences of 3–7 letters (letters were presented for 1 second), each presented in alternation with an arithmetic equation to verify [e.g., *(3 × 2) – 1 = 4*; half were true]. At recall, all 12 letters (used in the task but not in the specific trial) appeared in a grid; subjects recalled each letter by selecting it with a computer mouse. Each set length of 3–7 occurred three times in a random order for each subject. The variable used in analyses was the total number of letters recalled in correct serial position (of 75).

Symmetry Span. Subjects memorized sequences of 2–5 red squares appearing within a matrix. Each red square appeared in alternation (presented for 650 ms) with a black-and-white pattern made from an 8 × 8 grid to verify if it was vertically symmetrical (half were symmetrical). At recall, subjects saw an empty 4 × 4 matrix and mouse-clicked the red square locations. Each set length of 2–5 occurred three times in a random order for each subject. Each subject’s score was the total number of red-square locations recalled in correct serial position (of 42).

Personality Measures

I created a computerized questionnaire to measure the constructs of conscientiousness, neuroticism, and dispositional mindfulness. In this questionnaire, I included four infrequency items (e.g., “The word “statistics” has more letters than the word “math”). I dropped personality data from subjects who endorsed more than one infrequency item. Conscientiousness, neuroticism, dispositional mindfulness, and infrequency items were presented in a random order for all subjects. Subjects responded on a 5-point Likert scale (1 = *never true*, 5 = *always true*). Scores for all three constructs were computed by averaging Likert-scale scores (after reverse scoring had been done).

Conscientiousness and Neuroticism. For each construct, subjects completed 12 items from the Neuroticism-Extraversion-Openness Five Factor Inventory (NEO-FFI-3; McCrae & Costa, 2010).

Dispositional Mindfulness. Subjects completed the 24-item Five Facet Mindfulness Questionnaire – Short Form (FFMQ-SH; Bohlmeijer, ten Klooster, Fledderus, Veehof, & Baer, 2011). The 24 items load onto five facets: acting with awareness, nonjudging of internal experience, nonreactivity to internal experience, and describing and observing. Here, I performed analyses using both the composite overall measure and the five factors.

Data Analysis

My primary approach was to follow the analytic pipeline used by Seli et al. (2015). I augmented this pipeline by using Linear Mixed Models (LMMs) and Bayes Factors (BFs) where applicable. In places where I diverge from the preregistered analysis plan, the analyses should be considered exploratory. Unless otherwise specified, analyses were conducted in the R system for statistical analysis (R Core Team, 2017). LMMs were carried out using the lme4 package (Bates, Maechler, Bolker, & Walker, 2015). In the LMMs, predictors were centered on the grand mean of the sample and all models were random intercept only. P-values in the LMMs were computed using the Satterthwaite approximation contained in the lmerTest package (Kuznetsova, Brockhoff, & Christensen, 2016). The Satterthwaite approximation has been shown to produce *p* values in line with actual false positive rates (Luke, 2016). BFs for correlations used code provided in an online appendix to Wagenmakers, Verhagen, and Ly (2015; https://osf.io/cabmf/), and BFs for general linear models and ANOVAs were conducted with the BayesFactor package (Morey & Rouder, 2015). BFs compare the predictive performance of competing models (Etz & Wagnemakers, 2017; Kass & Raferty, 1995). I use BFs to determine which model is more likely given the data.

I followed Seli et al. (2015) in how I computed response time variability. Like Seli et al., response time variability was highly skewed, so following Seli et al., I used the natural log to transform the data (1st administration of metronome response task untransformed skew = 3.56, transformed skew = .67; 2nd administration of the metronome response task untransformed skew = 3.88, transformed skew = .38). Overall mind wandering rate (used in between-subject analyses) was the proportion of thought probes to which the subject endorsed mind wandering. For working memory capacity, I formed a composite measure by converting raw scores to z scores and averaging them. Data, analysis code, and outputs are available at the following link: https://osf.io/n9yxk/.

Data Loss

All data exclusions were made in accord with the preregistration. I dropped all data for six subjects were deemed by experimenters as noncompliant with instructions across tasks. These decisions were made without consulting the subjects' data. Due to computer or experimenter error, I was missing data from two subjects in the operation span task, six subjects in the symmetry span task and personality measures, four subjects in the first metronome task, and seven subjects in the second metronome task. In addition to these exclusions, I excluded 46 subjects in the operation span task and 40 subjects in the symmetry span task for not making the 85% processing criterion, 14 subjects' personality questionnaire data for endorsing more than one infrequency item, and 3 subjects' data from the first metronome response task for responding to less than 10% of the trials. To use a subject for working memory capacity-related analyses, I required them to have scores from both the operation span task and the symmetry span task to form a composite. Using a composite for working memory capacity is theoretically consistent with the notion that working memory capacity related cognitive performance differences are driven largely by domain-general processes (Kane et al., 2004). For all the following analyses, I used the maximum amount of data available (after these exclusions; therefore, Ns differ among analyses). Descriptive statistics for all tasks can be seen in Table 1. Intercorrelations between and internal consistency for measures can be seen in Table 2.

Replication Results

Correlations. Replication analyses are limited to the first administration of the metronome response task. I used two different BFs in analyses of these correlations. The first approach was to test if the data was more likely under the null hypothesis (r = 0) or an (two-sided) alternative hypothesis that r is different than 0. These analyses were structured so that numbers greater than one show that the data favor the alternative hypothesis and a BF of less than one shows the data support the null (values of 1 indicate that the data did not discriminate between models). For example, a BF of 3 in this setup shows that given the data the alternative hypothesis is 75% probable while a BF factor of .33 show that the null hypothesis is 75% probable. The second approach was to compute a replication BF, in these analyses, values less than one demonstrate that the probability of the null hypothesis (i.e., no relation) drops after updating the information from the initial study with the data from the replication study and values greater than one suggest that probability of the null hypothesis increases. For these replication BFs, the correlation from the initial study in the subscript brackets. For example, below I compare the correlation between mind wandering propensity and response time variability in the current study with that found by Seli et al. The subscript for that comparison (e.g., $BF_{0R[\ .31]}$) is the correlation reported by Seli et al.

Consistent with Seli et al. and previous studies (Kane & McVay, 2012; Kane et al., 2016; Seli, Levene, & Smilek, 2013; Unsworth, 2015) mind wandering propensity was positively associated with overall response time variability, r (291) = .18, p = .002, *95% Confidence Interval (CI)* [.07, .29], BF_{I0} = 10.77, $BF_{0R[\ .31]}$ = .02, and response time variability was not related to confidence, r (291) = -.11, p = .07, *95% CI* [-.22, .01], BF_{I0} = .39, $BF_{0R[.09]}$ = 1.53. But unlike Seli et al., confidence was not positively associated with mind wandering propensity, r (291) = .01, p = .82, *95% CI* [-.10, .13], BF_{I0} = .08, $BF_{0R[.27]}$ = 22.32.

Moderation analysis. Like Seli et al., I conducted hierarchical regressions with response time variability being regressed on mind wandering and confidence in the first step, and the interaction of mind wandering and confidence being entered in the second step (see Table 3). If the model with the mind wandering × confidence interaction fits the data better than the model without the interaction (and the interaction parameter is positive), it suggests that people with higher confidence show a greater distinction in response time variability according to their thought probe reports than lower-confidence people. More simply put, a positive interaction between mind wandering and confidence would show that some people (i.e., high-confidence people) are better able to determine if they are on-task or mind wandering than others. Contrary to Seli et al.'s result, the model with the interaction did not fit the data significantly better than the model without the interaction, $\Delta R^2 = .0002$, $F = (1, 287) = 0.06$, $p = .80$. Thus, confidence did not moderate the relation between mind wandering and response time variability. To further enhance my inference, I compared models with and without the mind wandering × confidence interaction within a Bayesian framework. The model without the interaction was preferred over the model with the interaction by a factor of 4.5 (i.e., *BF* = 4.5).

Repeated measures ANOVAs. Mirroring Seli et al., I conducted a set of two within-subject repeated measures ANOVAs. These ANOVAs were used to test if when a subject rates their thought probe with higher confidence if that thought probe is more valid (i.e., shows a greater distinction in response time variability between on-task and mind wandering reports) than probes that are followed with a lower confidence rating. Because ANOVAs require balanced cells and many subjects did not use all confidence ratings, confidence ratings were binned. Confidence reports of 1 and 2 were binned together as low confidence, reports of 3 were medium confidence, and reports of 4 or 5 were binned as high confidence. Seli et al. reported that binning reduced the percentage of missing cells from 41% to 26%. This same procedure reduced the missing cells in the current study from 44% to 28%.

The first ANOVA used data from all subjects who have response time variability data in all six cells (e.g. on-task and high confidence). I had 61 subjects who had data in all six cells (Seli et al. N = 22). Like Seli et al., the main effect of report type did not reach conventional significance, *F*(1, 60) = 2.43, MSE =0.79, *p* = .12, η^2_p= .04, but unlike Seli et al., neither the main effect of confidence, *F*(2, 120) = 1.76, MSE =0.67, *p* = .18, η^2_p = .03, nor the interaction between report type and confidence, *F*(2, 120) = 1.90, MSE =0.71, *p* = .15, η^2_p = .03, did (see Figure 1 Panel A for a visual depiction). In a Bayesian model comparison analysis, the intercept-only model (i.e., the model with subjects as the only predictor) was the most preferred model. The intercept-only model was preferred by a factor of 18 over the main-effects-only model, and by a factor of 59 over the main-effects and interaction model.

The second ANOVA imputed missing data by using SPSS's (IBM Corp., Armonk, NY) linear trend point estimation (this mirrors the approach used by Seli et al., 2015), thus allowing me to use the entire sample. In this ANOVA, report type, *F*(1, 290) = 4.56, MSE = 0.85, *p* = .03, η^2_p = .02, confidence level, *F*(2, 580) = 14.95, MSE = 0.73, *p* < .001, η^2_p = .05, and the report type and confidence level interaction, *F*(2, 580) = 5.31, MSE = 0.62, *p* = .005, η^2_p = .02, were all statistically significant (see Figure 1 Panel B). I decomposed the interaction by conducting dependent t-tests between reports of on-task and mind wandering at the different confidence levels. These t-tests revealed that only for high confidence reports did the mean (log) response time variability significantly differ between report types, *t*(290) = -4.24, *p* = .00003, *d* = .27, (medium confidence, *t*(290) = -1.8, *p* = .07, *d* = .12; low confidence, *t*(290) = .80, *p* = .42, *d* = .07). The main-effects only model (BF = 15781) and the main-effects-plus-interaction model (*BF* = 16377) were both strongly preferred over the intercept-only model. The BF from the comparison of the main-effects only model and main-effects-plus-interaction model was 1 suggesting that the data did not discriminate between these models and both are equally likely given the data.

Linear Mixed Models. Seli et al. (2015) performed their within-subject analyses with two ANOVAs (as I did above), one using data from all subjects who had data in all cells and one where missing data was imputed. Rather than using only a subset of the data or imputing missing data based on patterns present in the existing data, I sought a statistical approach that allowed me to ask the same questions of the data, use all subjects, but only use existing (rather than imputed) data. This resulted in using LMMs. LMMs accommodate unbalanced data without a loss of power and account for the non-independence of data by using subjects as a random variable (Kliegl, Wei, Dambacher, Yan, & Zhou, 2010). In the first LMM, response time variability was predicted from confidence, report type (on-task vs. mind wandering), and their interaction. Confidence was mean centered and entered as a continuous variable. Report type was effect coded, so the parameter value reflects the difference in response time variability between on-task and mind wandering reports. Report type, b = .21, SE = .04, t = 5.9, *p* < .001, confidence, b = -.08, SE = .02, t = -4.4, *p* < .001, and the report type by confidence interaction, b = .12, SE = .04, t = 3.7, *p* < .001, all predicted unique variance in response time variability.

To further explore where on the confidence scale report type was moderated, I next conducted an LMM with confidence coded as a factor with five levels. This model did not converge. Because there were few confidence responses of 1 (2.6%; see Table 4 for confidence report distributions by report type), I binned confidence responses of 1 and 2 together. In this dummy-coded model, the binned confidence level (confidence reports of one and two) was the reference level. At the lowest confidence level report type did not predict response time variability, b = .09, SE = .09, t = 1.0, *p* = .34. Confidence reports of 3 did not significantly differ from the lowest confidence level in predicting response time variability, b = -.10, SE = .06, t = -1.8, *p* = .08, but confidence reports of four and five did, b = -.27, SE = .06, t = 4.7, *p* < .001; b = -

.21, SE = .06, t = -3.5, *p* < .001, respectively, with these reports being associated with less response time variability. The slope of report type was not significantly different at confidence reports of three, b = -.02, SE = .11, t = -0.2, *p* = .87, or four, b = .07, SE = .11, t = 0.6, *p* = .58, from the slope at the lowest confidence level (i.e. the bin of confidence reports of one and two), but the slope at confidence reports of five was, b = .35, SE = .11, t = 3.1, *p* = .002, indicating that the confidence × report type interaction seen in the initial LMM was driven primarily by the highest confidence responses.

To qualify this inference that only the highest confidence responses were responsible for the report type by confidence interaction, I ran separate models with confidence reports of three and four as the reference levels. In these models, I focused on the parameter reflecting the effect of report type on response time variability and the interaction of how this relation changes at the different confidence levels (for comparisons not contained in the previous model). In the model with confidence level three as the reference, report type did not predict response time variability, b = .07, SE = .07, t = 1.0, *p* = .31, this relation did not significantly differ at confidence reports of four, b = -.10, SE = .06, t = -1.8, *p* = .08, but did significantly differ at confidence reports of five, b = .37, SE = .09, t = 4.0, *p* < .001. In the model with confidence level four as the reference, report type did predict response time variability, b = .15, SE = .07, t = 2.2, *p* = .03, and the slope of report type got steeper when a confidence level of five was indicated, b = .29, SE = .09, t = 3.2, *p* < .001.

Replication Discussion

Here, I attempted to replicate the between and within-subjects' effects of confidence ratings on thought probe validity. At the individual differences level, I found no evidence that some subjects were better at monitoring their thoughts than others. Although I found corroborating evidence that overall mind wandering propensity is positively associated with response time variability (Seli, Carriere, Levene, & Smilek, 2013, Seli et al., 2014, Seli, Cheyne, & Smilek, 2013; Seli, Cheyne, Xu, Purdon, & Smilek, 2015; Seli, Jonker, Cheyne, & Smilek, 2013), this effect was not moderated by overall confidence. Within-subject analyses revealed a more mixed pattern. The ANOVA with only complete cases did not contribute confirming evidence that when subjects give a higher confidence rating for their thought probe response, those thought probes have more validity. The ANOVA with imputed data and LMMs using all the subjects did provide evidence that probes associated with higher confidence have more validity. Moreover, the LMMs suggested that probe reports followed by the highest confidence ratings drive this relationship.

Two pieces of evidence suggest that these within-subjects effects are not robust. First, Bayesian model comparison on the ANOVA showed that the data did not distinguish between the model with the interaction and the model without the report type × confidence interaction. That is, although there is some signal from a frequentist perspective (as indicated by the p-value of .005), it is (almost) nonexistent through a Bayesian model-comparison perspective. Second, the effect sizes associated with report type × confidence rating interaction are quite modest (e.g., from the ANOVA with data from subjects with complete data η^2_p = .03; from the ANOVA with imputed data η^2_p = .02; these effect sizes were smaller those by those produced by Seli et al. who had an η^2_p = .20 for the complete cases ANOVA and an η^2_p = .07 for the imputed data ANOVA). The modest effect sizes are also reflected in the fact that in the ANOVA with fewer subjects (i.e., the ANOVA with complete cases), none of the predictors reached conventional statistical significance. Prior work by Seli et al.(2014) does provide evidence that subjects may be able to provide more detailed information about their thoughts if they are asked to respond to thought probes using a five-point Likert scale (1 = completely on task, 2 = mostly on task, 3 = both on the task and thinking about unrelated concerns, 4 = mostly thinking about unrelated concerns, 5 = completely thinking about unrelated concerns). Seli et al. (2014) found a strong association between the thought probe response and response time variability (and fidgeting) with more off task concerns associated with more response time variability (and fidgeting).

Based on the findings of Seli et al. (2015) and the findings presented here, I tentatively conclude that subjects cannot *authoritatively* comment on the validity of thought probes by completing retrospective confidence reports. To be clear, the evidence does suggest that confidence reports of four or five are associated with a greater distinction of response time variability between on-task and mind wandering reports but based on the effect sizes, Bayesian model comparison, and the different pattern of results found by Seli et al. (i.e., a quadratic effect) and those produced here (a more exponential effect) these effects appear relatively weak and noisy. I acknowledge that the current procedure differs from the original by having subjects first complete an operation span task**.** It is plausible that by having subjects complete a moderately demanding attentional task before the metronome task reduced their ability or motivation to monitor their thought in the metronome response task (Brewer et al., 2017), and that without first completing the operation span task these within-subjects effects would be more pronounced.

Although I found no evidence that mean confidence moderates the report type and response time variability relation, in the introduction I gave a rationale for other possible individual difference variables that may predict which subjects will be better able to monitor their thoughts than others. Below, I tested if these markers are effective at predicting thought probe validity and thought monitoring. In addition, I had questions as to what extent thought monitoring is a stable individual difference. To address this, I administered the metronome response task again. By administering the metronome response task again, I could assess test-retest validity on response time variability, overall mind wandering propensity, overall confidence (see Table 2 for intercorrelations). I also created an index of monitoring accuracy and compared this across the two administrations.

Extension Results

Metronome Response Task 2nd Administration Correlations. Here, when I calculated the replication BF, I entered the correlation I found in the first administration (rather than the one from Seli et al.). The question this second BF [0R] is answering is if the correlation obtained in the second administration is if the probability of the null hypothesis increases or decreases when incorporating the information from the second administration. Values less than 1 indicate the null hypothesis is less probable and values greater than 1 indicate the null is more probable (values close to 1 suggest that probability of the null is relatively unaffected by the incorporation of data from the 2nd administration). Mind wandering propensity was again positively associated with overall response time variability, r (291) = .21, p < .001, *95% CI* [.10, .32], BF_{I0} = 41.44, $BF_{0R[.18]}$ = .0026. Confidence was not associated with mind wandering propensity, r (291) = .04, p = .45, *95% CI* [-.07, .16], BF_{I0} = .10, $BF_{0R[.01]}$ = 1.15, and response time variability was not related to confidence, r (291) = -.01, p = .84, *95% CI* [-.13, .10], BF_{I0} = .07, $BF_{0R[-.11]}$ = 2.78. From these results, we can conclude that the data from the 2nd metronome response task are mostly consistent with data from the 1st administration of the metronome response task and the most robust association in this data is between mind wandering propensity and response time variability.

Moderation Analysis. Here, I again followed the approach by Seli et al. and used a hierarchical regression approach and compared a model with mind wandering and confidence as predictors to a model with mind wandering, confidence, and their interaction as predictors. As in the first administration, the fit of the model did not significantly improve with the addition of the interaction, $\Delta R^2 = .000002$, $F = (1, 287) = 0.19$, $p = .67$ (see Table 3 for all parameter estimates). With the Bayesian model comparison, the model without the interaction was preferred by a factor of 4 over the model with the interaction.

Metronome Response Task 2nd Administration Repeated Measures. The binning procedure reduced the percentage of missing cells from 50% to 37%. In the ANOVA with subjects who had data in all six cells (N = 41), when subjects reported mind wandering they had greater response time variability, $F(1, 40) = 9.36$, MSE =0.61, $p = .004$, $\eta^2_p= .19$, but the effect of confidence, $F(2, 80) = 2.71$, MSE =0.67, $p = .07$, $\eta^2_p = .06$, and the interaction between report type and confidence, $F(2, 80) = 1.90$, MSE =0.50, $p = .10$, $\eta^2_p = .06$, did not reach the conventional criterion for statistical significance (see Figure 2 Panel A). In the Bayesian model comparison (with the intercept-only model as the reference level), the model with the intercept and report type was the most preferred (*BF* = 16), followed by the main-effects model (i.e., report type and confidence; *BF* = 8), and then by the main-effects plus interaction model (BF = 4). When focusing on what the confidence by report type interaction adds to the model, the main-effects model was favored by a factor or 2 over the main-effects-plus-interaction model.

When using the same model with imputed data, report type $F(1, 290) = 100.75$, MSE =0.59, $p = < .001$, $\eta^2_p= .26$, and confidence, $F(2, 580) = 2.71$, MSE =5.43, $p = .005$, $\eta^2_p = .02$, both predicted response time variability, but the interaction between report type and confidence, $F(2, 580) = 1.80$, MSE =0.52, $p = .16$, $\eta^2_p = .01$, did not (see Figure 2 Panel B). In comparison to the intercept only model, Bayesian model comparison most favored the main-effects model (*BF* = 8.0e+19), followed by the model with the intercept and report type (*BF* = 3.2e+19) and the full-effects-plus-interaction model (*BF* = 4.86e+18). When comparing the main-effects model to the main-effects-plus-interaction model, the main-effects model was favored by a factor of 17.

Linear Mixed Model. This model predicted response time variability with report type, confidence, and their interaction. Report type, b = .28, SE = .04, t = 6.9, $p < .001$, and confidence, b = -.05, SE = .02, t = -2.8, $p = .006$, predicted response time variability**,** but the report type by confidence interaction did not, b = .03, SE = .04, t = 0.8, $p = .45$.

Working Memory Capacity Moderation Analysis. Like the moderation analyses with confidence ratings performed above, I compared models predicting response time variability. One model had report type and working memory capacity as predictors and the other had report type, working memory capacity, and their interaction (i.e., working memory capacity × report type) as predictors (for parameter estimates from both task administrations see Table 5). The question I attempted to answer with this analysis is: Are people with higher working memory capacity better at providing more valid responses to their thought probes than people with lower working memory capacity? I did this analysis for both administrations of the metronome response task (N for both = 220). In the first administration, the model with the interaction did not fit the data better than the model without $\Delta R^2 = .007$, $F = (1, 216) = 1.63$, $p = .20$. In the Bayesian model comparison, the model without the interaction was favored the data by a factor of 4 over the model with the interaction. In the second administration, the model with the interaction also did not improve the model fit over the model without the interaction $\Delta R^2 = .002$, *F* = (1, 216) = 0.46, *p* = .50. For the second administration, the Bayesian model comparison favored the model without the interaction by a factor of 3.5 over the model with the interaction. These results provide evidence that thought probe validity is not a function of working memory capacity variation.

Working Memory Capacity LMM. To directly assess working memory capacity related differences in thought monitoring (rather than indirectly through probe validity like the immediately preceding analysis), I conducted LMMs on both metronome response task administrations. The key parameter in these models is the three-way interaction among report type, confidence, and the individual difference marker. For example, in this model with working memory capacity, if we see a significant three-way interaction it may indicate that when higher working memory capacity people are more confident, we see a larger distinction between on-task and mind wandering reports in terms of response time variability.

In these LMMs predicting response time variability, I entered report type, confidence, and working memory capacity and their interactions as predictors. Report type was effect coded. Confidence and working memory capacity were mean centered and treated as continuous variables. In the first metronome response task, report type, b = .18, SE = .04, t = 4.3, *p* < .001, confidence, b = -.06, SE = .02, t = -3.0, *p* = .003, working memory capacity, b = -.23, SE = .06, t = -4.1, *p* < .001, and the report type by confidence interaction, b = .15, SE = .04, t = 4.0, *p* < .001, all predicted unique variance in response time variability. The report type × working memory capacity, b = -.04, SE = .05, t = 0.7, *p* = .46, the confidence × working memory capacity, b = .01, SE = .02, t = 0.4, *p* = .67, and the critical report type × confidence × working memory capacity interaction, b = .00, SE = .05, t = 0.1, *p* = .95, did not.

In the second metronome response task, report type, b = .28, SE = .05, t = 5.9, *p* < .001, confidence, b = -.04, SE = .02, t = -2.31, *p* = .03, and working memory capacity, b = -.18, SE = .06, t = -2.93, *p* = .004, again predicted unique variance in response time variability. The two-way interactions of report type and confidence, b = .01, SE = .04, t = 0.2, *p* = .82, report type and working memory capacity, b = -.05, SE = .06, t = -0.8, *p* = .42, and confidence and working memory capacity, b = -.03, SE = .03, t = -1.1, *p* = .27, did not. In this model, the three-way interaction among report type, confidence, and working memory capacity, did reach conventional statistical significance, b = .10, SE = .05, t = 2.0, *p* = .046.

To interpret this interaction, I conducted separate models for mind wandering and on-task reports. In the model examining just reports of on-task, working memory capacity did, b = -.18, SE = .07, t = -2.5, *p* = .01, and confidence did not, b = -.03, SE = .03, t = -0.8, *p* = .42, predict unique variance in response time variability. Most importantly, the confidence × working memory capacity was nonsignificant but had a negative slope, b = -.05, SE = .04, t = -1.3, *p* = .21. In the model of mind wandering reports, working memory capacity was again associated with less response time variability, b = -.17, SE = .07, t = -2.4, *p* = .02. Here, confidence just missed the statistical significance criterion, b = -.10, SE = .03, t = 2.0, *p* = .051, and the confidence × working memory capacity was nonsignificant but positive, b = .03, SE = .04, t = 0.7, *p* = .47. Recall that this analysis was conducted to better understand the three-way interaction among working memory capacity, confidence, and report type. When on-task was reported, the parameter estimate for the working memory capacity × confidence interaction had a negative slope providing evidence that when higher working memory capacity subjects reported with high confidence, they exhibited less response time variability than lower-working memory capacity subjects when they provided high confidence reports. When mind wandering was reported, the parameter estimate for the working memory capacity × confidence interaction was positive suggesting that when higher working memory capacity subjects reported mind wandering with high confidence they showed a greater increase in response time variability than did lower working memory capacity subjects when going from low to high confidence. This result does provide some evidence for working memory capacity related differences in thought monitoring. But I consider this evidence very weak because it is only found in one task administration and p-values that close to .05 have been shown to provide very weak evidential value (Benjamin et al., 2017).

Conscientiousness Moderation Analysis. Adding the conscientiousness × report type interaction did not significantly improve the fit of the model over the model that had conscientiousness and report type as predictors, $\Delta R^2 = .009$, $F = (1, 273) = 2.55$, $p = .11$ (for parameter estimates from both task administrations see Table 6). Here, the BF was not able to distinguish these two models (BF = 1.35 in favor of the model without interaction). The conscientiousness × report type interaction also did not significantly improve the model fit in the 2nd administration of the metronome response task, $\Delta R^2 = .005$, $F = (1, 273) = 1.46$, $p = .23$. The model without the conscientiousness and report type interaction was favored by a factor of 2 (*BF* = 2.26) over the model with the interaction.

Conscientiousness LMM. In the first metronome response task, in a model with report type, confidence, and conscientiousness predicting response time variability, report type, b = .21, SE = .04, t = 5.6, $p < .001$, confidence, b = -.07, SE = .02, t = -3.8, $p < .001$, and the report type × confidence interaction, b = .12, SE = .03, t = 3.6, $p < .001$, were all significantly associated with response time variability. Conscientiousness, b = -.03, SE = .07, t = -0.4, $p = .69$, the report type × conscientiousness interaction, b = -.09, SE = .07, t = -1.3, $p = .18$, the confidence ×

conscientiousness interaction, b = .02, SE = .03, t = 0.6, *p* = .56, nor the three-way interaction among report type, confidence and conscientiousness, b = .01, SE = .06, t = 0.1, *p* = .93, did.

The model with conscientiousness, report type, and confidence from the second administration of the metronome response task found again that report type, b = .29, SE = .04, t = 3.6, *p* < .001, and confidence, b = -.05, SE = .02, t = -2.8, *p* = .004, predicted unique variance in response time variability. But unlike the model on the first administration, the report type × confidence interaction did not, b = .03, SE = .04, t = 0.7, *p* = .48. In addition, in this model the parameter representing the report type × conscientiousness interaction was significantly associated with response time variability, b = -.20, SE = .08, t = -2.5, *p* < .001. In separate models for on-task and mind wandering reports, the parameter estimate for conscientiousness was nonsignificantly positive for reports of on-task, b = .05, SE = .09, t = 0.6, *p* = .58, and nonsignificantly negative for reports of mind wandering, b = -.16, SE = .09, t = -1.7, *p* = .09. These conscientiousness parameters from these two models are the opposite of what one would expect if conscientiousness was associated with more accurate thought monitoring and thus more valid thought probe reporting. The most direct measure of monitoring, the three-way interaction among report type, confidence, and conscientiousness was nonsignificant, b = -.08, SE = .06, t = -1.2, *p* = .23.

Neuroticism Moderation Analysis. Neuroticism did not moderate the relation between report type and response time variability in the first administration, ΔR^2 = .001, *F* = (1, 273) = .33, *p* = .59 (for parameter estimates from both task administrations see Table 7). The model without the interaction was favored by a factor of 4 (*BF* = 3.86) over the model with the interaction. In the second administration of the metronome response task, the neuroticism and report type interaction also did not significantly improve the model fit, ΔR^2 = .008, *F* = (1, 273) =

2.50, p = .12. Bayesian model comparison was not able to distinguish between the models with and without the interaction (BF = 1.4 in favor of the model without the interaction).

Neuroticism LMM. Following a now familiar pattern, in both metronome response tasks administrations report type, 1st administration: b = .21, SE = .04, t = 5.5, p < .001, 2nd administration: b = .28, SE = .04, t = 6.6, p < .001, and confidence, 1st administration: b = -.07, SE = .02, t = -3.9, p < .001, 2nd administration: b = -.05, SE = .02, t = -2.8, p = .004, both predicted unique variance in response time variability, and in only the first administration the report type × confidence interaction did, 1st administration: b = .12, SE = .03, t = 3.5, p < .001, 2nd administration: b = .02, SE = .04, t = 0.7, p = .51. In neither the first administration nor the second administration did any of the other two-way interactions (neuroticism × confidence, 1st administration: b = .02, SE = .03, t = 0.7, p = .50, 2nd administration, b = .01, SE = .03, t = 0.3, p = .78;report type × neuroticism, 1st administration: b = .05, SE = .06, t = 0.9, p = .39,2nd administration: b = .08, SE = .07, t = 1.0, p = .30) or the three-way interaction among report type, confidence, and neuroticism reach statistical significance, 1st administration: b = .00, SE = .06, t = 0.0, p = .98, 2nd administration: b = .03, SE = .06, t = 0.5, p = .62.

Dispositional Mindfulness Moderation Analysis. Dispositional mindfulness also did not moderate the report type and response time variability relation in the first administration, ΔR^2 = .00005, F = (1, 273) = .01, p = .90 (for parameter estimates from both task administrations see Table 8). Here, the model without the interaction was favored by a factor of 4 (BF = 4.41) over the model with the interaction. Dispositional mindfulness again did not moderate the relation in the 2nd administration, ΔR^2 = .006, F = (1, 273) = 1.74, p = .19. The model without the interaction was favored by a factor of 2 over the model with the interaction. In addition to the total dispositional mindfulness score not moderating the report and response time variability relation, neither did any of the five facets of this scale (BFs all in favor of the model without the interaction; Acting with Awareness in the first metronome response task: $\Delta R^2 = .0003$, $F = (1, 273) = .09$, $p = .76$, $BF = 4.2$; Acting with Awareness in the second metronome response task: $\Delta R^2 = .00003$, $F = (1, 273) = .01$, $p = .93$, $BF = 4.4$; Nonjudging of Internal Experience in the first metronome response task: $\Delta R^2 = .000001$, $F = (1, 273) = .0003$, $p = .98$, $BF = 4.5$; Nonjudging of Internal Experience in the second metronome response task: $\Delta R^2 = .0002$, $F = (1, 273) = .07$, $p = .80$, $BF = 4.4$; Nonreactivity to Internal Experience in the first metronome response task, $\Delta R^2 = .004$, $F = (1, 273) = 1.2$, $p = .28$, $BF = 2.6$; Nonreactivity to Internal Experience in the second metronome response task, $\Delta R^2 = .007$, $F = (1, 273) = 2.1$, $p = .15$, $BF = 1.7$; Observation in the first metronome response task administration, $\Delta R^2 = .006$, $F = (1, 273) = 1.7$, $p = .20$, $BF = 2.1$; Observation in the second metronome response task administration, $\Delta R^2 = .0002$, $F = (1, 273) = .05$, $p = .82$, $BF = 1.1$; Description in the first metronome response task administration, $\Delta R^2 = .006$, $F = (1, 273) = 1.7$, $p = .20$, $BF = 2.2$; Description in the second metronome response task administration, $\Delta R^2 = .005$, $F = (1, 273) = 1.6$, $p = .21$, $BF = 2.2$).

Dispositional Mindfulness LMMs. The LMMs for the overall dispositional mindfulness score also provided no evidence for dispositional mindfulness-related individual differences in thought monitoring. In both models, report type, 1st administration: b = .20, SE = .04, t = 5.4, $p < .001$, 2nd administration: b = .29, SE = .04, t = 6.7, $p < .001$, and confidence were associated with unique variance in response time variability. In the first model, the report type × confidence interaction was a significant predictor, b = .12, SE = .03, t = 3.5, $p < .001$, and in the second model it was not, b = .03, SE = .04, t = 0.8, p =.43. The parameter estimate for dispositional mindfulness did not reach statistical significance in either model, 1st administration: b = -.15, SE = .12, t = -1.3, $p = .21$, 2nd administration: b = -.13, SE = .13, t = -1.1, $p = .29$, but the dispositional mindfulness × confidence parameter did in the 2nd administration, b = .10, SE = .05, t = 2.0, *p* = .044, but not the 1st administration, b = -.01, SE = .05, t = -0.2, *p* = .87. The interaction between report type and dispositional mindfulness was not associated with response time variability in either the first, b = -.01, SE = .11, t = -0.1, *p* = .89, or second metronome response task administration, b = -.23, SE = .12, t = -1.8, *p* = .06, nor was the three-way interaction among dispositional mindfulness, confidence, and report type, 1st administration: b = .09, SE = .09, t = 1.0, *p* = .32, 2nd administration: b = .00, SE = .10, t = 0.0, *p* = .97.

Because this manuscript is already results laden, I only briefly comment on the most germane aspects of the LMMs on the five separate facets of mindfulness measured. Supplemental tables including all parameter estimates, standard errors, t values, and p values for these models are available at the following link: https://osf.io/n9yxk/. None of the critical three-way interactions among the mindfulness facets (i.e., Acting with Awareness, Nonjudging of Internal Experience, Nonreactivity to Internal Experience, Observing, and Describing), confidence, and report type reached statistical significance. With few exceptions the models followed the now familiar pattern of report type, confidence and their interaction reaching statistical significance in the first administration, and only report type and confidence reaching statistical significance in the second administration.

Omnibus Linear Mixed Model. In this model, I predicted response time variability with report type, working memory capacity, conscientiousness, neuroticism, and dispositional mindfulness and their interactions as the predictors. Working memory capacity, conscientiousness, neuroticism, and dispositional mindfulness were entered as mean-centered continuous variables. Report type was effect coded. I ran separate models for each metronome response task administration. In the first administration model (see Table 9 for all parameter estimates, standard errors, degrees of freedom, t-values, and p-values), three parameters reached statistical significance: Report type, b = .19, SE = .05, t = 3.8, *p* < .001, with reports of mind wandering associated with more response time variability, and working memory capacity, b = -.20, SE = .07, t = -2.8, *p* = .006, with higher working memory capacity subjects exhibiting less response time variability than subjects with lower working memory capacity. In addition, the four-way interaction among report type, dispositional mindfulness, neuroticism and working memory capacity predicted response time variability, b = -.91, SE = .30, t = -3.1, *p* = .002. I judged this interaction to be uninterpretable. In the second administration (see Table 10 for details), report type again predicted response time variability, b = .27, SE = .05, t = 5.0, *p* < 001, (i.e., report of mind wandering being associated with greater response time variability) as did the same four-way interaction of report type, dispositional mindfulness, neuroticism and working memory capacity, b = -.81, SE = .37, t = -2.2, *p* = .03, which I again judged to be uninterpretable.

Monitoring Ability Index Point-biserial correlations. As an index of monitoring ability, for every subject I computed the point-biserial correlation between the report type (0 = on-task, 1 = mind wandering) and response time variability with the idea being that subjects who are better than others at monitoring would show a stronger positive relation. More simply put, subjects who are more accurate at monitoring their thoughts would exhibit a greater distinction in response time variability between their on-task and mind wandering reports. This monitoring ability index was only able to be computed for subjects who chose both report types (in the 18 thought probes) within a single administration. Descriptive statistics for this monitoring index for both administrations can be found in Table 1. These indexes correlated with each other weakly and nonsignificantly, *r* (230) = .10, *p* = .12, *95% CI* [-.03, .23], BF_{10} = .27, providing no evidence that monitoring thoughts (as measured here) is a stable individual difference. In accord with the preregistration, I did not pursue further analyses with this attempted monitoring index because the correlation between the two administrations was lower than the .30 threshold. Psychometric issues with this measurement are briefly considered in the General Discussion.

Extension Discussion

The correlational analyses that repeated from the first to second administration of the metronome response tasks lead to the same inferences as made for the first administration. Mind wandering is positively associated with response time variability, but only weakly (and non-significantly) related to confidence while confidence and response time variability were not associated. Again, in the hierarchical regressions, confidence ratings did not moderate the association between mind wandering and response time variability suggesting that some subjects were not able to access or report fine-grained metacognitive information about the thought report validity better than others. In this second administration, none of the within-subject analyses provide evidence that subjects were either willing or able to continue to discriminate the validity of their subjective self-reports by confidence ratings.

I assessed putative individual difference markers of monitoring ability in both metronome response administrations. Neither working memory capacity, conscientiousness, neuroticism, nor dispositional mindfulness (composite or individual facets) moderated the mind wandering and response time variability relation in either task administration. In the LMMs, the only statistically significant three-way interaction (representing individual differences in monitoring) was for working memory capacity in the 2nd administration of the metronome response task. Finally, I computed an index of monitoring ability and assessed whether monitoring (as measured) is a stable individual difference. Only a weak relation was found, and the Bayesian model comparison suggested the null hypothesis over the alternative hypothesis by slightly greater than a factor of three. Because the association between monitoring indexes did not meet the a priori criterion of *r* = .30, I did not pursue any further analyses with this monitoring index.

General Discussion

Seli et al. found evidence for individual differences in the ability to monitor thoughts. In their study, high-confidence subjects showed a greater difference in response time variability between reports of being on-task or mind wandering suggesting that some people either have higher fidelity thoughts (i.e., less noisy representations) or more fine-tuned assessment of these thoughts. With a large sample of subjects, I found no evidence of robust individual differences in conscious thought monitoring or in thought probe validity. In two administrations of the metronome response task, I found the established association of mind wandering with response time variability, but neither retrospective confidence ratings, working memory capacity, conscientiousness, neuroticism, nor dispositional mindfulness moderated this relation.

Although I found no evidence for individual differences in thought report validity and only very weak evidence for individual differences in thought monitoring, I did find support for two other findings that Seli et al. described as key: 1) Variation in the confidence ratings that subjects provide with the distributions of confidence reports being remarkably similar to those found in Seli et al. (see Table 4 for comparisons across studies and administrations); 2) Some support for the within-subjects' finding that when subjects reported low confidence, response time variability did not significantly differ between reports of on-task and mind-wandering. However, regarding the within-subjects' result, this support was only found in the first administration of the metronome response task and even there, when considering Bayesian model comparisons and effect sizes, does not appear robust.

If we see individual differences in rates of mind wandering but cannot find individual differences in the consciously available monitoring of mind wandering, control mechanisms other than conscious monitoring are implicated. In other words, if I assume that for a subject to redirect their thoughts back to the primary task, they first notice they are off task and then shift their attention, the result of no robust individual differences in noticing suggests the differences lie in the ability to volitionally shift. Thus, this finding narrows the control processes under consideration in the control failures × concerns model of mind wandering. Adding more specificity to this model will allow it to make more precise predictions (and progress) in future work.

Examining the relation between thought probes, confidence, and response time variability is not the only way to assess meta-awareness. Another way is to contrast rates of self-caught to experimenter-caught mind wandering. In studies using this approach, in addition to encountering thought probes within the task, subjects are also instructed to press a designated button when they catch themselves mind wandering. Meta-awareness is gauged as the ratio of self-caught to experimenter-caught probes. That is, more aware subjects (should) have a higher rate of self-caught to probe-caught mind wandering than less aware subjects. Until recently, the findings in the current study regarding individual differences in monitoring of thoughts jibed well with work that has used this method. Prior to the most recently published work with this method (Seli et al., in press), only experimenter-caught mind wandering had shown a relationship with task performance (Levinson, Smallwood, & Davidson, 2012; Sayette, Reichle, & Schooler, 2009; Sayette, Schooler, & Reichle, 2010). Seli et al. (in press) have provided the first evidence in the self-caught paradigm for an association between self-caught probes and task performance suggesting that this may yet be a fruitful avenue of research into aspects of mind wandering awareness. One salient difference between Seli et al. (in press) and the prior self-caught probe studies is the sample size. Seli et al. found evidence of an association with 105 subjects (in two separate samples); the three earlier studies cited above that failed to detect the association had sample sizes ranging from 40-52.

Likewise, recent work has highlighted the role psychometric properties of tasks play in replication attempts of individual differences work (Cooper, Gonthier, Barch, & Braver, 2017). Specifically, the role of psychometric reliability in limiting the magnitude of potential correlations seems relevant here. For instance, if Seli et al.'s (2015) finding of a positive correlation between confidence and mind wandering was substantially different in the replication attempt and the reliability of these measures was substantially lower in this attempt, the discrepancy between the two studies can be attributed to the lower reliability. Low reliability does not seem to be an issue in the current experiment, as seen in Tables 1 and 2, the measures used in the replication analyses had ample variation and are not affected by floor or ceiling effects. These measures all had Cronbach's alphas of greater than .80 (deemed as adequate for basic research by Nunnaly, 1978), and test-retest correlations of greater than .70. However, psychometric issues may very well account for the lack of evidence of stable individual differences found with the monitoring ability index. Although these indexes showed ample variation, their test-retest correlation was only .10 and nonsignificant. My attempt to assess this was by creating within-subject correlations between thought reports and response time variability. Because these correlations were based on only 18 observations, these correlations are noisy measures (Schönbrodt, & Perugini, 2013), thus limiting the statistical power. Future work in investigating individual differences in monitoring ability must use high-powered designs to produce evidential value. Although the attempt to create an index of monitoring ability was ineffective, it does not subtract from the evidential value provided by the other analyses.

Because the truth is not (usually) revealed by any one study and sampling error can either effect the original, the replication, or both studies, my interpretation of these two studies will be constrained. There are many positive attributes of this replication and extension study that may cause one to weigh its results more heavily than the original. The larger sample size allows the estimation of effects with more precision, preregistration limited the analytical and data collection flexibility (making the p-values interpretable), and analyses performed by Seli et al. were supplemented by incorporating a Bayesian approach and LMMs. In addition to these attributes, I am confident in the results because the expected pattern of results in associations that were not the critical focus here were found. For example, response time variability and mind wandering were significantly positively associated in both metronome response task administrations (first administration, $r = .18$, $p = .001$, $BF_{10} = 11$; second administration, $r = .20$, $p = .0004$, $BF_{10} = 41$), neuroticism and conscientiousness were significantly negatively associated ($r = -.43$, $p < .00001$, $BF_{10} = 56{,}470{,}715{,}959$), and working memory capacity and response time variability were negatively associated in both administrations (first administration, $r = -.26$, $p = .00007$, $BF_{10} = 220$; second administration, $r = -.19$, $p = .005$, $BF_{10} = 4$). These expected patterns work as positive controls, strongly suggesting that if individual differences variables (in this sample) did moderate the relation between subjective and objective measures of mind wandering, they would have been found (i.e., this study does not suffer from measurement problems in general).

To foster a more efficient accumulation of knowledge, I provide constraints on the generalizability of the findings (Simons, Shoda, & Lindsay, 2016). I expect the results reported here to generalize to lab-based measures of mind wandering and meta-awareness. To my knowledge, these relations have not been tested outside of the lab. Recently, Kane et al., (2017) have communicated evidence that relations among mind wandering and other constructs differ between these two settings (i.e., lab versus out of lab). Moreover, the claims reported here are specific to younger adults. Older and younger adults (in the lab) have shown divergent patterns in their reporting of mind wandering (Frank, Nara, Zavagnin, Touron, & Kane, 2015; Giambra, 1989; Jackson & Balota, 2012; McVay, Meier, Touron, & Kane, 2013). Because the present analyses included a measure of cognitive ability (i.e., working memory capacity) that is associated with fluid intelligence and very little moderation was found, these results are expected generalize no matter the cognitive ability of the sample.

Conclusion

Seli et al. suggest that subjects “are able to provide more valid and reliable reports” about the certainty with which they respond to thought probes. The evidence contained here calls this assertion into doubt and suggests that all subjects monitor their mind wandering with only limited resolution. Thus, as has already been established, thought probes are valid (but noisy) measures of on versus off-task thinking, but subjects are limited in their ability to qualify these reports. Although the study of mind wandering is continually yielding new findings and building out the nomological network, it appears there are some topics of study (e.g., qualification of thought probe responses) that are not well suited to be revealed by subjective thought reports alone. Because subjects cannot comment authoritatively on some aspects of mind wandering experiences, more objective measures like pupillometry (Konishi, Brown, Battaglini, &

Smallwood, 2017) may be most efficient path forward. To be clear, I am not claiming that individual differences in thought monitoring do not exist, I am claiming that the evidence provided in Seli et al. (2015) and the current study (considered together) offer insufficient support for this claim.

REFERENCES

Adam, K. C., & Vogel, E. K. (2017). Confident failures: lapses of working memory reveal a metacognitive blind spot. *Attention, Perception, & Psychophysics*, 1-18.

Andrews-Hanna, J. R., Smallwood, J., & Spreng, R. N. (2014). The default network and self-generated thought: component processes, dynamic control, and clinical relevance. *Annals of the New York Academy of Sciences*, *1316*(1), 29-52.

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., & Singmann, H. (2015). lme4: Linear mixed-effects models using Eigen and S4, 2014. *R package version*, *1*(4).

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., ... & Cesarini, D. (2017). Redefine statistical significance. Nature Human Behaviour, 1.

Bishop, S. R., Lau, M., Shapiro, S., Carlson, L., Anderson, N. D., Carmody, J., ... & Devins, G. (2004). Mindfulness: A proposed operational definition. *Clinical Psychology: Science and Practice*, *11*(3), 230-241.

Bohlmeijer, E., ten Klooster, P. M., Fledderus, M., Veehof, M., & Baer, R. (2011). Psychometric properties of the five facet mindfulness questionnaire in depressed adults and development of a short form. *Assessment*, *18*(3), 308-320.

Brewer, Gene A., Kevin KH Lau, Kimberly M. Wingert, B. Hunter Ball, and Chris Blais. "Examining depletion theories under conditions of within-task transfer." *Journal of Experimental Psychology: General* 146, no. 7 (2017): 988-1008.

Bühler, K. (1907). *Tatsachen und Probleme zu einer Psychologie der Denkvorgänge*. Engelmann.

Chambers, R., Lo, B. C. Y., & Allen, N. B. (2008). The impact of intensive mindfulness training on attentional control, cognitive style, and affect. *Cognitive Therapy and Research*, *32*(3), 303-322.

Cai, W., Chen, T., Ryali, S., Kochalka, J., Li, C. S. R., & Menon, V. (2015). Causal interactions within a frontal-cingulate-parietal network during cognitive control: Convergent evidence from a multisite–multitask investigation. *Cerebral Cortex*, 26(5), 2140-2153.

Callard, F., Smallwood, J., Golchert, J., & Margulies, D. S. (2013). The era of the wandering mind? Twenty-first century research on self-generated mental activity. *Frontiers in Psychology*, 4.

Cooper, S. R., Gonthier, C., Barch, D. M., & Braver, T. S. (2017). The role of psychometrics in individual differences research in cognition: A case study of the AX-CPT. *Frontiers in Psychology*, *8*.

Coull, J. T., Frith, C. D., Frackowiak, R. S. J., & Grasby, P. M. (1996). A fronto-parietal network for rapid visual information processing: a PET study of sustained attention and working memory. *Neuropsychologia*, 34(11), 1085-1095.

Darki, F., & Klingberg, T. (2014). The role of fronto-parietal and fronto-striatal networks in the development of working memory: a longitudinal study. *Cerebral Cortex*, *25*(6), 1587-1595.

Etz, A., & Wagenmakers, E. J. (2017). JBS Haldane’s contribution to the Bayes factor hypothesis test. *Statistical Science*, *32*(2), 313-329.

Fox, E., Dutton, K., Yates, A., Georgiou, G. A., & Mouchlianitis, E. (2015). Attentional control and suppressing negative thought intrusions in pathological worry. *Clinical Psychological Science, 3*(4), 593-606.

Frank, D. J., Nara, B., Zavagnin, M., Touron, D. R., & Kane, M. J. (2015). Validating older adults' reports of less mind-wandering: An examination of eye movements and dispositional influences. *Psychology and Aging, 30*(2), 266-278.

Franklin, M. S., Mrazek, M. D., Anderson, C. L., Johnston, C., Smallwood, J., Kingstone, A., & Schooler, J. W. (2017). Tracking distraction: The relationship between mind-wandering, meta-awareness, and ADHD symptomatology. *Journal of Attention Disorders*, *21*(6), 475-486.

Giambra, L. M. (1989). Task-unrelated thought frequency as a function of age: A laboratory study. *Psychology and Aging*, *4*(2), 136.

Gulbinaite, R., van Rijn, H., & Cohen, M. X. (2014). Fronto-parietal network oscillations reveal relationship between working memory capacity and cognitive control. *Frontiers in Human Neuroscience*, *8*.

Ham, T., Leff, A., de Boissezon, X., Joffe, A., & Sharp, D. J. (2013). Cognitive control and the salience network: an investigation of error processing and effective connectivity. *Journal of Neuroscience, 33*(16), 7091-7098.

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

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: a latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology:* General, *133*(2), 189-217.

Kane, M. J., & McVay, J. C. (2012). What mind wandering reveals about executive-control abilities and failures. *Current Directions in Psychological Science*, *21*(5), 348-354.

Kane, M. J., Meier, M. E., Smeekens, B. A., Gross, G. M., Chun, C. A., Silvia, P. J., & Kwapil, T. R. (2016). Individual differences in the executive control of attention, memory, and thought, and their associations with schizotypy. *Journal of Experimental Psychology: General*, *145*(8), 1017.

Kane, M. J., Gross, G. M., Chun, C. A., Smeekens, B. A., Meier, M. E., Silvia, P. A., & Kwapil, T. R. (in press). For whom the mind wanders, and when, II: Individual differences in subjective experience vary across laboratory and daily-life settings. *Psychological Science*.

Konishi, M., Brown, K., Battaglini, L., & Smallwood, J. (2017). When attention wanders: Pupillometric signatures of fluctuations in external attention. *Cognition*, 168, 16-26.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430), 773-795.

Kliegl, R., Wei, P., Dambacher, M., Yan, M., & Zhou, X. (2010). Experimental effects and individual differences in linear mixed models: Estimating the relationship between spatial, object, and attraction effects in visual attention. *Frontiers in Psychology*, 1.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2015). Package ‘lmerTest’. R package version, 2(0).

Levinson, D. B., Smallwood, J., & Davidson, R. J. (2012). The persistence of thought: evidence for a role of working memory in the maintenance of task-unrelated thinking. *Psychological Science,* 23(4), 375-380.

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods*, 49(4), 1494-1502.

McCrae, R. R., & Costa, P. T. (2010). NEO Inventories professional manual. Lutz, FL: Psychological Assessment Resources.

McVay, J. C., & Kane, M. J. (2009). Conducting the train of thought: working memory capacity, goal neglect, and mind wandering in an executive-control task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(1), 196-204.

McVay, J. C., & Kane, M. J. (2010). Does mind wandering reflect executive function or executive failure? Comment on Smallwood and Schooler (2006) and Watkins (2008). *Psychological Bulletin*,

McVay, J. C., & Kane, M. J. (2012a). Drifting from slow to “d'oh!”: Working mmemory capacity and mind wandering predict extreme reaction times and executive control errors. *Journal Of Experimental Psychology: Learning, Memory, and Cognition*, 38(3), 525-549.

McVay, J. C., & Kane, M. J. (2012b). Why does working memory capacity predict variation in reading comprehension? On the influence of mind wandering and executive attention. *Journal of Experimental Psychology: General, 141*(2), 302-320.

McVay, J. C., Meier, M. E., Touron, D. R., & Kane, M. J. (2013). Aging Ebbs the Flow of Thought: Adult Age Differences in Mind Wandering, Executive Control, and Self-Evaluation. *Acta psychologica, 142*(1), 136 -147.

Meier, M. E., Smeekens, B. A., Silvia, P. J., Kwapil, T. R., & Kane, M. J. (2018). Working memory capacity and the antisaccade task: A microanalytic–macroanalytic investigation of individual differences in goal activation and maintenance. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(1), 68 -84.

Moore, A., & Malinowski, P. (2009). Meditation, mindfulness and cognitive flexibility. *Consciousness and Cognition*, *18*(1), 176-186.

Morey, R. D., Rouder, J. N., & Jamil, T. (2015). BayesFactor: Computation of Bayes factors for common designs. R package version 0.9, 9.

Mrazek, M. D., Franklin, M. S., Phillips, D. T., Baird, B., & Schooler, J. W. (2013). Mindfulness training improves working memory capacity and GRE performance while reducing mind wandering. *Psychological Science*, 24(5), 776-781.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than I can know: Verbal reports on mental processes. *Psychological Review*, 84(3), 231-259.

Nunnally, J. (1978). Psychometric Methods.

Perkins, A. M., Arnone, D., Smallwood, J., & Mobbs, D. (2015). Thinking too much: self-generated thought as the engine of neuroticism. *Trends in Cognitive Sciences*, 19(9), 492-498.

Psychology Software Tools, Inc. [E-Prime 2.0]. (2012). Retrieved from http://www.pstnet.com.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

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. *Proceedings of the National Academy of Sciences*, 98(2), 676-682.

Robison, M. K., Gath, K. I., & Unsworth, N. (2017). The neurotic wandering mind: An individual differences investigation of neuroticism, mind-wandering, and executive control. *The Quarterly Journal of Experimental Psychology*, *70*(4), 649-663.

Robinson, M. D., & Tamir, M. (2005). Neuroticism as mental noise: a relation between neuroticism and reaction time standard deviations. *Journal of Personality and Social Psychology,* 89(1), 107.

Sayette, M. A., Reichle, E. D., & Schooler, J. W. (2009). Lost in the sauce: The effects of alcohol on mind wandering. *Psychological Science*, 20(6), 747-752.

Sayette, M. A., Schooler, J. W., & Reichle, E. D. (2010). Out for a smoke: The impact of cigarette craving on zoning out during reading. *Psychological Science,* 21(1), 26-30.

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize?. Journal of Research in Personality, 47(5), 609-612.

Seli, P. (2016). The attention-lapse and motor decoupling accounts of SART performance are not mutually exclusive. *Consciousness and Cognition*, *41*, 189-198.

Seli, P., Carriere, J. S., Levene, M., & Smilek, D. (2013). How few and far between? Examining the effects of probe rate on self-reported mind wandering. *Frontiers in Psychology,* 4.

Seli, P., Carriere, J. S., 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.

Seli, P., Cheyne, J. A., & Smilek, D. (2013). Wandering minds and wavering rhythms: Linking mind wandering and behavioral variability. *Journal of Experimental Psychology: Human Perception and Performance*, 39(1), 1-5.

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

Seli, P., Jonker, T. R., Cheyne, J. A., & Smilek, D. (2013). Enhancing SART validity by statistically controlling speed-accuracy trade-offs. *Frontiers in Psychology, 4*.

Seli, P., Ralph, B. C. W., Smilek, D., & Schacter, D. L. (in press). The awakening of the attention: Evidence for a link between the monitoring of mind wandering and prospective goals. *Journal of Experimental Psychology: General.*

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, 1359-1366.

Simons, D. J., Shoda, Y., & StephenLindsay, D. (2016). Constraints on Generality (COG): A Proposed Addition to All Empirical Papers.

Tangney, J. P., Baumeister, R. F., & Boone, A. L. (2004). High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. *Journal of Personality,* 72(2), 271-324.

Teper, R., Segal, Z. V., & Inzlicht, M. (2013). Inside the mindful mind: How mindfulness enhances emotion regulation through improvements in executive control. *Current Directions in Psychological Science*, 22, 449-454.

Unsworth, N. (2015). Consistency of attentional control as an important cognitive trait: A latent variable analysis. *Intelligence,* 49, 110-128.

Unsworth, N., & McMillan, B. D. (2013). Mind wandering and reading comprehension: examining the roles of working memory capacity, interest, motivation, and topic experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(3), 832-842.

Unsworth, N., & Robison, M. K. (2017). A locus coeruleus-norepinephrine account of individual differences in working memory capacity and attention control. *Psychonomic Bulletin & Review*, 1-30.

Wagenmakers, E. J., Verhagen, J., & Ly, A. (2016). How to quantify the evidence for the absence of a correlation. *Behavior Research Methods*, 48, 413-426.

Watson, J. B. (1913). Psychology as the behaviorist views it. *Psychological Review*, *20*(2**), 158.**

Weinstein, Y., De Lima, H. J., & van der Zee, T. (2017). Are you mind-wandering, or is your mind on task? The effect of probe framing on mind-wandering reports. *Psychonomic Bulletin & Review*, 1-7.

Weinstein, Y. (2017). Mind-wandering, how do I measure thee with probes? Let me count the ways. *Behavior Research Methods*, 1-20.

Wundt, W. M. (1907). *Outlines of psychology*. W. Engelmann.

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.

Zanto, T. P., & Gazzaley, A. (2013). Fronto-parietal network: flexible hub of cognitive control. *Trends in Cognitive Sciences*, 17(12), 602-603.

Zeidan, F., Johnson, S. K., Diamond, B. J., David, Z., & Goolkasian, P. (2010). Mindfulness meditation improves cognition: Evidence of brief mental training. *Consciousness and Cognition*, 19(2), 597-605.

Footnotes

1. In the preregistration, the thesis project mentioned was never completed. Also, in the preregistered analysis plan, I stated that the dependent variable in the linear mixed model analysis would be the point-biserial correlation. Before beginning the analyses included here, I realized this would not make sense as the dependent variable and that only response time variability does. Using the point-biserial correlation as the dependent variable in these analyses does not make sense because it is a constant for each subject. This would reduce the information available per subject thus robbing the LMMs of their ability to provide precise estimates.

Table 1. Descriptive Statistics

Measure	M	SD	Min	Max	Skew	Kurtosis	N
MRT 1 Response Time Variability	8.77	0.81	7.08	11.44	0.67	0.48	291
MRT 2 Response Time Variability	8.97	0.82	7.00	11.67	0.38	0.00	291
MRT 1 Mind Wandering	0.55	0.26	0.00	1.00	-0.30	-0.69	291
MRT 2 Mind Wandering	0.63	0.30	0.00	1.00	-0.57	-0.75	291
MRT 1 Confidence Rating	3.77	0.67	1.67	5.00	-0.29	-0.16	291
MRT 2 Confidence Rating	3.80	0.76	1.28	5.00	-0.36	-0.08	291
Operation Span	50.81	14.68	3.00	75.00	-0.65	0.00	250
Symmetry Span	28.53	6.83	9.00	41.00	-0.44	-0.33	252
Working Memory Capacity	0.04	0.79	-2.27	1.64	-0.35	-0.45	221
Conscientiousness	3.66	0.56	2.00	5.00	-0.14	-0.38	278
Neuroticism	2.84	0.60	1.25	4.83	0.15	0.06	278
Dispositional Mindfulness Total	3.28	0.36	2.21	4.38	-0.15	0.25	278
Describing	3.28	0.50	2.00	4.60	0.01	-0.06	278
Nonreactivity	3.27	0.50	1.80	4.60	-0.09	-0.28	278
Nonjudging	3.29	0.50	1.80	4.60	-0.37	0.08	278
Observing	3.24	0.54	1.75	4.50	-0.12	-0.51	278

Acting with Awareness	3.31	0.51	1.60	4.80	-0.12	0.24	278
MRT 1 Point-biserial Correlation	0.09	0.25	-0.55	0.69	-0.24	-0.22	277
MRT 2 Point-biserial Correlation	0.11	0.24	-0.54	0.78	0.10	0.01	239

Note. *Note.* M = Mean of subject means; SD = Standard deviation of subject means; Min = Minimum; Max = Maximum; MRT 1 = First administration of metronome response task; MRT 2 = Second administration of metronome response task.

Table 2. Correlations Among Measures, with Coefficient Alphas (for Uncombined Measures) Presented on the Diagonal

Measure	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1. MRT 1 Response Time Variability	0.85																		
2. MRT 2 Response Time Variability	0.78	0.87																	
3. MRT 1 Mind Wandering	0.18	0.16	0.84																
4. MRT 2 Mind Wandering	0.11	0.21	0.74	0.91															
5. MRT 1 Confidence Rating	-0.11	0.00	0.01	0.04	0.89														
6. MRT 2 Confidence Rating	0.01	-0.01	0.08	0.04	0.72	0.92													
7. Operation Span	-0.20	-0.18	0.03	0.03	0.08	0.01	0.83												
8. Symmetry Span	-0.19	-0.14	-0.06	-0.06	0.08	0.04	0.36	0.61											
9. Working Memory Capacity	-0.26	-0.19	-0.05	0.00	0.08	0.03	0.82	0.82	-										
10. Conscientiousness	-0.07	-0.02	-0.14	-0.12	0.05	-0.04	-0.08	-0.13	-0.12	0.78									
11. Neuroticism	0.09	0.07	0.11	0.08	-0.07	0.03	-0.05	0.01	-0.03	-0.43	0.78								
12. Dispositional Mindfulness Total	-0.10	-0.10	-0.18	-0.15	0.14	0.04	0.04	-0.06	-0.03	0.45	-0.65	0.72							
13. Describing	-0.11	-0.07	-0.13	-0.07	0.13	0.07	-0.01	-0.03	-0.03	0.29	-0.44	0.70	1.00						
14. Nonreactivity	0.01	-0.05	-0.08	-0.13	0.15	0.07	0.09	0.01	0.03	0.32	-0.40	0.69	0.33	1.00					
15. Nonjudging	-0.11	-0.10	-0.12	-0.08	0.11	-0.04	0.02	-0.06	-0.07	0.31	-0.45	0.74	0.38	0.45	1.00				
16. Observing	-0.11	-0.10	-0.17	-0.12	0.01	-0.03	0.02	-0.09	-0.04	0.34	-0.51	0.65	0.33	0.27	0.39	1			
17. Acting with Awareness	-0.04	-0.04	-0.14	-0.14	0.10	0.07	0.02	-0.06	0.00	0.33	-0.48	0.72	0.41	0.36	0.39	0.36	1.00		
18. MRT 1 Point-biserial Correlation	0.11	0.08	-0.04	-0.01	-0.04	-0.01	-0.10	-0.07	-0.08	-0.07	0.02	0.00	-0.01	0.02	0.02	-0.02	-0.01	-	
19. MRT 2 Point-biserial Correlation	0.08	0.16	-0.05	-0.11	-0.03	0.02	-0.12	0.00	-0.07	-0.04	0.01	-0.02	-0.10	-0.01	-0.02	0.07	0.00	0.07	-

Note. MRT 1 = First administration of metronome response task; MRT 2 = Second administration of metronome response task.

Table 3. Hierarchical Multiple Regression Models with Report Type, Confidence, and their Interaction as Predictors of Response Variability

		Seli et al. (2015) N = 100			1st Administration N = 291			2nd Administration N = 291		
	Predictor	semi-partial correlation	t value	*p*	semi-partial correlation	t value	*p*	semi-partial correlation	t value	*p*
Step 1	Report Type	0.29	3.04	.003	0.19	3.24	.001	0.21	3.6	.0003
	Confidence	0.01	0.85	.932	-0.11	-1.91	.060	-0.03	-0.4	.720
Step 2	Report Type	0.24	2.53	.013	0.19	3.12	.002	0.22	3.6	.0004
	Confidence	-0.01	-0.13	.895	-0.11	-1.91	.060	-0.02	-0.3	.750
	Report Type × Confidence	0.25	2.62	.010	-0.01	-0.25	.800	-0.02	-0.4	.670

Table 4. Frequency Distributions of Confidence Ratings for Reports of On-Task and Mind Wandering

	Confidence	Seli et al. (N = 100)			1st MRT Administration (N = 291)			2nd MRT Administration N = (291)		
Report Type	Rating	Frequency	%	Cumulative %	Frequency	%	Cumulative %	Frequency	%	Cumulative %
On-task	1	13	1.6	1.6	35	1.5	1.5	45	2.3	2.3
	2	112	13.9	15.5	281	11.8	13.3	264	13.6	15.9
	3	257	31.8	47.3	633	26.7	40	507	26.1	42.1
	4	211	26.1	73.5	655	27.6	67.6	486	25.1	67.1
	5	214	26.5	100	768	32.4	100	637	32.9	100
Mind Wandering	1	31	3.1	3.1	100	3.5	3.5	137	4.2	4.2
	2	129	13	16.1	354	12.4	15.8	339	10.3	14.5
	3	215	21.7	37.8	660	23	38.9	718	21.8	36.2
	4	267	26.9	64.7	740	25.8	64.7	819	24.8	61.1
	5	351	35.3	100	1012	35.3	100	1286	39	100

Table 5. Hierarchical Multiple Regression Models with Report Type, Working Memory Capacity, and their Interaction as Predictors of Response Time Variability

		1st Administration N = 220			2nd Administration N = 220		
	Predictor	semi-partial correlation	t value	*p*	semi-partial correlation	t value	*p*
Step 1	Report Type	0.14	2.10	.030	0.21	3.20	.0003
	Working Memory Capacity	-0.26	-3.98	<.0001	-0.18	-2.90	.004
Step 2	Report Type	0.13	2.01	.050	0.21	3.60	.0004
	Working Memory Capacity	-0.26	-3.96	.0001	-0.18	-2.80	.005
	Report Type × Working Memory Capacity	-0.08	1.30	.200	0.04	0.70	.500

Table 6. Hierarchical Multiple Regression Models with Report Type, Conscientiousness, and their Interaction as Predictors of Response Time Variability

		1st Administration N = 277			2nd Administration N = 277		
	Predictor	semi-partial correlation	t value	*p*	semi-partial correlation	t value	*p*
Step 1	Report Type	0.18	3.10	.002	0.20	3.40	.001
	Conscientiousness	-0.04	-0.68	.500	-0.01	0.00	.998
Step 2	Report Type	0.19	3.20	.002	0.20	3.50	.001
	Conscientiousness	-0.04	-0.74	.460	0.00	0.00	.998
	Report Type × Conscientiousness	-0.09	1.60	.110	0.07	-1.20	.230

Table 7. Hierarchical Multiple Regression Models with Report Type, Neuroticism, and their Interaction as Predictors of Response Time Variability

		1st Administration N = 277			2nd Administration N = 277		
	Predictor	semi-partial correlation	t value	*p*	semi-partial correlation	t value	*p*
Step 1	Report Type	0.18	3.10	.002	0.2	3.40	.001
	Neuroticism	0.07	1.10	.260	0.05	0.87	.390
Step 2	Report Type	0.18	3.10	.002	0.21	3.60	.000
	Neuroticism	0.06	1.10	.270	0.04	0.77	.440
	Report Type × Neuroticism	0.03	0.50	.580	0.09	1.60	.120

Table 8. Hierarchical Multiple Regression Models with Report Type, Dispositional Mindfulness, and their Interaction as Predictors of Response Time Variability

		1st Administration N = 277			2nd Administration N = 277		
	Predictor	semi-partial correlation	t value	*p*	semi-partial correlation	t value	*p*
Step 1	Report Type	0.17	3.00	.003	0.20	3.40	.001
	Dispositional Mindfulness	-0.07	-1.10	.260	0.05	0.87	.390
Step 2	Report Type	0.18	3.00	.002	0.21	3.60	.000
	Neuroticism	-0.07	-1.10	.260	0.04	0.77	.440
	Report Type × Neuroticism	-0.01	-0.12	.900	0.09	1.60	.120

Table 9. Linear Mixed Models with Report Type, Mindfulness, Conscientiousness, Neuroticism, Working Memory Capacity, and their Interactions as Predictors of Response Time Variability in the First Metronome Response Task

Predictor	Estimate	Std. Error	df	t value	*p*
(Intercept)	8.17	0.05	193.4	156.06	0.000
Report Type	0.19	0.05	3692.3	3.84	0.000
Mindfulness	-0.31	0.20	194.2	-1.57	0.120
Conscientiousness	-0.02	0.10	193.1	-0.24	0.810
Neuroticism	0.03	0.11	194.6	0.29	0.770
WMC	-0.20	0.07	195.8	-2.79	0.010
Report Type × Mindfulness	0.09	0.18	3686.8	0.5	0.620
Report Type × Conscientiousness	-0.02	0.09	3691.9	-0.27	0.790
Mindfulness × Conscientiousness	-0.42	0.32	196.0	-1.31	0.190
Report Type × Neuroticism	0.10	0.10	3685.3	1.06	0.290
Mindfulness × Neuroticism	0.15	0.23	196.2	0.66	0.510
Conscientiousness × Neuroticism	-0.31	0.18	194.6	-1.78	0.080
Report Type × WMC	-0.10	0.07	3691.6	-1.48	0.140
Mindfulness × WMC	-0.07	0.24	198.2	-0.3	0.760
Conscientiousness × WMC	0.22	0.13	193.9	1.64	0.100
Neuroticism × WMC	-0.02	0.15	197.1	-0.13	0.890
Report Type × Mindfulness × Conscientiousness	-0.22	0.30	3692.2	-0.72	0.470
Report Type × Mindfulness × Neuroticism	0.37	0.21	3692.9	1.76	0.080

Report Type × Conscientiousness × Neuroticism	-0.16	0.16	3682.5	-0.99	0.320
Mindfulness × Conscientiousness × Neuroticism	-0.40	0.27	200.5	-1.46	0.150
Report Type × Mindfulness × WMC	0.03	0.22	3690.2	0.14	0.890
Report Type × Conscientiousness × WMC	0.11	0.12	3692.6	0.84	0.400
Mindfulness × Conscientiousness × WMC	-0.25	0.47	198.2	-0.52	0.600
Report Type × Neuroticism × WMC	-0.03	0.14	3692.9	-0.25	0.800
Mindfulness × Neuroticism × WMC	-0.05	0.32	199.4	-0.16	0.870
Conscientiousness × Neuroticism × WMC	0.06	0.27	196.8	0.23	0.820
Report Type × Mindfulness × Conscientiousness × Neuroticism	0.17	0.27	3670.9	0.66	0.510
Report Type × Mindfulness × Conscientiousness × WMC	-0.43	0.44	3690.1	-0.96	0.340
Report Type × Mindfulness × Neuroticism × WMC	-0.91	0.30	3689.8	-3.07	0.000
Report Type × Conscientiousness × Neuroticism × WMC	-0.14	0.24	3669.5	-0.57	0.570
Mindfulness × Conscientiousness × Neuroticism × WMC	0.16	0.47	217.7	0.34	0.740
Report Type × Mindfulness × Conscientiousness × Neuroticism × WMC	-0.55	0.49	3564.8	-1.12	0.26

Note. WMC = working memory capacity.

Table 10. Linear Mixed Models with Report Type, Mindfulness, Conscientiousness, Neuroticism, Working Memory Capacity, and their Interactions as Predictors of Response Time Variability in the Second Metronome Response Task

Predictor	Estimate	Std. Error	df	t value	*p*
(Intercept)	8.22	0.06	197.5	141.18	0.000
Report Type	0.27	0.05	3641.0	5.04	0.000
Mindfulness	-0.23	0.22	197.2	-1.05	0.290
Conscientiousness	-0.05	0.11	196.5	-0.43	0.670
Neuroticism	0.14	0.12	200.6	1.18	0.240
WMC	-0.12	0.08	202.5	-1.54	0.120
Report Type × Mindfulness	-0.34	0.20	3659.2	-1.67	0.100
Report Type × Conscientiousness	-0.09	0.10	3636.1	-0.91	0.360
Mindfulness × Conscientiousness	-0.34	0.36	195.4	-0.97	0.330
Report Type × Neuroticism	0.06	0.11	3583.9	0.56	0.570
Mindfulness × Neuroticism	0.13	0.25	201.7	0.51	0.610
Conscientiousness × Neuroticism	-0.32	0.20	203.4	-1.60	0.110
Report Type × WMC	-0.13	0.07	3663.2	-1.74	0.080
Mindfulness × WMC	-0.07	0.27	207.1	-0.27	0.790
Conscientiousness × WMC	-0.01	0.15	199.6	-0.09	0.930
Neuroticism × WMC	-0.16	0.17	205.9	-0.96	0.340
Report Type × Mindfulness × Conscientiousness	-0.05	0.34	3548.0	-0.14	0.890
Report Type × Mindfulness × Neuroticism	0.28	0.25	3527.4	1.10	0.270

Report Type × Conscientiousness × Neuroticism	-0.19	0.19	3604.6	-0.96	0.330
Mindfulness × Conscientiousness × Neuroticism	-0.57	0.30	201.4	-1.89	0.060
Report Type × Mindfulness × WMC	0.20	0.27	3433.3	0.73	0.470
Report Type × Conscientiousness × WMC	0.02	0.14	3631.7	0.16	0.870
Mindfulness × Conscientiousness × WMC	-0.64	0.52	199.3	-1.23	0.220
Report Type × Neuroticism × WMC	0.07	0.16	3605.4	0.46	0.650
Mindfulness × Neuroticism × WMC	-0.04	0.36	211.0	-0.11	0.920
Conscientiousness × Neuroticism × WMC	0.03	0.30	205.8	0.10	0.920
Report Type × Mindfulness × Conscientiousness × Neuroticism	-0.32	0.31	3188.3	-1.03	0.300
Report Type × Mindfulness × Conscientiousness × WMC	-0.27	0.52	3313.1	-0.51	0.610
Report Type × Mindfulness × Neuroticism × WMC	-0.81	0.37	3470.8	-2.19	0.030
Report Type × Conscientiousness × Neuroticism × WMC	0.16	0.28	3687.5	0.56	0.570
Mindfulness × Conscientiousness × Neuroticism × WMC	-0.30	0.52	213.5	-0.58	0.560
Report Type × Mindfulness × Conscientiousness × Neuroticism × WMC	-0.68	0.55	3015.1	-1.23	0.220

Note. WMC = working memory capacity.

AUTHOR NOTE

For project management, I thank Adam Lyons. For assistance in data collection, I thank Adam Lyons, Elizabeth Moss, Madison Pruitt, Josh Ritch, and Megan Donchenko. I thank Michael Kane for thoughtful comments on an earlier draft of this manuscript.

Correspondence should be addressed to Matt E. Meier, Department of Psychology, Room 302I Killian Bldg, Western Carolina University, Cullowhee, NC 28723, e-mail: mmeier@wcu.edu

Figure Captions

Figure 1. First administration of the metronome response task violin (density) and box plots of log-transformed response variability as a function of binned confidence ratings for subjects who had data in all bins (N = 61; Panel A) and for all subjects with imputed data (N = 291; Panel B). Dots are the distribution mean. Bars are 95% confidence intervals.

Figure 2. Second administration of the metronome response task violin (density) and box plots of log-transformed response variability as a function of binned confidence ratings for subjects who had data in all bins (N = 41; Panel A) and for all subjects with imputed data (N = 291; Panel B). Dots are the distribution mean. Bars are 95% confidence intervals.

Figure 1.

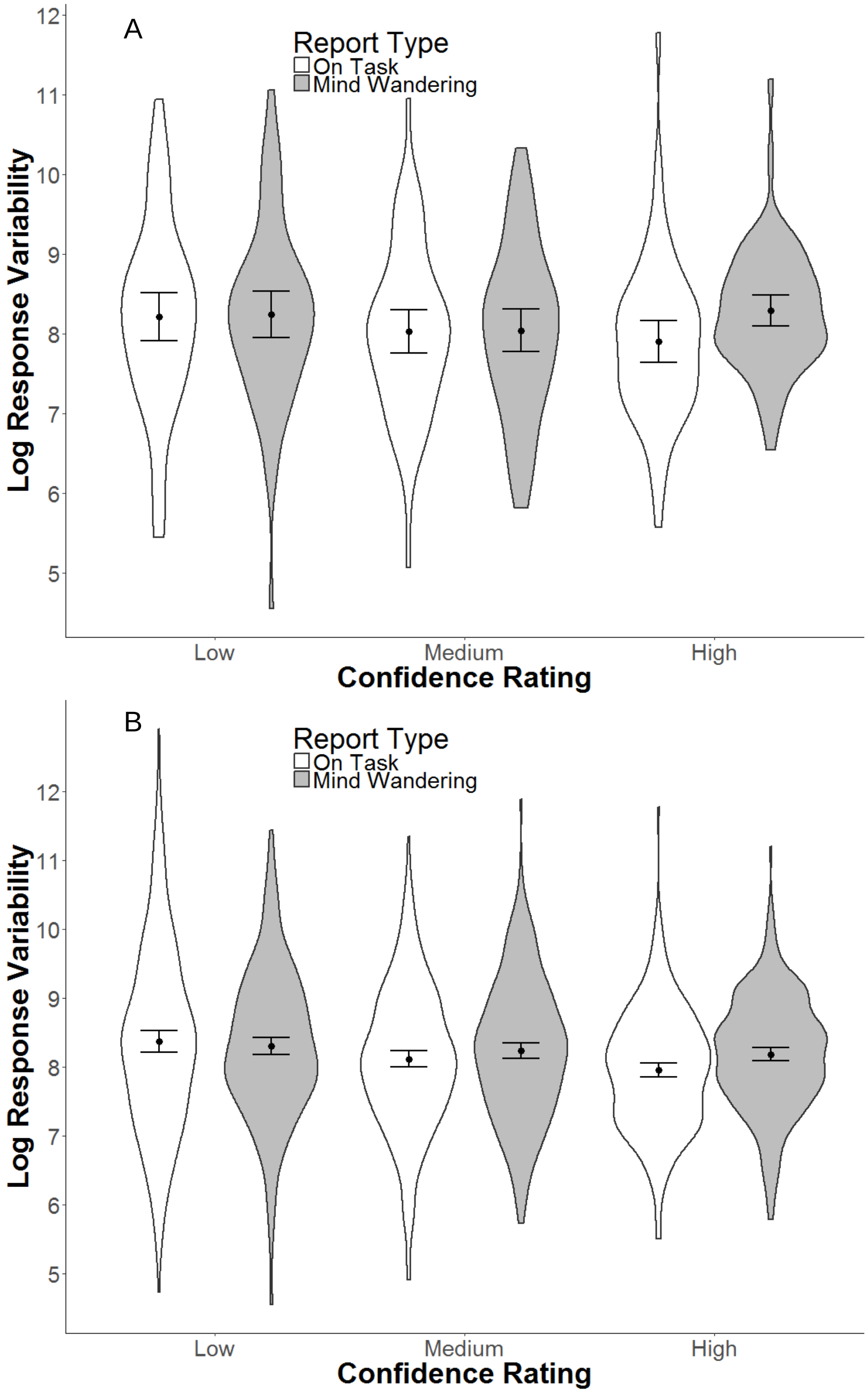
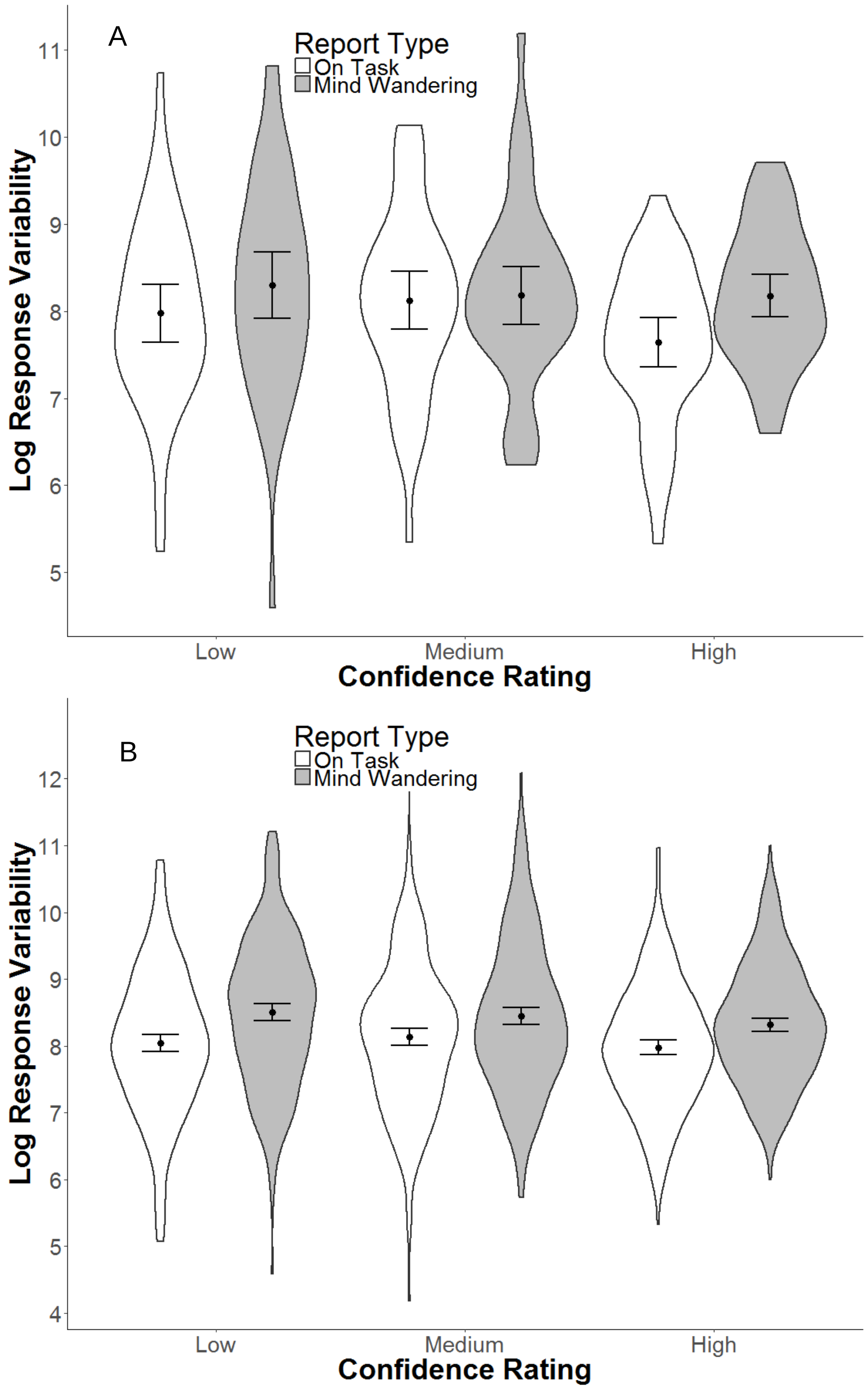

A
Report Type
□ On Task
■ Mind Wandering
Log Response Variability
12
11
10
9
8
7
6
5
Low
Medium
High
Confidence Rating
B
Report Type
□ On Task
■ Mind Wandering
Log Response Variability
12
11
10
9
8
7
6
5
Low
Medium
High
Confidence Rating

Figure 2.


A
Report Type
□ On Task
▨ Mind Wandering
Log Response Variability
11
10
9
8
7
6
5
Low
Medium
High
Confidence Rating
B
Report Type
□ On Task
▨ Mind Wandering
Log Response Variability
12
11
10
9
8
7
6
5
4
Low
Medium
High
Confidence Rating