What are the benefits of mind wandering to creativity?

Forthcoming in *Psychology of Creativity, Aesthetics, and the Arts*

*Samuel Murray[a,b], Nathan Liang[c], Nicholaus Brosowsky[a], and Paul Seli[a]

[a]Department of Psychology & Neuroscience, Duke University, Durham, NC
[b]Departamento de Psicología, Facultad de Ciencias Sociales, Universidad de los Andes, Bogotá, Colombia
[c]Department of Psychology, Princeton University, Princeton, NJ

Author Note

Correspondence may be directed to Samuel Murray, Department of Psychology & Neuroscience, Duke University, 417 Chapel Dr., Durham, NC, 27708. Email: samuel.f.murray@duke.edu.

Abstract

A primary aim of mind-wandering research has been to understand its influence on task performance. While this research has typically highlighted the *costs* of mind wandering, a handful of studies have suggested that mind wandering may be beneficial in certain situations. Perhaps the most-touted benefit is that mind wandering during a creative-incubation interval facilitates creative thinking. This finding has played a critical role in the development of accounts of the adaptive value of mind wandering and its functional role, as well as potential mechanisms of mind wandering. Thus, a demonstration of the replicability of this important finding is warranted. Here, we attempted to conceptually replicate results of a highly cited laboratory-based experiment supporting this finding. However, across two studies ($N = 443$), we found no evidence for the claim that mind wandering during a creative-incubation interval facilitates a form of creativity associated with divergent thinking. We suggest that our failed conceptual replication stems from an inadequate characterization of mind wandering (task-unrelated thought), and that there are good reasons to think that task-unrelated thought is unlikely to be causally related to creativity. Our results cast doubt on the claim that task-unrelated thought during an incubation interval enhances divergent creativity while also offering some prescriptions for how future research might further elucidate the cognitive benefits of mind wandering.

Keywords: Mind wandering; Task-Unrelated Thought; Incubation; Creativity; Divergent Thinking

> "Don't think. Thinking is the enemy of creativity…You can't try to do things. You simply must do things."
> *–Ray Bradbury*

Introduction

What makes somebody or something creative? Typically, something—like an idea, product, or solution—is considered creative when it is new. The creative musician generates a unique combination of sounds, the creative chef discovers an original combination of flavors, and so on. But novelty isn't everything: according to many, the invention must exhibit some value to somebody (Dietrich, 2019; Hennessey & Amabile, 2009; Kenett et al., 2020). After all, any hack can bang piano keys randomly or throw together random ingredients from their kitchen. What sets the musician or chef apart from the hack—true creativity—is the genesis of something that is both innovative and instrumental.

Different kinds of problem spaces require different kinds of creativity. Some problems have unique solutions, and in these spaces, the creative individual is one who can navigate the problem space to converge on this single solution. For example, the mathematician building a proof is searching for a single conclusion. This is a paradigmatic example of *convergent* creativity. Other problems require generating numerous potential solutions, and the creative individual is judged to be the one who can generate many unique solutions. For example, the screenwriter searching for new story ideas benefits from being able to flexibly generate a sizeable list of narratives. This exemplifies *divergent* creativity.

Creativity is crucial for many of the unique and valuable endeavors we undertake. It's important and often necessary for facilitating scientific, artistic, and political achievements that underscore the heights of human progress. But creativity also plays a role in mundane interactions, from improvising a recipe when lacking ingredients to maintaining engaging conversations over dinner. Creative people report experiencing better mood (Nadler et al., 2010), higher self-esteem (Barbot, 2018), and score higher on several dimensions of well-being (Conner et al., 2016). However, despite its demonstrated importance and benefits, much remains unknown about the cognitive mechanisms of creativity.

Anecdotal evidence about the mannerisms and discoveries of highly creative individuals provide some clues about the underlying processes supporting creative cognition. Many creative insights seem to occur in moments wherein the person is not actively focusing on generating a creative insight. Kekulé, for example, claims to have discovered the ring structure of the benzene molecule in a dream. The Indian mathematician Ramanujan was struck with numerous thoughts, the source of which he could not explain (so much so that he attributed his insights to a local Hindu goddess of good fortune; Cheng, 2017). Poincaré, Einstein, and Edison, among many others, all made similar claims. The collection of such anecdotal examples suggests an interesting phenomenon: setting a problem aside can lead to creative breakthroughs. This temporary disengagement is known as *incubation*, and there is a wealth of evidence to suggest that it is beneficial to creative thinking. For example, Gable et al. (2019) recently found that, when asked to report the creative quality and context of their most salient daily "aha" moments, physicists and writers exhibited a pattern of more frequently generating enlightening solutions during mind wandering relative to focused thinking.

The idea that incubation is readily conducive to creativity invites the question of what cognitive modes might promote creativity during incubation. This is important to understand, as such cognitive modes might be crucial for facilitating or even enhancing creativity. One relevant mode seems to be mind wandering. Consider that during incubation, an individual is not fixated on a particular problem. Thus, when incubating, a person is able to disengage and undergo cognitive exploration. Mind wandering during such periods might be particularly beneficial to certain kinds of creativity. Divergent creativity in particular, which requires forming novel associations, might benefit from mind wandering, because mind wandering is associated with hippocampally-mediated episodic memory reactivation (Ellamil et al., 2016; McCormick et al., 2018), which can lead to distinct contents being reactivated jointly in a novel context. This reactivation can thereby facilitate forming novel associations between contents that are less likely to form during periods of goal-directed thinking (Mills et al., 2018-a). Altogether, this implies that mind wandering during an incubation interval should enhance divergent creativity.

There is some correlational evidence for this relationship between mind wandering and divergent creativity. Recently, Yamaoka and Yukawa (2020) found a small but significant correlation between performance on a task that measures divergent creativity and self-reported susceptibility to mind wandering (as measured by the Japanese version of the Mind Wandering Questionnaire; Kajimura and Nomura, 2016). They also showed, using multiple linear regression, that scores on a divergent creativity task predicted trait mind wandering to a small degree, even after controlling for age, sex, self-reported depressive symptoms, and self-reported symptoms of schizotypal personality. Agnoli et al. (2018) identified significant positive correlations between scores on a divergent creativity task, trait susceptibility to deliberate mind wandering, and responses to the Creative Achievement Questionnaire. Finally, Tan et al. (2015)

found that participants who reported greater amounts of mind wandering during an incubation interval improved at a significantly greater rate on the number reduction task (a measure of both divergent and convergent creativity) relative to participants who reported less mind wandering. However, these results are correlational and, with the exception of Tan et al., rely on trait measures of mind wandering.

One influential study found experimental evidence for the predicted relationship between mind wandering and divergent creativity: Baird et al. (2012) examined performance on the Alternate Uses Task (AUT; a widely used measure of divergent creativity; Guilford, 1967) across several conditions following participants' initial efforts to think of novel and creative uses for a common object (e.g., brick). These conditions included: (a) performing a demanding (2-back) task, (b) performing a less-demanding (0-back) task, (c) having a period of rest, or (d) performing an immediate repetition of the AUT, with no intermediate activity. The authors hypothesized that mind wandering during an incubation interval would lead to greater divergent creativity. Based on this hypothesis, they predicted that participants in the less-demanding task would mind wander more and perform better on the AUT relative to other conditions. Consistent with their hypotheses, the less-demanding task was found to be associated with more mind wandering than the demanding task and, critically, the less-demanding task was also associated with greater subsequent production of additional creative 'uses' responses relative to other task conditions (demanding, rest, and immediate repetition). Thus, the authors concluded that mind wandering led to the discovery of new uses beyond those imagined in the first exposure.

This finding represented a major advance in the study of both creativity and mind wandering (as evidenced by the 914 citations Baird et al.'s paper has garnered as of April 9, 2021). On the creativity side, this presented a promising intervention to facilitate creative-idea

generation and possible cognitive mechanisms of certain kinds of creative cognition. For mind wandering, this finding provided evidence for conceptual accounts of the adaptive value (Sripada, 2018) and functional role (Shepherd, 2019) of mind wandering, and contributed to the development of novel frameworks of mind wandering (see Andrews-Hanna et al., 2014).

Smeekens and Kane (2016) conducted a series of conceptual replications of Baird et al. (2012). Across several multi-session studies measuring how working-memory capacity and other individual-differences variables affected mind wandering and creativity, Smeekens and Kane failed to find the expected significant relationships between mind wandering and AUT performance, despite utilizing various scoring procedures. In their Experiment 3 (the one modelled most closely on Baird et al.), they had participants perform the AUT, followed by an undemanding incubation-interval task, followed by a repeated AUT problem. Even here, Smeekens and Kane (2016) failed to find any relation between mind wandering during the incubation-interval and AUT scores. More recently, Steindorf et al. (forthcoming) conducted a conceptual replication of the results from Baird et al. and Smeekens and Kane. Participants completed both verbal and figural AUTs prior to a 12-min. incubation interval (0-back task). After the incubation interval, participants completed another round of the AUTs. Steindorf et al. also failed to find significant correlations between post-incubation AUT score and proportion of mind wandering.

Notably, Leszczynski et al. (2017) found that mind wandering during an incubation interval positively correlates with improved scores on a creativity task. Participants completed the SART between iterations of the Compound Remote Associates Test (Mednick, 1962). In the latter, participants are given three words and asked to find the shared associate (e.g., “cottage--Swiss--cake” would have “cheese” as its solution). Participants who reported more mind

wandering during the SART also produced more solutions in the second round of the task relative to those who reported less mind wandering. However, unlike Baird et al. (2012), Leszczynski et al. used a test of *convergent* creativity rather than *divergent* creativity. Hence, these results cannot be considered a close replication of Baird et al.

In the interest of replication, we tested the hypothesis from Baird et al. (2012) that mind wandering during an incubation interval would improve divergent creativity by conducting a conceptual replication Our replication adopted three design features from Smeekens and Kane. First, we explicitly prompted participants to generate creative and useful responses to the AUT. Second, Baird et al. assessed AUT performance using uniqueness scores, where unique responses (relative to the entire set) are assigned a score of 1 (non-unique responses are assigned a score of 0). However, given criticisms of uniqueness scoring for measuring creativity (Silvia et al., 2008; Smeekens and Kane, 2016), we used subjective ratings of creativity generated by three independent raters (we also conducted exploratory analyses using alternative scoring systems to ensure that analyses were robust to different scoring procedures). Finally, we measured mind wandering using thought probes rather than a retrospective report. Baird et al. measured mind wandering with a single retrospective report collected at the end of the incubation-interval. However, this method is likely to be less reliable than *in situ* probes because accurately responding to a single retrospective measure requires significantly more working memory and recollection than accurately responding to a probe about where attention was directed over the previous few seconds.

Our studies more closely conceptually replicates Baird et al. (2012) relative to Smeekens and Kane (2016) or Steindorf et al. (forthcoming) for several reasons. First, neither Smeekens and Kane nor Steindorf et al. manipulate proportions of mind wandering during the incubation

interval and measure for effects of this manipulation; instead, they used a incubation-difficulty condition looking at predictive relations between proportion of mind wandering during an undemanding incubation-interval task and AUT responses. In this experiment, we use a task-difficulty manipulation from Baird et al. to manipulate proportions of mind wandering across conditions. Second, Smeekens and Kane did not measure the content of task-unrelated thoughts during the incubation-interval task. This is important, as thoughts about the AUT during the incubation interval would count as mind wandering according to the probes used by Smeekens and Kane because such thoughts are unrelated to the incubation interval task. However, such thoughts would *not* constitute incubation, as incubation about some non-focal task or problem precludes fixating on some task or problem. Neither of these are meant to indicate problems with Smeekens and Kane, as their studies examined more general relationships between executive control, working memory, and different forms of creativity, or Steindorf et al., who examined the role of different probing techniques on creative performance. Additionally, unlike either Baird et al. or Smeekens and Kane, we (a) collected data from a larger sample, (b) explicitly informed participants about the post-incubation AUT, which should encourage incubation effects, and; (c) used Bayesian analyses to supplement significance testing (these match aspects of the sampling, design, and analytic approach of Steindorf et al. (forthcoming)).

Study 1

Pre-registration of sample size, primary outcome measures, exclusion criteria, experimental materials (including experiment programs and stimuli), raw data, and analysis scripts can be found at https://osf.io/dwec2/. In accordance with the recommendations of Simmons et al. (2012), we report how we determined our sample size, all data exclusions, and all

measures in our study. All procedures were approved by the Duke University Internal Review Board.

Methods

Participants

We recruited 200 participants (M_{age} = 40.24, SD_{age} = 11.22, female = 104) through Amazon's Mechanical Turk. We recruited only participants located in the United States with a HIT approval rate greater than or equal to 98% and at least 5000 previously approved HITs. We determined our target sample size based on the results of an a priori power analyses conducted in the G*Power 3.1.9.7 software (Faul et al., 2007) for independent-samples *t*-tests with a medium effect size (d = 0.40), standard two-tailed alpha value ($p < .05$) at 80% power. Per the criteria described in the pre-registration, we excluded data from 6 participants based on their self-reported use of online resources while completing the AUT. We further excluded 8 participants who performed at chance or worse on the n-back (final N = 186).[1]

Materials

N-back Task. We induced differential proportions of mind wandering by implementing a between-groups n-back manipulation (Smallwood et al., 2011; Baird et al., 2012; Konishi et al., 2015; Smeekens & Kane, 2016; Brosowsky et al., Forthcoming). In this paradigm, participants are asked to respond to target stimuli and withhold responses to non-target stimuli. The stimuli consisted of eight digits (1-8) presented serially on-screen, with target stimuli displayed in red and non-target stimuli displayed in black. In the *0-back condition*, participants were asked to

[1] While this departs from our pre-registered exclusion criteria, none of our results change significantly when we included these 8 participants.

indicate whether the red digit was even or odd. In the *2-back condition*, participants were asked to indicate whether the digit presented two trials prior to the red digit was even or odd.

Participants completed 15 blocks of 16 trials (240 trials total). At the start of each trial, a fixation cross was presented in the center of the screen for 1500ms followed by a blank screen presented for 500ms. The target stimulus was displayed on screen for 1500ms followed by a blank screen for 500ms. If a response was not registered within the 1500ms window, the trial was counted as a miss. Participants completed one practice block containing 24 trials with 4 target stimuli. If participants responded to a non-target (false alarm), they were instructed to withhold responses to non-targets. If participants did not respond or responded incorrectly to a target (miss), they were instructed to respond according to the instructions. Participants had to respond correctly to 3 out of 4 targets to move on to the experimental trials. Participants were not given feedback on performance in the experimental trials. Target stimuli were evenly split between even and odd, so that performance at chance would be 50%. The ratio of target to non-target stimuli differed slightly between experimental and practice trials. There were 24 practice trials with 4 targets (1 target per 6 trials), while the experimental task had 240 trials with 30 targets (1 target per 8 trials).

Alternate Uses Task. The Alternate Uses Task (Guilford, 1967) is a prominent measure of divergent creativity (Plucker & Makel, 2010) that assesses individual ability to access semantically distant concepts relative to a mundane cue. Participants were provided the name of a single everyday object (“brick”) and asked to generate creative and unusual uses for the object. Three sample responses were provided. Additionally, participants were told to generate responses that are creative, useful, and specific to the object.

Thought Probes. At semirandom intervals, participants were presented with thought probes to assess the content of thought (Smallwood & Schooler, 2015; Weinstein, 2018). Probes were never presented immediately after a target stimulus. A thought-probe trial was presented once in every 16 trials (15 n-back trials/1 thought-probe trial), randomly presented between trials 7 to 11. The minimum time between probes was 48 seconds and the maximum time between probes was 80 seconds (64-second intervals, on average). These probes inquired about what the participant was thinking about just prior to seeing the probe with three options: (1) thinking about the even/odd task; (2) thinking about the upcoming creativity task, and; (3) thinking about something unrelated to the experiment. The second option was included because participants were explicitly instructed about the AUT to be completed. Reaction times were recorded for each thought probe response.

In keeping with the experimental procedures from Baird et al. (2012), we operationalized mind wandering in terms of task-unrelated thought. Thus, when participants reported either thinking about the upcoming creativity task or thinking about something unrelated to the experiment, these were recorded as mind wandering. This matches the procedures used in Smeekens & Kane (2016). However, because the hypothesis from Baird et al. is that mind wandering during an *incubation* interval improves AUT performance, we ran all analyses with a separate measure of mind wandering that included only the instances where participants reported thinking about something unrelated to the experiment.

Procedure

After reading instructions on the n-back task, AUT, and thought probes, participants practiced the n-back task and could not advance without responding correctly to three targets. Before beginning the experimental trials, participants were again told about completing the AUT

immediately after the n-back. Participants completed 15 minutes of n-back trials before doing the AUT for 2.5 minutes. A timer was visible for the AUT, indicating how much time was remaining (no timer was visible for the n-back task).

Creativity Scoring

Three raters (undergraduate research assistants trained in AUT scoring but blind to the hypotheses of the study and conditions) independently scored each participant's individual responses on a scale of 1-5. To assign scores, raters were told to assess responses on their novelty and creativity and to exclude nonsense answers with the following scoring system (full instructions are available at https://osf.io/dwec2/): 1 = very obvious/ordinary use; 2 = somewhat obvious use; 3 = non-obvious use; 4 = somewhat imaginative use; 5 = very imaginative/re-contextualized use; 99 = invalid (all invalid responses were excluded from final analyses). Raters exhibited strong reliability ($\alpha = 0.906$). Creativity scores were calculated by taking the arithmetic mean of all three ratings.

Results

We supplemented null hypothesis significance tests with Bayes Factor analyses using the BayesFactor package in R with default settings (Morey & Rouder, 2018). To simplify interpretation, we report the Bayes Factor in the direction the data supports (BF_{01} when there is more evidence in favor of the null over alternative hypothesis and BF_{10} when there is more evidence in favor of the alternative over null hypothesis). Bayes Factors are interpreted according to the proposal from Jeffreys (1961), with $BF > 3$ indicating moderate evidence, $BF > 10$ indicating strong evidence, $BF > 100$ indicating overwhelming evidence, and $BF < 3$ indicating anecdotal evidence (see Lee & Wagenmakers, 2013).

Task difficulty manipulation

All descriptive statistics for Study 1 are reported in Table 1. To assess the effectiveness of the task difficulty manipulation, we compared performance and proportion of mind wandering across conditions. Based on Baird et al. (2012), we hypothesized that participants in the 0-back condition would perform better and mind wander more than participants in the 2-back condition. Performance was calculated as the proportion of correct responses to total responses. Participants in the 0-back condition performed significantly better (M = .91, SD = 0.09, n = 95) than participants in the 2-back (M = .81, SD = 0.13, n = 91) condition ($t_{Welch}(159) = 5.94$, $p < .001$, $d = 0.88$, $CI[0.57, 1.17]$), with overwhelming evidence in favor of the alternative hypothesis over the null, $BF_{10} > 100$ (Figure 1).

Table 1. Descriptive statistics for Study 1

Condition	***Accuracy***	***MW Overall***	***MW Unrelated***	***AUT***	***Correlations***
0-back (n = 95)	M = .91(*.09*)	M = .39(*.30*)	M = .28(*.26*)	M = 2.53(*0.39*)	MW (Overall): -.11 (p = .30) MW (Unrelated) -.09 (p = .41)
2-back (n = 91)	M = .81(*.13*)	M = .18(*.18*)	M = .11(*.14*)	M = 2.48(*0.40*)	MW (Overall): -.11 (p = .28) MW (Unrelated): -.07 (p = .49)

MW Overall = proportion of mind wandering calculated using all thoughts unrelated to the n-back task. MW Unrelated = proportion of individual mind wandering calculated using thoughts unrelated to the n-back task or the AUT. AUT = average score per participants averaged across three independent ratings. Correlations = Spearman's rank correlation between AUT score and different measures of mind wandering.

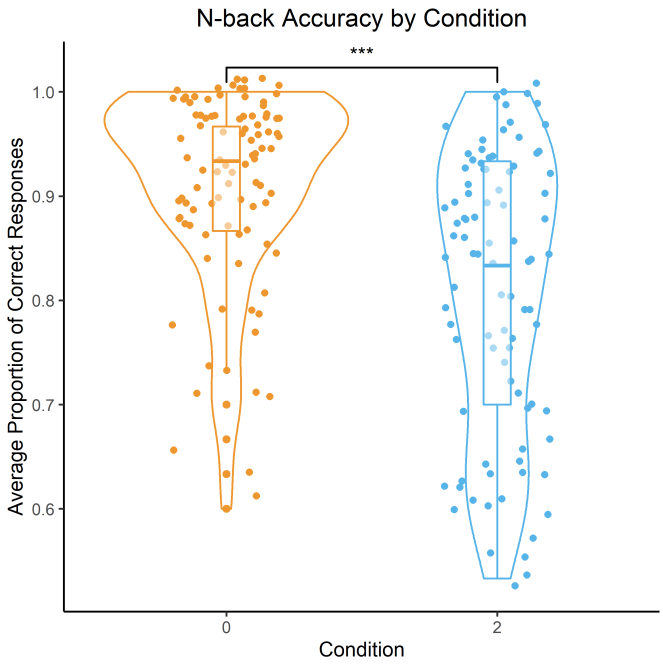

Figure 1. Proportion of correct responses between n-back conditions (0-back or 2-back)

As expected, participants reported significantly more mind wandering in the 0-back (39% of thought probes) relative to the 2-back (18% of thought probes) condition ($t_{Welch}(157) = 5.59$, $p < .001$, $d = 0.81$, $CI[0.52, 1.11]$), with overwhelming evidence in favor of the alternative hypothesis $BF_{10} > 100$ (Figure 2). When calculating mind wandering as thoughts that are unrelated either to the n-back task or upcoming AUT, participants in the 0-back group report significantly more mind wandering (28% of thought probes) relative to participants in the 2-back group (11% of thought probes) ($t_{Welch}(147) = 5.51$, $p < .001$, $d = 0.80$, $CI[0.51, 1.10]$), with overwhelming evidence in favor of the alternative hypothesis, $BF_{10} > 100$. Both performance and wind wandering measures resemble results found in other studies utilizing task difficulty manipulations to influence rates of mind wandering (Baird et al., 2011; Brosowsky et al., Forthcoming; Smeekens & Kane, 2016).

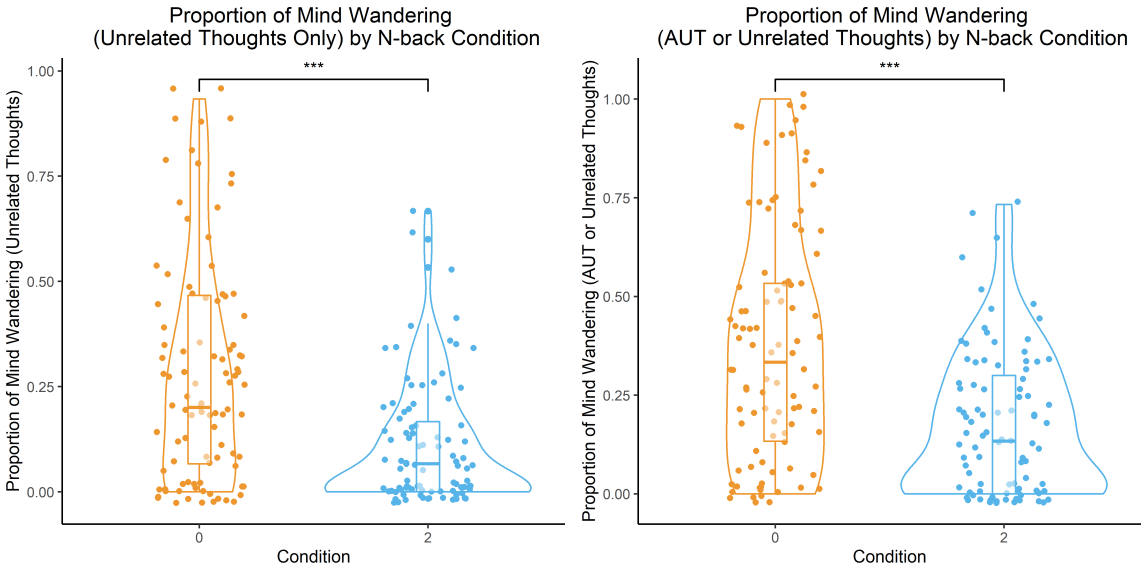

Figure 2. Proportion of mind wandering between n-back conditions (0-back or 2-back)

The relationship between mind wandering and creativity

To assess whether mind wandering is associated with AUT scores, we compared AUT performance across conditions. Participants in the 0-back group performed better on the AUT (M = 2.53, SD = 0.39) relative to participants in the 2-back group (M = 2.48, SD = 0.40), though an independent samples t-test indicated that this difference was not significant ($t(184) = 0.82$, p = .41, d = 0.12, CI[-0.17, 0.41]; see Figure 3). Bayesian analyses yielded modest evidence in favor of the null hypothesis $BF_{01} = 4.59$.

Figure 3. Average AUT scores (1-5) between n-back conditions (0-back and 2-back).

We computed correlations between AUT performance and different thought probe responses. A Shapiro-Wilk test of normality for average AUT scores indicated non-normal distribution (p = .02), so we calculated Spearman's rank correlation coefficients. For participants in the 0-back group, there was no significant correlation between AUT performance and mind wandering calculated as either overall mind wandering ($\log_e(S)$ = 11.97, p = .30, ρ = -.11, *CI*[-.30, .10], BF_{01} = 4.34) or thoughts unrelated to the n-back or AUT ($\log_e(S)$ = 11.95, p = .41, ρ = -.09, *CI*[-.28, .12], BF_{01} = 3.57) (see Figure 4). We also found no significant correlation between AUT performance and mind wandering calculated as either overall mind wandering ($\log_e(S)$ = 11.85, p = .28, ρ = -.11, *CI*[-.31, .09]) or thoughts unrelated to the n-back or AUT ($\log_e(S)$ = 11.15, p = .49, ρ = -.07, *CI*[-.27, .14]) for participants in the 2-back group. Notably, we found only anecdotal evidence in favor of the null hypothesis over the alternative hypothesis that there

is a significant correlation between AUT performance and overall mind wandering (BF_{01} = 2.14) or thinking that is not about the n-back or AUT (BF_{01} = 1.90).

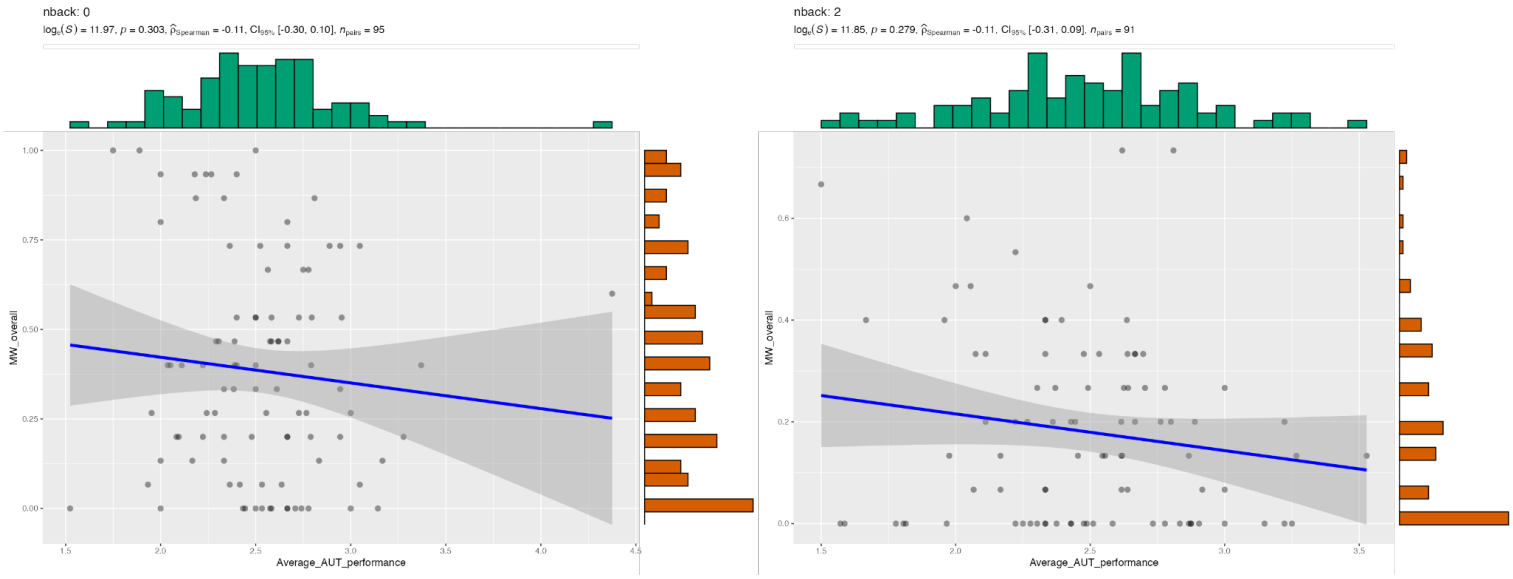

Figure 4. Correlation between average AUT score and individual proportion of mind wandering (calculated as thoughts that are unrelated to the n-back or AUT).

As in Baird et al. (2012), we split participants' data according to whether they registered at least one AUT-related thought during the n-back task or not. In the 0-back condition, 50 participants registered at least one AUT-related thought (45 participants registered none), whereas in the 2-back condition, 38 participants registered at least one AUT-related thought (53 participants registered none). In the 0-back group, participants who did not think about the AUT at all performed better on it (M = 2.56, SD = .32) relative to participants who thought at least once about the AUT (M = 2.50, SD = .45). In the 2-back group, participants who did not think about the AUT performed slightly better on it (M = 2.48, SD = .43) relative to those who thought about the AUT (M = 2.47, SD = .36).

A between-subjects ANOVA with n-back group and AUT-related thought as factors revealed no significant interaction between n-back group and AUT-related thought (F(1, 182) =

0.24, $p = .63$, $\eta_p^2 = .00$), and no main effects of n-back group ($F(1, 182) = .73$, $p = .39$, $\eta_p^2 = .00$) or AUT-related thought ($F(1, 182) = 0.43$, $p = .51$, $\eta_p^2 = .00$) with moderate evidence for the null hypothesis over the hypothesis that there is an effect of AUT-related thoughts on AUT score ($BF_{01} = 5.35$).

Exploratory Analyses

Finally, for exploratory purposes, we used a recently developed automated AUT scoring program (SemDis; Beaty & Johnson, 2020) to assess whether the predicted effect of mind wandering on creativity obtains when using automated creativity scores. SemDis computes the semantic distance between the cue and the participant response relative to a semantic space. The program enables computing semantic distance within five different semantic spaces (cbowukwacsubtitle_nf_m, cbowsubtitle_nf_m, cbowBNCwikiukwac_nf_m, TASA_nf_m, and glove_nf_m). We chose to analyze responses using automated ratings indexed to each of the five semantic spaces, as well as a composite score based on all five sets of ratings (SemDis_MEAN). We first wanted to assess correlations between subjective ratings and automated ratings. Because subjective and automated ratings are generated on different scales, we first normalized both sets of ratings using MinMaxScaler with the scikit-learn package in Python. The default setting is to map values to a [0, 1] range. However, given the 1-5 scale used by raters, we adjusted the parameters to map all values onto a common [1, 5] range. After normalization, the internal reliability between average AUT scores across all SemDis dictionaries and manually coded AUT responses was strong ($\alpha = .90$), We computed Spearman's rank correlation coefficients between manually coded AUT scores and automated AUT scores (see Figure 5). For most semantic spaces, we found moderate significant correlations between manually coded AUT scores and automated AUT scores.

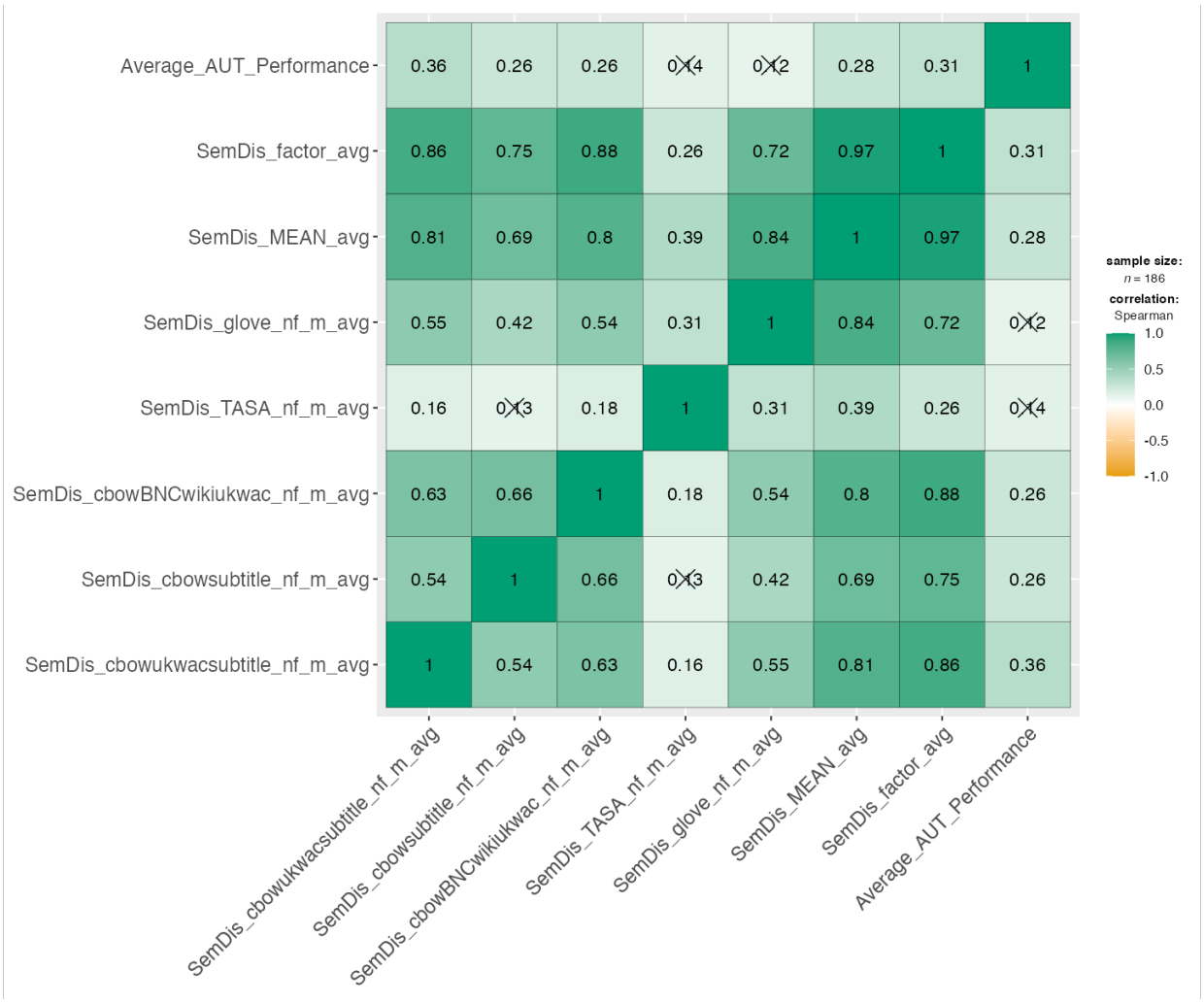

Figure 5. Correlations between manual and automated AUT scores. Correlation coefficients are Spearman's ρ, and boxes with an '×' indicate p > .05 *(p-values are corrected for multiple comparisons using Bonferroni correction).*

Independent-samples *t*-tests revealed no significant differences and moderate evidence for the null hypothesis over the alternative hypothesis that there are significant differences in AUT scores by n-back group (see Table 3).

Bayesian Independent Samples T-Test

	BF_{01}	error %
SemDis_cbowukwacsubtitle_nf_m_avg	3.78	8.27e-6
SemDis_cbowsubtitle_nf_m_avg	4.75	9.41e-6
SemDis_cbowBNCwikiukwac_nf_m_avg	5.45	1.01e-5
SemDis_TASA_nf_m_avg	4.62	9.26e-6
SemDis_glove_nf_m_avg	6.18	1.08e-5
SemDis_MEAN_avg	6.20	1.08e-5
SemDis_factor_avg	5.84	1.05e-5

Note. The alternative hypothesis is that AUT scores differ by n-back group. When the alternative hypothesis is changed to match the prediction from Baird et al. (2012) that AUT scores in the 0-back group are greater than AUT scores in the 2-back group, there is still moderate to strong evidence against the alternative hypothesis (all BF_{01} from 2.95 - 12.01).

Table 2. Bayesian independent samples t-test comparing automatically generated AUT scores by n-back group

To test the robustness of our results against different scoring procedures, we further analyzed AUT responses through a second automated scoring package (Mildner, 2020). The software calculates four conceptually unique facets of creativity: (1) fluency, (2) elaboration, (3) flexibility, and (4) originality. The composite AUT score for this package is computed as the arithmetic mean of z-scored subscale scores. The first three subscales are determined using function calls from the SpaCy natural language processing (NLP) library in Python while the

originality index is evaluated by *k*-means clustering using the scikit-learn package in Python. To approximate semantic similarity for the flexibility subscale, a further bootstrapping analysis was conducted to correct for the disproportionately higher similarity scores to be expected for longer words in each response: for words of length *n* characters, 10,000 words with a corresponding length of *n* characters were randomly sampled from a vocabulary corpus from a trained SpaCy English NLP pipeline to compare against them for similarity.

We first examined whether fluency scores differed by condition. Fluency is a measure of the number of items generated during the AUT. Participants in the 0-back group exhibited roughly equal fluency (M = 7.52, SD = 3.8) to participants in the 2-back group (M = 7.53, SD = 4.0) ($t(184) = -0.02$, $p = .98$ $d = -0.00$, $CI[-0.29, 0.29]$), with moderate evidence for the null hypothesis over the alternative hypothesis that there is a difference in fluency scores across conditions, $BF_{01} = 6.28$. We also found no significant difference by condition for elaboration ($p = .55$, $BF_{01} = 5.30$), flexibility ($p = .86$, $BF_{01} = 6.20$), or originality scores ($p = .98$, $BF_{01} = 6.28$), with moderate evidence for the null hypotheses over the alternative hypotheses that there are significant differences by condition for these scores.

To account for potential interactions with AUT-related thought, we conducted a series of between-subjects ANOVA with automatically generated fluency, elaboration, flexibility, and originality scores with n-back group and AUT-related thought as factors. There was no evidence for a significant interaction between n-back group and AUT-related thought for elaboration ($F(1, 182) = 0.19$, $p = .67$, $\eta_p^2 = .00$, $BF_{01} = 33.63$), flexibility ($F(1, 182) = 2.34$, $p = .13$, $\eta_p^2 = .01$, $BF_{01} = 39.19$), originality ($F(1, 182) = 0.13$, $p = .72$, $\eta_p^2 = .00$, $BF_{01} = 4.70$), and fluency ($F(1, 182) = 2.84$, $p = .09$, $\eta_p^2 = .02$, $BF_{01} = 3.59$), with moderate to very strong evidence for the null

hypotheses over the alternative hypothesis that there is a significant interaction of n-back group and AUT-related thought on automatically generated AUT scores.

Discussion

Study 1 attempted to replicate results reported in Baird et al. (2012) that an increased proportion of mind wandering during an incubation interval correlates with better AUT performance. We replicated the finding that task-difficulty manipulations modulate rates of mind wandering: harder tasks tend to elicit less overall mind wandering than easier tasks. However, unlike Baird et al., we did not find evidence for an effect of n-back group on AUT scores. Moreover, we did not find that individuals' proportions of mind wandering (based on two different ways of calculating proportions of mind wandering) were significantly associated with their AUT scores.

These results align with the failed conceptual replication reported in Smeekens and Kane (2016) and Steindorf et al. (forthcoming). However, the former did not measure the content of mind wandering during the incubation interval. Hence, Smeekens and Kane could not confirm that the n-back task functioned as an incubation period for the creativity task because they could not rule out participants explicitly thinking about the creativity task during the n-back. In this study, we split participants' responses according to whether they thought of the subsequent AUT at all during the n-back task or not. This allowed us to remove people who might not have been incubating. Moreover, neither group manipulated rates of mind wandering to measure for effects on AUT performance. While our task-difficulty manipulation successfully modulated mind wandering, we found no evidence for an effect on AUT performance.

Even after splitting the data among participants who thought about the AUT during the n-back task, we did not find any significant relationships between mind wandering and AUT

performance in the undemanding condition. This might seem surprising in one of two directions, so it's important to understand what this result indicates. On the one hand, isolating the people who did not think about the AUT at all during the n-back selects people for whom the n-back might function as an incubation interval. Thus, we should expect that isolating the potential incubators in the undemanding condition would generate the predicted effect of mind wandering on AUT performance. However, even after making these post-hoc distinctions, we did not find that individual proportion of mind wandering predicts subsequent AUT performance. On the other hand, it might seem surprising that people who report thinking about the AUT during the n-back do not perform better on it relative to those who don't. While this result appears surprising, it should be interpreted cautiously. We did not account for how much time one spent thinking about the AUT, so measure of AUT-related thought might not be fine-grained enough to reveal an effect of thinking on AUT performance. However, the number of AUT-related thoughts was so low that it seems unlikely that a more sophisticated tool would yield different results. Additionally, perseverative thinking might strengthen a narrow range of associations, resulting in constrained AUT responding. Thus, thinking more about the AUT likely might not facilitate better performance on it.

Study 2

Study 1 failed to replicate the effect of mind wandering during an incubation interval on AUT performance reported in Baird et al. (2012). However, our design differed from Baird et al. in a crucial respect, as participants did not complete an initial round of the AUT prior to the incubation interval. Without a pre-incubation task, there is nothing on which participants can incubate. In order to more closely replicate the design of Baird et al., we conducted another study

that included a pre-incubation AUT and a novel post-incubation AUT to compare post-incubation AUT performance between a novel and repeated prompt.

Methods

Participants

We recruited 284 participants (M_{age} = 42.1, SD_{age} = 12.5, female = 131) through Amazon's Mechanical Turk. We used the same screening criteria from Study 1. Target sample size was determined with an a priori power analysis conducted with G*Power for a linear regression with 4 predictors with 95% power to detect a medium effect size (f^2 = 0.15) at a one-tailed standard p-value threshold (< .05). The analysis indicated that 129 participants are needed per condition (258 total). We over-recruited by 10% to account for attrition and exclusions. Per the criteria described in the pre-registration, we excluded 27 participants based on either self-reported use of outside resources when completing the AUT, not completing part of the experiment, not following instructions, or failing to perform better than chance (50%) on the n-back (final N = 257).

Materials

All materials were the same as Study 1.

Procedure

After reading instructions on the n-back task, AUT, and thought probes, participants practiced the n-back task and could not advance without responding correctly to three targets. Before beginning the experiment, participants were prompted about completing another round of the AUT after the n-back task.

Participants were randomly assigned to complete an initial AUT for 2 minutes with either 'marble' or 'balloon' as the prompt. Participants were then randomly assigned to complete 12 minutes of either 0- or 2-back trials. Afterward, two AUTs (2 minutes each) were presented in a random order: one repeat AUT from the beginning of the experiment and one novel AUT (either 'marble' or 'balloon'). A timer was visible during AUT problems, indicating how much time was remaining (no timer was visible for the n-back task).

Creativity scoring

We used the same scoring procedure from Study 1. Raters exhibited strong reliability for both AUT items (balloon: $\alpha = .80$; marble: $\alpha = .86$). AUT scores for each participant were calculated by taking the arithmetic mean of all three ratings.

Results

Descriptive statistics for AUT scores are reported in Table 4.

Table 3. Descriptive statistics for AUT scores in Study 2

Condition	***Prompt***	***Pre-Incubation AUT***	***Post-Incubation Repeat AUT***	***Post-Incubation Novel AUT***
0-back (n = 131)	Balloon	M = 2.20(*.48*)	M = 2.28(*.55*)	M = 2.26(*.45*)
	Marble	M = 2.59(*.44*)	M = 2.66(*.49*)	M = 2.55(*.47*)
2-back (n = 125)	Balloon	M = 2.30(*.56*)	M = 2.40(*.57*)	M = 2.18(*.49*)
	Marble	M = 2.54 (*.48*)	M = 2.58(*.55*)	M = 2.51(*.50*)

Task difficulty manipulation

Descriptive statistics for task difficulty and mind wandering are reported in Table 5. To assess the effectiveness of the task difficulty manipulation, we compared performance and proportion

of mind wandering across conditions. Performance was calculated as the proportion of correct responses to total responses. Participants in the 0-back condition performed significantly better (M = .92, SD = .07, n = 131) than participants in the 2-back condition (M = .85, SD = .12, n = 125) ($t_{Welch}(210) = 6.25$, $p < .001$, $d = .76$) with overwhelming evidence in favor of the alternative hypothesis that there is a difference in accuracy by n-back group over the null, $BF_{10} > 100$ (Figure 6).

Table 4. Descriptive statistics for task difficulty and mind wandering in Study 2

Condition	***Accuracy***	***MW Overall***	***MW Unrelated***	***Post-incubation Repeat AUT***
0-back (n = 131)	M = .92(*.07*)	M = .44(*.33*)	M = .32(*.29*)	M = 2.48(*.55*)
2-back (n = 125)	M = .85(*.12*)	M = .13(*.19*)	M = .09(*.16*)	M = 2.51(*.56*)

MW Overall = proportion of mind wandering calculated using all thoughts unrelated to the n-back task. MW Unrelated = proportion of individual mind wandering calculated using thoughts unrelated to the n-back task or the AUT. Post-incubation Repeat AUT = average score per participant averaged across three independent ratings and collapsed across prompts for the repeated post-incubation AUT problem. Correlations = Spearman's rank correlation between post-incubation repeated AUT score and different measures of mind wandering.

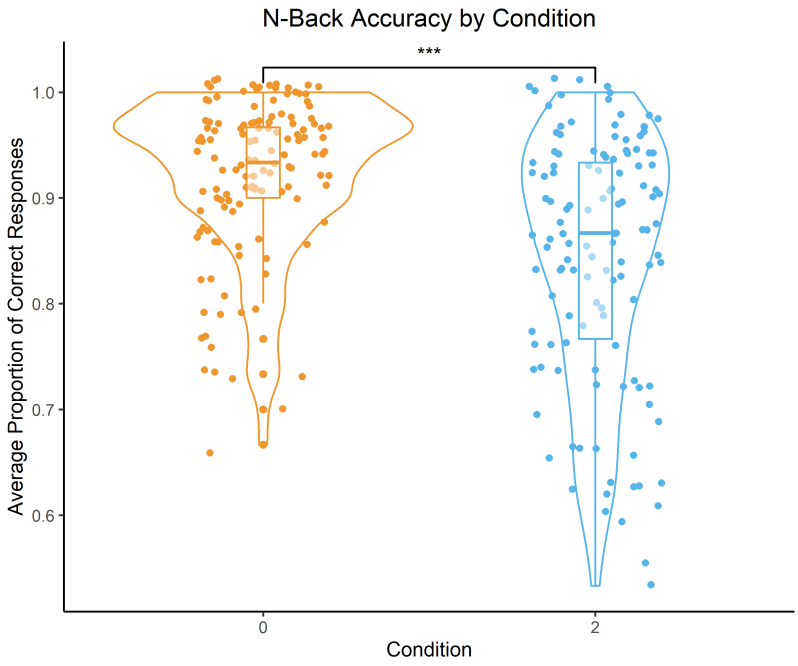

*Figure 6. N-back performance by condition. *** denotes significant difference between conditions ($p < .001$).*

Based on results from Study 1, we predicted that participants in the 0-back condition would report significantly more mind wandering than participants in the 2-back condition. To assess this prediction, we calculated the proportion of mind wandering in two ways. First, we calculated the overall proportion of mind wandering, which includes probe responses that indicate thinking either about the AUT or something unrelated to the n-back task. Second, we calculated the proportion of mind wandering only including probes that indicate thinking about something other than the n-back or AUT. According to overall proportion of mind wandering, participants reported significantly more mind wandering in the 0-back condition (44% of thought probes) relative to participants in the 2-back condition (13% of thought probes) ($t_{Welch}(208) = 9.33$, $p < .001$, $d = 1.16$) with overwhelming evidence in favor of the alternative, $BF_{10} > 100$. When isolating thoughts that are unrelated either to the n-back or the AUT, participants in the 0-back condition still report significantly more mind wandering (32% of probes) relative to the 2-

back condition (9% of probes) ($t_{Welch}(202) = 7.75$, $p < .001$, $d = 0.96$), with overwhelming evidence in favor of the alternative hypothesis, $BF_{10} > 100$ (Figure 7).

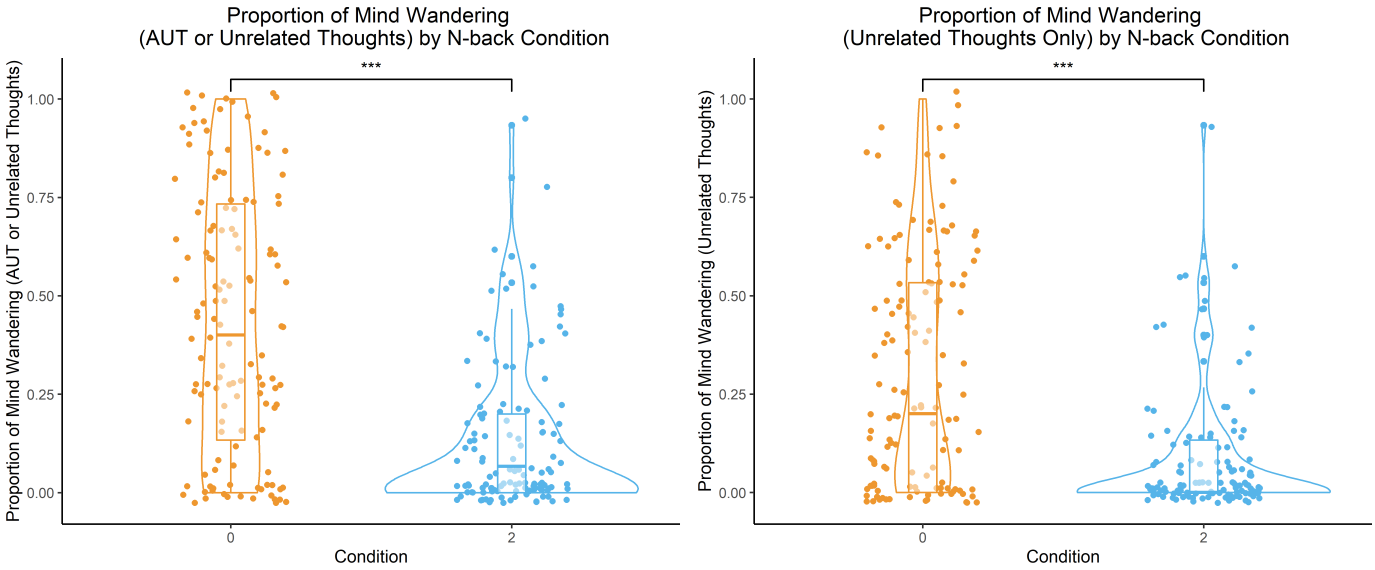

Figure 7. The left panel shows proportions of mind wandering calculated as including thoughts about the AUT by condition. The right panel shows proportions of mind wandering calculated as including thoughts unrelated to either the n-back or AUT by condition.

Both performance and mind wandering measures replicated the results from Study 1.

The relationship between mind wandering and AUT scores

To assess whether mind wandering during an incubation interval has an effect on AUT performance, we first conducted two ANCOVAs to assess effects of n-back group on post-incubation AUT scores when controlling for pre-incubation AUT scores. Additionally, because the hypothesis from Baird et al. (2012) is that mind wandering during an incubation-interval facilitates improved AUT performance, we also added AUT-related thought as a factor. There was a significant effect of pre-incubation AUT score on post-incubation repeated AUT problems ($F(1, 251) = 58.15$, $p < .001$, $\eta_p^2 = .19$). There was also a significant interaction between n-back group and AUT-related thought ($F(1, 251) = 4.38$, $p = .04$, $\eta_p^2 = .02$. There was no main effect of AUT-related thought ($F(1, 251) = 1.32$, $p = .25$, $\eta_p^2 = .01$) or n-back group ($F(1, 251) = 0.46$, $p =$

.50, η_p^2 = .00). Correcting for multiple comparisons, we found no significant differences in post-incubation AUT performance for repeated problems as a function of n-back group by AUT-related thought (all p > .22). In the ANCOVA for post-incubation novel AUT problems, we again found a significant effect of pre-incubation AUT score (F(1, 251) = 16.47, p < .001, η_p^2 = .06). There was no significant interaction between n-back group and AUT-related thought (F(1, 251) = 0.34, p = .56, η_p^2 = .00. There was no main effect of AUT-related thought (F(1, 251) = 2.68, p = .47, η_p^2 = .00) or n-back group (F(1, 251) = 0.46, p = .10, η_p^2 = .01).

We computed linear regressions to test models for predicting post-incubation AUT scores. A model for post-incubation repeated AUT problems that included pre-incubation AUT scores and overall mind wandering as covariates accounted for 18% of the variance in post-incubation AUT scores (F(2, 253) = 27.6, p < .001, R^2 = .18). However, overall mind wandering did not have significant partial effects in the model (β = 0.01, CI[-0.20, 0.21], p = .96), with moderate evidence for the null hypothesis over the alternative hypothesis that overall mind wandering has significant partial effects in the model (BF_{01} = 6.98). When the model included mind wandering that encompasses thoughts unrelated to the n-back or AUT, it accounted for 18% of the variance in post-incubation AUT scores (F(2, 253) = 27.8, p < .001, R^2 = .18) and mind wandering again did not have significant partial effects in the model (β = 0.07, CI[-0.17, 0.30], p = .59), with moderate evidence for the null hypothesis over the alternative hypothesis that overall mind wandering has significant partial effects in the model (BF_{01} = 6.95).

A model for post-incubation novel AUT problems that included pre-incubation AUT scores and overall mind wandering as covariates accounted for 7% of the variance in post-incubation novel AUT scores (F(2, 253) = 9.28, p < .001, R^2 = .07). However, overall mind wandering did not have significant partial effects in the model (β = 0.16, CI[-0.03, 0.35], p =

.10), with anecdotal evidence for the null hypothesis over the alternative hypothesis that overall mind wandering has significant partial effects in the model (BF_{01} = 2.91). When the model included mind wandering that encompasses thoughts unrelated to the n-back or AUT, it accounted for 7% of the variance in post-incubation AUT scores ($F(2, 253) = 9.45$, $p < .001$, $R^2 = .07$) and mind wandering again did not have significant partial effects in the model ($\beta = 0.20$, CI[-0.03, 0.43], $p = .08$), with anecdotal evidence for the null hypothesis over the alternative hypothesis that overall mind wandering has significant partial effects in the model (BF_{01} = 2.20).

Exploratory analyses

To explore for possible effects using automated AUT scoring procedures, we analyzed responses using automated ratings indexed to each of the five semantic spaces available in SemDis using the same normalization procedure as Study 1. For ease of presentation, we report results of analyses using SemDis_MEAN, which incorporates all scores across the different SemDis dictionaries (see Beaty and Johnson, 2020). After normalization, manually coded pre-incubation AUT scores and SemDis scores exhibited poor reliability (α = .15). However, we found a weak but significant correlation between manually coded pre-incubation AUT scores and SemDis scores ($\log_e(S) = 14.60$, $p < .001$, $\rho = .22$, CI[.10, .33], with overwhelming evidence in favor of the hypothesis that there is a significant correlation between manual and automatically scored pre-incubation AUT responses, $BF_{10} > 100$.

Manual and automatic scores of post-incubation AUT responses for repeated AUT problems exhibited low reliability (α = .11). However, we again found a weak but significant correlation between manually coded AUT scores and SemDis scores ($\log_e(S) = 14.70$, $p = .03$, $\rho = .13$, CI[.01, .25]), though there was anecdotal evidence for the null hypothesis over the

alternative hypothesis that there is a significant correlation between manual and automatically scored post-incubation AUT responses, $BF_{01} = 2.24$.

We conducted the same analyses using the average SemDis score to measure for effects with an alternative scoring system. We first conducted ANCOVAs with post-incubation SemDis score on repeated AUT problems with n-back group and AUT-related thought as factors and controlling for pre-incubation SemDis AUT score. There was a significant effect of pre-incubation SemDis AUT score on post-incubation SemDis score for repeated AUT problems ($F(1, 251) = 6.39$, $p = .01$, $\eta_p^2 = .03$). There was no significant interaction between n-back group and AUT-related thought ($F(1, 251) = 2.67$, $p = .10$, $\eta_p^2 = .01$. There was no main effect of AUT-related thought ($F(1, 251) = 0.01$, $p = .91$, $\eta_p^2 = .00$) or n-back group ($F(1, 251) = 0.55$, $p = .46$, $\eta_p^2 = .00$). In the ANCOVA for post-incubation SemDis score on novel AUT problems, we again found a significant effect of pre-incubation SemDis AUT score ($F(1, 251) = 4.32$, $p = .04$, $\eta_p^2 = .02$). There was no significant interaction between n-back group and AUT-related thought ($F(1, 251) = 0.33$, $p = .57$, $\eta_p^2 = .00$. There was no main effect of AUT-related thought ($F(1, 251) = 1.87$, $p = .17$, $\eta_p^2 = .01$) or n-back group ($F(1, 251) = 0.73$, $p = .39$, $\eta_p^2 = .00$).

We computed linear regressions to test models for predicting post-incubation SemDis AUT scores. A model for post-incubation SemDis scores for repeated AUT problems that included pre-incubation SemDis AUT scores and overall mind wandering as covariates accounted for 2% of the variance in post-incubation AUT scores ($F(2, 253) = 3.16$, $p = .04$, $R^2 = .02$). However, overall mind wandering did not have significant partial effects in the model ($\beta = 0.00$, $CI[-0.04, 0.04]$, $p = .95$), with moderate evidence for the null hypothesis over the alternative hypothesis that overall mind wandering has significant partial effects in the model ($BF_{01} = 7.12$). When the model included mind wandering that encompasses thoughts unrelated to

the n-back or AUT, it accounted for 3% of the variance in post-incubation SemDis AUT scores for repeated problems ($F(2, 253) = 3.30$, $p = .04$, $R^2 = .03$) and mind wandering again did not have significant partial effects in the model ($\beta = 0.01$, $CI[-0.04, 0.06]$, $p = .52$), with moderate evidence for the null hypothesis over the alternative hypothesis that overall mind wandering has significant partial effects in the model ($BF_{01} = 6.11$).

A model for post-incubation SemDis scores for novel AUT problems that included pre-incubation SemDis AUT scores and overall mind wandering as covariates accounted for 3% of the variance in post-incubation SemDis scores for novel AUT problems ($F(2, 253) = 4.28$, $p = .02$, $R^2 = .03$). However, overall mind wandering did not have significant partial effects in the model ($\beta = -0.03$, $CI[-0.07, 0.02]$, $p = .25$), with moderate evidence for the null hypothesis over the alternative hypothesis that overall mind wandering has significant partial effects in the model ($BF_{01} = 4.61$). When the model included mind wandering that encompasses thoughts unrelated to the n-back or AUT, it accounted for 3% of the variance in post-incubation SmeDis AUT scores for novel problems ($F(2, 253) = 4.28$, $p = .02$, $R^2 = .03$) and mind wandering again did not have significant partial effects in the model ($\beta = -0.05$, $CI[-0.10, 0.00]$, $p = .06$), with anecdotal evidence for the null hypothesis over the alternative hypothesis that overall mind wandering has significant partial effects in the model ($BF_{01} = 1.71$).

Discussion

In Study 2, we more closely replicated the original methods of Baird et al. (2012). We replicated our main results from Study 1. We found that while a task difficulty manipulation influenced the proportion of mind wandering experienced during an incubation interval, there was no evidence of an effect of condition for repeated AUT problems. This was robust to

different scoring procedures. Notably, in our exploratory analyses, we found that automatically scored AUT responses were only weakly associated with manually scored AUT responses, yet we still failed to find an effect of condition on AUT performance for repeated problems.

General Discussion

The effects of incubation on creativity are well-known (see Gilhooly, 2016). Incubation requires that one refrain from fixating on a problem and, during incubation, there are reasons to think that mind wandering can facilitate creative insights. This seems especially plausible for problems that require divergent creativity, which require cognitive exploration and forming novel associations. As mentioned in the Introduction, mind wandering facilitates forming novel associations and pattern learning (see Sripada, 2016). Why, then, do we fail to find any effects of task-unrelated thought during incubation on divergent creativity?

One issue concerns the theoretical operationalization of mind wandering as task-unrelated thought. Baird et al. (2012), following standard procedures in research on mind wandering, assessed mind wandering in terms of thoughts unrelated to a focal task. However, this category is too heterogeneous to function as a useful tool for measuring the kind of mind wandering that is likely beneficial, in some circumstances, for divergent creativity. Consider, for example, that the probes used by Baird et al. (and in our replication) rule *in* a wide range of thoughts as mind wandering. Someone might not be thinking about the n-back or AUT, but instead fixating on an upcoming test or recalling a stressful conversation with a loved one. This kind of fixation counts as mind wandering according to standard probes, but lacks the exploratory and unconstrained character that is characteristic of the kind of mind wandering that is likely beneficial for divergent creativity. The fact that task-unrelated thought encompasses both rumination (or

perseverative thinking) and unconstrained exploration makes the construct too unwieldy to reliably assess the relationship between mind wandering and divergent creativity.

This heterogeneity problem ramifies into a more general issue that applies both to our replication and to the experiment in Baird et al. (2012). During the n-back task, some participants might be fixating on concrete mental contents or episodes. This would not constitute incubation, because incubation requires a release from fixation. Thus, the observation of task-unrelated thought is not thereby evidence that a person is incubating. However, if it were the case that task-unrelated thoughts do not gravitate toward thinking about concrete mental episodes, then this would be good (though not definitive) evidence that the observation of task-unrelated thought indicates that a person is also incubating (holding fixed some other situational factors). In fact, though, a wealth of evidence indicates precisely the opposite: task-unrelated thought has a general prospective orientation and often takes planned behaviors as part of its content (Baird et al., 2011). Thus, the experience of task-unrelated thought is likely often incompatible with incubating, as thinking about planned behaviors is relatively fixated (see Irving, 2016). Thus, we have further reason to think that task-unrelated thought is not a good measure for assessing the relationship between mind wandering and divergent creativity.

This is not to imply that there is no interesting relationship between mind wandering and divergent creativity. We are claiming only that different measures of mind wandering are likely required to assess this relationship. In particular, measures that assess the dynamics of mind wandering are better able to capture the kinds of mind wandering that are likely related to divergent creativity. One suggestion that follows from our results is that researchers should shift from using content-based measures of mind wandering to constraint-based measures of mind wandering (Mills et al., 2018-b, based on the framework articulated in Christoff et al., 2016).

These measures have recently been shown to assess mind wandering episodes that are neurally dissociable from task-unrelated thought (Kam et al., 2021), and better capture the kinds of mind wandering that likely facilitate divergent creativity (see Girn et al., 2020).

Some have noted that the AUT might not be an adequate measure of divergent creativity (see Zeng et al., 2011). While there are legitimate issues surrounding the construct validity of the AUT, we believe our discussion extends beyond these familiar criticisms in two ways. First, it might be the case that AUT performance measures only a narrow dimension of divergent creativity. Even if this were the case, however, we would expect to find an effect of mind wandering on the dimension of divergent creativity that the AUT putatively measures. Insofar as we do not find these effects, we think our results raise concerns about the use of task-unrelated thought measures of mind wandering to assess divergent creativity in laboratory settings. Second, perhaps the AUT does not measure any facet of divergent creativity whatsoever. Though a possibility, our criticism raises a distinct issue: even with more valid measures of divergent creativity, there is reason to believe that task-unrelated thought measures of mind wandering are not useful for assessing the kinds of mind wandering that likely facilitate divergent creativity.

As noted in the Introduction, the connection between mind wandering and creativity plays an important role in views about the functional role and adaptive value of mind wandering. Because the study from Baird et al. (2012) was the first to experimentally manipulate mind wandering to modulate creativity, it became a cornerstone of such theoretical accounts. Sripada (2018, p. 25), for instance, cites only Baird et al. (2012) to support his claim that mind wandering is an exploratory mode of cognition (see Shepherd, 2019). Because of this failed replication (coupled with Smeekens and Kane, 2016), must we cede that these theoretical accounts are misguided?

We think that conclusion would be too rash. However, we do think some of the conceptual and methodological limitations noted above call for a more nuanced approach to studying the effects of mind wandering on creativity. One important methodological change is to reconsider the kind of creative benefits mind wandering is likely to generate. Potential effects of mind wandering on creativity are likely to appear for problems with some personal relevance to the individual (see Klinger, 2013). Hence, studies of mind wandering and creativity might consider using the Personal Concerns Inventory (Cox & Klinger, 1988) to assess areas of concern where people might experience creative breakthroughs. Second, researchers should reconsider the temporal scale along which effects of mind wandering on creativity are likely to occur. For Baird et al. (2012), mind wandering was supposed to influence creativity on an impersonal word association task within 15 minutes. However, breakthroughs on personally relevant problems or issues might require hours, days, or even weeks to occur. Hence, rather than using impersonal tests of creativity over short time scales, we suggest understanding creativity applied to the personal concerns of individuals over longer timescales. For this reason, we find the methodology of Gable et al. (2019) exemplary. They used experience-sampling techniques to assess the context, phenomenology, and content of thought during creative breakthroughs and how these moments of creativity related to the overall project. While this study focused on the creative insights of professional scientists and writers, it is also possible to use experience-sampling methods to assess the relationship between creativity, mind wandering, and everyday concerns. We see no reason to think that mind wandering plays a fundamentally different role in the psychic economy of academics than non-academics.

The results presented here indicate that the experience of task-unrelated thought during an incubation interval is not associated with improved AUT performance. Ultimately, the failure

to replicate Baird et al. (2012), coupled with the failed replications reported in Smeekens and Kane (2016) and Reinsdorf et al. (forthcoming), indicate that the results should no longer be used to support theories of mind wandering or hypotheses about mind wandering. As indicated in the Discussion, this makes sense. There are good reasons to think that the standard characterization of mind wandering (task-unrelated thought) used by Baird et al. is unlikely to bear interesting relationships to divergent creativity. In this way, the failure to replicate is unsurprising.

We should note some important limitations to the current research. We utilized a single measure (AUT performance) of one kind of creativity (divergent creativity) and a single measure (task-unrelated thought) of mind wandering. Different measures or different constructs might relate differentially to mind wandering, conceptualized either as task-unrelated thought or relatively unconstrained thinking. As noted in the Introduction, Leszcyzynski et al. (2017) found that task-unrelated thought during an incubation interval was associated with improved performance on a measure of convergent creativity. And Tan et al. (2015) found that task-unrelated thought during an incubation interval positively correlated with improvement on a subsequent round of the number reduction task. Thus, different tasks that measure different kinds of creativity might yield different results. However, both studies should be interpreted carefully. Tan et al. did not manipulate task-unrelated thought, and their thought probes assessed thought in terms of *drifting* away from the focal task. This means that participants might have been prompted to report on the dynamic characteristics of mind wandering that we argue above are more useful for exploring the relationship between mind wandering and divergent creativity. Leszcyzynski et al. did not separately assess whether task-unrelated thought was related to the upcoming creativity task or not (in their Study 3, when they do measure for these different kinds of content, they do not compute separate measures of mind wandering that include or exclude

thoughts related to the upcoming creativity task). Hence, it is possible that participants who exhibit greater improvement on the Compound Remote Associates Test simply thought more about the upcoming task rather than incubating. In all, more work is needed to carefully assess different measures and tasks to determine more precisely the relationship between different kinds of mind wandering and different kinds of creativity.

While alternative methods, outlined in the Discussion, might reveal interesting relationships between divergent creativity and mind wandering, that's a matter of what evidence there might be, not what evidence there is. This point bears repeating. Since 2017, the failed replications in Smeekens and Kane (2016) have been cited 93 times. Meanwhile, Baird et al. have steadily *increased* their citation rate, garnering over half of their citations (537) since 2017. We hope that our efforts induce some restraint among researchers looking to use the results from Baird et al.

Acknowledgements

This research was supported by a Charles Lafitte Foundation Undergraduate Research Grant to Nathan Liang through the Duke University Department of Psychology & Neuroscience. Thanks to Zachary Irving, Michael Kane and an anonymous reviewer for helpful feedback on the manuscript.

Author Contributions

SM and NL conceptualized the study, NB collected data, SM and NL performed analyses, wrote the manuscript, and reviewed the final manuscript.

References

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

Andrews-Hanna, J., Smallwood, J., and 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.

Baird, B., Smallwood, J., Mrazek, M. D., Kam, J. W. Y., Franklin, M. S., & Schooler, J. W. (2012). Inspired by Distraction: Mind Wandering Facilitates Creative Incubation. *Psychological Science*. https://doi.org/10.1177/0956797612446024

Baird, B., Smallwood, J., & Schooler, J. W. (2011). Back to the future: Autobiographical planning and the functionality of mind-wandering. *Consciousness and Cognition*, *20*(4), 1604–1611. https://doi.org/10.1016/j.concog.2011.08.007

Barbot, B. (2018). The Dynamics of Creative Ideation: Introducing a New Assessment Paradigm. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.02529

Beaty, R., & Johnson, D. R. (2020). *Automating Creativity Assessment with SemDis: An Open Platform for Computing Semantic Distance*. https://doi.org/10.31234/osf.io/nwvps

Brosowsky, N., Murray, S., Schooler, J. W., & Seli, P. (Forthcoming). Thought dynamics under task demands: Evaluating the influence of task difficulty on unconstrained thought.

Cheng, E. (2017). *Beyond Infinity: An Expedition to the Outer Limits of Mathematics* (1st Edition). Basic Books.

Christoff, K., Irving, Z. C., Fox, K. C., Spreng, R. N. & Andrews-Hanna, J. R. (2016) Mind-wandering as spontaneous thought: a dynamic framework. Nature Rev. Neurosci. 17, 718–731.

Conner, T. S., DeYoung, C. G., & Silvia, P. J. (2016). Everyday creative activity as a path to flourishing. *The Journal of Positive Psychology*, *13*(2), 181–189. https://doi.org/10.1080/17439760.2016.1257049

Cox, W. M., & Klinger, E. (1988). A motivational model of alcohol use. *Journal of Abnormal Psychology*, *97*(2), 168–180. https://doi.org/10.1037/0021-843X.97.2.168

Dietrich, A. (2019). Types of creativity. *Psychonomic Bulletin & Review*, *26*(1), 1–12. https://doi.org/10.3758/s13423-018-1517-7

Ellamil, M., Fox, K. C. R., Dixon, M. L., Pritchard, S., Todd, R. M., Thompson, E., & Christoff, K. (2016). Dynamics of neural recruitment surrounding the spontaneous arising of thoughts in experienced mindfulness practitioners. *NeuroImage*, *136*, 186–196. https://doi.org/10.1016/j.neuroimage.2016.04.034

Gable, S. L., Hopper, E. A., & Schooler, J. W. (2019). When the Muses Strike: Creative Ideas of Physicists and Writers Routinely Occur During Mind Wandering. *Psychological Science*, *30*(3), 396–404. https://doi.org/10.1177/0956797618820626

Gilhooly, K. J. (2016). Incubation and Intuition in Creative Problem Solving. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01076

Girn, M., Mills, C., Roseman, L., Carhart-Harris, R.L., and Christoff, K. 2020. Updating the dynamic framework of thought: creativity and psychedelics. *NeuroImage* 213:116726. doi: 10.1016/j.neuroimage.2020.116726.

Guilford, J. P. (1967). *The nature of human intelligence*. McGraw-Hill.
Hennessey, B. A., & Amabile, T. M. (2009). Creativity. *Annual Review of Psychology*, *61*(1), 569–598. https://doi.org/10.1146/annurev.psych.093008.100416
Irving, Z.C. (2016). Mind-wandering is unguided attention: Accounting for the 'Purposeful' Wanderer. *Philosophical Studies* 173:2, 547-71.
Jeffreys, H (1961). *Theory of probability*, 3rd Edn. Oxford: Oxford University Press.
Kajimura, S. and Nomura, M. 2016. Development of Japanese versions of Daydream Frequency Scale and the Mind Wandering Questionnaire. *Shinrigaku Kenkyu* 87, 79-88. PMID: 27180516.
Kam, J.W.Y., Irving, Z.C., Mills, C., Patel, S., Gopnik, A., and Knight, R.T. 2021. Distinct electrophysiological signatures of task-unrelated and dynamic thoughts. *Proceedings of the National Academy of Sciences* 118:4, 32011796118. https://doi.org/10.1073/pnas.2011796118.
Kenett, Y. N., Kraemer, D. J. M., Alfred, K. L., Colaizzi, G. A., Cortes, R. A., & Green, A. E. (2020). Developing a neurally informed ontology of creativity measurement. *NeuroImage*, *221*, 117166. https://doi.org/10.1016/j.neuroimage.2020.117166
Klinger, E. (2013). Goal Commitments and the content of thoughts and dreams: basic principles. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00415
Konishi M, McLaren DG, Engen H, and Smallwood J (2015) Shaped by the Past: The Default Mode Network Supports Cognition that Is Independent of Immediate Perceptual Input. PLOS ONE 10(6): e0132209. https://doi.org/10.1371/journal.pone.0132209
Lee, M.D. and Wagenmakers, E.J. 2014. *Bayesian cognitive modeling: A practical course.* Cambridge: Cambridge University Press.
Leszczynski, M., Chaieb, L., Reber, T.P. *et al.* Mind wandering simultaneously prolongs reactions and promotes creative incubation. *Sci Rep* 7, 10197 (2017). https://doi.org/10.1038/s41598-017-10616-3
McCormick, C., Rosenthal, C. R., Miller, T. D., & Maguire, E. A. (2018). Mind-Wandering in People with Hippocampal Damage. *Journal of Neuroscience*, *38*(11), 2745–2754. https://doi.org/10.1523/JNEUROSCI.1812-17.2018
Mednick, S. (1962). The associative basis of the creative process. *Psychological Review* 69, 220–232.
Mildner, J. (2020). Unusual Uses Analysis. GitHub. https://github.com/jnmildner/unusual-uses-analysis
Mills, C., Herrera-Bennett, A., Faber, M., & Christoff, K. (2018-a). Why the Mind Wanders. *The Oxford Handbook of Spontaneous Thought*. https://doi.org/10.1093/oxfordhb/9780190464745.013.42
Mills, C., Raffaelli, Q., Irving, Z.C., Stan, D., and Christoff, K. 2018-b. Is an off-task mind a freely-moving mind? Examining the relationship between different dimensions of thought. *Consciousness and cognition* 58, 20-33.
Morey, R. D., & Rouder, J. N. (2018). *BayesFactor: Computation of Bayes Factors for Common Designs* (0.9.12-4.2) [Computer software]. https://CRAN.R-project.org/package=BayesFactor
Nadler, R. T., Rabi, R., & Minda, J. P. (2010). Better Mood and Better Performance: Learning Rule-Described Categories Is Enhanced by Positive Mood. *Psychological Science*. https://doi.org/10.1177/0956797610387441

Plucker, J. A., & Makel, M. C. (2010). Assessment of creativity. In *The Cambridge handbook of creativity* (pp. 48–73). Cambridge University Press. https://doi.org/10.1017/CBO9780511763205.005

Shepherd, J. (2019). Why does the mind wander? *Neuroscience of Consciousness, 2019*(1), Article niz014. https://doi.org/10.1093/nc/niz014

Silvia, P. J., Winterstein, B. P., Willse, J. T., Barona, C. M., Cram, J. T., Hess, K. I., Martinez, J. L., & Richard, C. A. (2008). Assessing creativity with divergent thinking tasks: Exploring the reliability and validity of new subjective scoring methods. *Psychology of Aesthetics, Creativity, and the Arts 2*(2), 68–85. https://doi.org/10.1037/1931-3896.2.2.68

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). *A 21 Word Solution* (SSRN Scholarly Paper ID 2160588). Social Science Research Network. https://doi.org/10.2139/ssrn.2160588

Smallwood, J., Schooler, J.W., Turk, D.J., Cunningham, S.J., Burns, P., and Macrae, C.N. 2011. Self-reflection and the temporal focus of the wandering mind. *Consciousness and Cognition* 20:4, 1120-26.

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

Smeekens, B. A., & Kane, M. J. (2016). Working Memory Capacity, Mind Wandering, and Creative Cognition: An Individual-Differences Investigation into the Benefits of Controlled Versus Spontaneous Thought. *Psychology of Aesthetics, Creativity, and the Arts*, *10*(4), 389–415. https://doi.org/10.1037/aca0000046

Sripada, C.S. (2016). Imaginative guidance: A mind forever wandering. In *Homo Prospectus* (Oxford: Oxford University Press), 103-31.

Sripada, C. S. (2018). *An Exploration/Exploitation Trade-off Between Mind-Wandering and Goal-Directed Thinking*. The Oxford Handbook of Spontaneous Thought. https://doi.org/10.1093/oxfordhb/9780190464745.013.28

Steindorf, L., Hammerton, H. A., & Rummel, J. (forthcoming). Mind wandering outside the box—About the role of off-task thoughts and their assessment during creative incubation. *Psychology of Aesthetics, Creativity, and the Arts.* Advance online publication. https://doi.org/10.1037/aca0000373

Tan, T., Zou, H., Chen, C., and Luo, J. 2015. Mind wandering and the incubation effect in insight problem solving, *Creativity Research Journal* 27:4, 375-82.

Weinstein, Y. 2018. Mind-wandering, how do I measure thee with probes? Let me count the ways. *Behavior Research Methods* 50:2, 642-61.

Yamaoka, A. and Yukawa, S. 2020. Mind wandering in creative problem-solving: relationships with divergent thinking and mental health. *PLoS ONE* 15(4): e0231946. https://doi.org/10.1371/journal.pone.0231946.

Zeng, L., Proctor, R.W., and Salvendy, G. 2011. Can traditional divergent thinking tests be trusted in measuring and predicting real-world creativity? *Creativity Research Journal* 23:1, 24-37.