Social desirability bias obscures mind-wandering reports

Shivang Shelat[1,2], Alexis R. Clubb[1], Barry Giesbrecht[1,2]*, Jonathan W. Schooler[1]*

[1]Department of Psychological and Brain Sciences, University of California, Santa Barbara, Santa Barbara, CA 93106, USA

[2]Institute for Collaborative Biotechnologies, University of California, Santa Barbara, Santa Barbara, CA 93106, USA

*These authors share senior authorship.

Correspondence:
Shivang Shelat (sshelat@ucsb.edu)
Psychology East, Santa Barbara, CA 93117

Abstract

The study of mind-wandering relies on the assumption that people accurately introspect and report on their attentional state. Yet self-reports may be distorted by extraneous motives, such as a desire to appear attentive. Across two sustained attention tasks, we demonstrate that social desirability bias is systematically linked with task-unrelated thought (TUT). In Experiment 1, higher social desirability bias scores predicted fewer subjective TUT reports during a reaction time task, despite no relationship with objective response times (RTs). This suggests a self-presentation effect rather than true attentional differences. Time pressure during thought probes further weakened RT–TUT coupling. In Experiment 2, controlling for social desirability bias enhanced the ability of an objective, RT-based real-time triggering procedure to predict deeper subjective TUT reports in the Sustained Attention to Response Task (SART). The alignment between objective triggers and subjective reports was strong for low-bias individuals, but weakened for those high in bias. Triggers were also less effective at capturing working memory lapses in participants prone to deliberate mind-wandering, suggesting that intentional off-task thinking manifests differently in behavior. The results show that social desirability bias can obscure TUT reports and should be considered in a science of attention and thought.

Keywords: mind-wandering, sustained attention, introspection, social desirability bias

Introduction

The science of mind-wandering relies on the accurate introspection and truthful report of thoughts during cognitive tasks. However, self-reports of attentional states may be confounded by a general unwillingness to admit the occurrence of mind-wandering. Such biases might be broadly present in work that measures lapses using self-report. Although subjective reports are limited, they remain valuable when studying phenomenological states and when coupled with complementary objective measures such as accuracy and response times (RTs; Schooler & Schreiber, 2004; see also Corneille & Gawronski, 2024). The thought probe method that prompts participants to report on qualities of internal experience is ubiquitous in the field of mind-wandering (Smallwood & Schooler, 2015). This method has contributed to a now-burgeoning field with implications for consciousness, creativity, sustained attention, and more.

Although the field has surged, more foundational work on the fundamental methods remains an utmost priority (Kane et al., 2021). The goal in the present study is to assess whether *social desirability bias* (SDB) is linked with self-reports of mind-wandering (i.e., task-unrelated thoughts, or TUTs; Smallwood & Schooler, 2006) during sustained attention tasks. SDB refers to a tendency to self-present desirable traits and underreport undesirable ones (Nederhof, 1985). To address this bias, researchers often include scales with lure items, particularly for stigmatized constructs (e.g., psychopathology; Crowne & Marlowe, 1960). For example, the original Marlowe-Crowne social desirability scale was vaguely titled “Personal Reaction Inventory” and had items such as “I have never intensely disliked anyone” and “I'm always willing to admit it when I make a mistake.” It is highly improbable that any given person embodies all these virtuous traits, so a high score indicates greater levels of SDB. Extensive discussion has been dedicated to ways of coping with SDB in psychological science (Nederhof, 1985), and one method recommended by Larson (2019) is to include SDB scores in statistical models (e.g., Garg et al., 2025a).

Surprisingly, no study to our knowledge has assessed the role of SDB in state reports of TUT. Since mind-wandering is often responsible for performance failures in several domains (Mooneyham & Schooler, 2013), it may be perceived as an undesirable experience, leading a participant to underreport it to appear favorable to themselves or to the experimenter. Indirect evidence comes from Zhang et al. (2015), who administered the Sustained Attention to Response Task (SART; Robertson et al., 1997) with thought probes to a sample of children while measuring attitudes toward mind-wandering using a questionnaire (e.g., “My daydreaming is good for my planning for the future.”). Those with positive attitudes toward mind-wandering

showed no coupling between objective SART indices and subjective TUT reports, whereas those with negative attitudes did. Moreover, although those with positive attitudes reported more TUT overall, their behavioral performance was comparable to those with negative attitudes. This suggests that valid self-reports of attentional state are obscured by attitudes toward mind-wandering's functionality for human cognition.

If SDB is tied to dishonesty (e.g., Leisge et al., 2024), then manipulations that promote honesty might provide a test of its influence on TUT reports. Vinski & Watter (2012) tested the effects of an implicit honesty prime to encourage truthful TUT self-reports in a subsequent SART. One group received a prime in which they selected synonyms of honesty-related words. For example, if the word was "genuine", they would have to select across "real", "straightforward", and "true." The other group received a control task with words such as "common" (having to pick from "frequent", "routine", "average"). The results suggested that the honesty prime actually reduced TUT, meaning that TUT might be inflated when participants are dishonest due to demand characteristics. The prime also improved coupling between SART performance and TUT reports. Nonetheless, this needs more rigorous consideration since Vinski & Watter used a small sample and an "honesty-enhancing" technique that has been questioned in later literature (Pashler et al., 2013).

Most relevant to the current research, a psychometric analysis of questionnaires designed to assess day-to-day tendencies for deliberate and spontaneous mind-wandering revealed that trait SDB was negatively correlated with dispositional spontaneous mind-wandering (Marcusson-Clavertz & Kjell, 2018). In other words, individuals who claimed more desirable traits were also less likely to report spontaneous TUT. This pattern can be interpreted in two ways. On one hand, those who are prone to attentional failures may also tend to engage in undesirable behaviors; on the other hand, high-SDB individuals may just be less likely to admit that their attention fails them in daily life. State reports collected during tasks, especially when paired with standard objective performance markers, offer a way to contrast these explanations: if SDB covaries with subjective TUT but not with behavioral lapse indices, this favors a reporting-bias explanation over any actual attentional differences. This collection of evidence is unclear, and to date, there is no published literature examining the link between SDB and phenomenological *state* reports of TUT.

While SDB has historically been measured using standardized surveys, it is possible that it can be causally amplified by applying time pressure. Protzko et al. (2019) showed that SDB scale scores increased when participants were forced to respond in just a few seconds. Therefore, one plausible method to increase SDB in TUT reports is to add a rapid deadline cue

that exerts time pressure during thought probes. Alternatively, pressure might decouple TUT and objective task performance through some separate mechanism (e.g., not offering enough time to effectively introspect), or might even improve coupling if it benefits metacognitive accuracy (Stone et al., 2025). It is informative to test whether time pressure shapes objective–subjective coupling of attention measures and whether it alters the role of SDB in TUT reports.

Besides SDB, trait differences in mind-wandering and motor impulsivity may contribute to associations between objective and subjective measures of attention. General proneness for spontaneous vs. deliberate TUT have different contributions to how lapses manifest in behavior (Seli et al., 2013), and similarly, impulsivity is argued to underpin behavioral patterns in some attention tasks (Helton, 2009). One of our background goals was to also explore the degree to which such traits moderate the predictive utility of objective indices like RT-based triggers for subjective attention reports. Because dispositional spontaneous TUT represents a tendency to unintentionally lapse whereas deliberate TUT represents a tendency to intentionally disengage, their trait-level associations with behavioral variability may differ. Likewise, higher motor impulsivity might reduce the usefulness of RT fluctuations since fast responses could be explained more by an impulsive response style than by waning attention.

The present study

Here, we conducted two experiments to test the role of SDB in mind-wandering reports across two attention tasks. The central goal was to move beyond prior trait-level correlations (Marcusson-Clavertz & Kjell, 2018) and use state reports combined with objective performance to distinguish whether SDB represents actual differences in attention or just presentation biases. If SDB tracks subjective TUT but not behavioral markers of TUT, this would point to the latter.

In Experiment 1, participants completed a boring sustained attention task where they reacted to target dots after long interstimulus intervals (ISI) and reported their TUT either under time pressure or not. Because slow reactions to targets in this task are a lapse marker (Nedergaard et al., 2023), we were able to ask whether trait SDB survey scores predicted fewer TUT reports even when RTs remained unchanged. We hypothesized that (1) SDB and time pressure would affect TUT reports, (2) SDB would moderate the effect of time pressure on TUT reports, and (3) SDB would moderate time pressure's decoupling of TUT reports from objective performance (RTs).

In Experiment 2, we leveraged recent methodological developments with a real-time RT speed-based triggering procedure (see Shelat et al., 2024 for a review) to better sample high and low attention states on the fly during a SART. Prior work has shown that relatively fast RTs

in this task precede reduced memory encoding (deBettencourt et al., 2018) and TUT reports (Franklin et al., 2011; Smallwood et al., 2008). In addition to thought probes, we triggered the insertion of working memory (WM) reports to replicate prior work that sustained attention and WM encoding lapse concurrently (deBettencourt et al., 2019). This enables a precise test of whether SDB alters the alignment between objective lapses and subjective reports when attention is sampled at its extremes. Besides the SDB survey from Experiment 1, participants filled out other individual difference surveys so we could examine the predictive utility of the triggering procedure across traits. We hypothesized that (1) fast (vs. slow) triggers would capture reduced WM encoding and deeper TUT, (2) SDB would affect TUT reports, (3) SDB would moderate the association between TUT reports and objective performance (fast/slow triggers), and (4) other individual differences (trait spontaneous/deliberate mind-wandering, motor impulsivity) would shape the coupling between TUT and objective performance (exploratory).

In both experiments, the hypotheses were partially supported. In Experiment 1, SDB was associated with a reduced likelihood of reporting TUT but not with RTs. This pattern is consistent with biased reporting rather than true differences in attentional state. There was no evidence that SDB moderated RT–TUT coupling nor time pressure's effect on RT–TUT coupling. In Experiment 2, triggers predicted flux in WM encoding but did not predict TUT until SDB was entered into the statistical model. SDB had no overall association with TUT depth, but it moderated triggers' ability to predict TUT such that high-bias individuals had worse objective–subjective coupling. Interestingly, triggers did not work well in catching WM encoding for participants prone to deliberate mind-wandering. We demonstrate that a state-level approach reveals that the fidelity of introspective thought reports regarding mind-wandering is partially dependent on biases related to social desirability and not true dynamics of attention.

Experiment 1

Methods

Participants

To identify an appropriate sample size at power ≈ .975 and α = .05, we performed a simulation-based power analysis on pilot data (n = 41; discarded). We modeled RT–TUT coupling (the correlation per participant between RT and TUT) by SDB scores × speeded group assignment using a linear model. We generated 10,000 simulations using the pilot sample's estimates for the intercept, main effects, interaction term, and residual error. For each simulated dataframe, continuous predictor values for SDB scores were drawn from a normal distribution based on the pilot mean and standard deviation. The categorical variable (speeded group) was

randomly assigned. The outcome, RT–TUT coupling, was computed using the pilot estimates and a normally distributed error term. We then fitted the same linear model to each simulated dataset and tallied the number of times that the interaction term p-value was < .05. We calculated the power as the proportion of simulations yielding a significant interaction effect. Power was .9767 for a sample of 97 participants. To account for potential exclusions and ensure high power, we doubled and rounded the intended *n* to 200.

A total of 200 pre-exclusion participants completed the online experiment on Prolific (www.prolific.com). They were paid at a rate of US$9 per hour. To be eligible for the study, participants were required to be 18 years of age or older, have normal or corrected-to-normal color vision, and have access to the internet and a computer. We applied preregistered exclusion criteria in two steps. First, we removed those who admitted that they did not take the study seriously (*n* = 0), failed an attention check (*n* = 1), or had a high omission error rate (5 or more misses; *n* = 33). Second, we removed participants whose average probe RT exceeded three SD from the mean probe RT of their respective condition (*n* = 1 from the speeded group; *n* = 3 from the unspeeded group). The final sample consisted of 162 participants (M_{age} = 36.1 ± 10.8 years; 80 female, 79 male, 3 other). All procedures were approved by the Institutional Review Board at the University of California, Santa Barbara.

Materials

Sustained attention task. We administered 3 practice trials and 20 main trials of a target reaction task (Figure 1) with long ISIs, based on the paradigm in Nedergaard et al. (2023). Each trial began with the presentation of a grey fixation cross at the center of the screen. The fixation cross remained onscreen for a duration determined via random sampling without replacement from 5, 10, or 15 seconds (s) for the practice block and 20, 25, 30, 35, or 40s four times each for the main block. Following the fixation period, a circular black target appeared at the center of the display. Participants were instructed to click the target as quickly as possible once it appeared. After a click, the target turned from black to grey to indicate that the response was received. Participants had 1.5s to make their click before the target dot disappeared. The grey response feedback was presented for 0.5s. After, there was a short 100 millisecond (ms) post-feedback interval before a thought probe appeared. The thought probe asked participants to categorize their thoughts as “on-task” or “off-task” in the moments before the last target appeared. Both of these radio buttons were equally distanced from the midpoint of the target dot in order to reduce the likelihood of participants clicking the nearest option on the display when under time pressure.

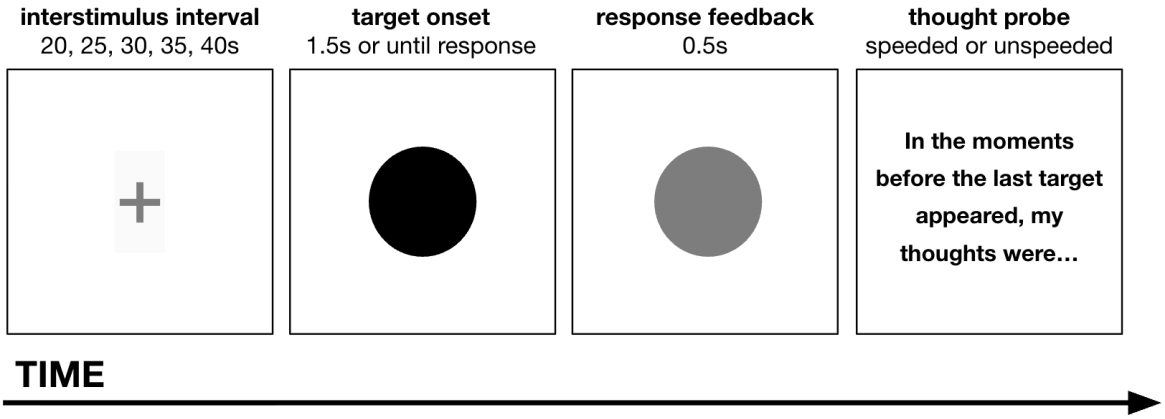

Figure 1. A schematic of a trial sequence in Experiment 1's target reaction task. After a long ISI, a dot appeared and participants had to click it as fast as they could. After a click, the dot turned grey to communicate to the participant that their response was registered. Participants then responded to a speeded or unspeeded binary thought probe.

In the speeded condition, participants saw a black bar rapidly ticking down over 1000ms. When the bar was empty, "RESPOND!" was presented in bold text. Participants were not prevented from responding after the deadline cue. In the unspeeded condition, the bar was present but did not tick down. This was to approximately equate visual load across conditions during the thought report.

We determined the 1000ms deadline cue by reanalyzing probe RTs from two large prior datasets. We analyzed the 5th percentile of participants' median probe RTs because it reflects the fastest typical responses while ensuring most participants would experience the deadline as pressuring. Martindale et al. (2023) had a 5th percentile median of 1137.2ms with a binary probe, and Kane et al. (2016) reported 1287.5ms with a content-based probe. Since both values were just above 1000ms, we chose it as a round, conservative cutoff that would reliably impose time pressure in the speeded group.

Oath. We administered an oath right before the main trials of the sustained attention task to increase the participants' adherence to the speeded thought probe response instructions. Based on a large-scale study by Zickfeld & colleagues (2025) on the efficacy of oaths, we designed ours to orient participants toward a specific behavior. Participants were directed to read the following statement carefully and type it in the box below to commit to the statement; they were allowed to proceed to the next page without committing if they wanted, although we expected rates of non-commitment to be low based on Zickfeld et al.'s results. In the speeded condition, participants were prompted to type out "I hereby declare that I will respond to the

thought probes quickly, in less than 1 second." In the control condition, they typed out "I hereby declare that I will respond to the thought probes, taking however much time I need." We disabled the ability to copy and paste the oath into the textbox.

Social desirability bias survey. The Marlowe-Crowne Social Desirability (MCSD) Short Form was developed by Vésteinsdóttir et al. (2017) to measure SDB (Crowne & Marlowe, 1960). Participants rated 10 true/false statements, such as "No matter who I'm talking to, I'm always a good listener." A total score was computed by summing all items together, where true equals 1 and false equals 0. A high score indicates greater biases related to positive self-presentation. We also embedded an attention check after the 6th item that read "I will press True to show that I'm paying attention."

Seriousness check. At the end of the experiment, participants were presented with a seriousness check (Aust et al., 2013). Participants were told "It would be very helpful if you could tell us at this point whether you have taken this experiment seriously, so that we can use your data for our scientific analysis." They selected either "I did not participate seriously, please throw my data away" or "I participated seriously."

Procedure

Participants first acknowledged an informed consent form. Then, they were introduced to the sustained attention task and given a definition of mind-wandering. We adapted our definition from Kane et al. (2021): "After each dot, you will be interrupted by a probe that asks you to indicate if you were experiencing task-unrelated thoughts (i.e., mind-wandering) just before the dot appeared on the screen. It is perfectly normal to think about things that are not related to the task. For instance, you may think about off-task things such as something you did recently or will be doing later, your current emotional or physical state, personal worries, daydreams, or your external environment. Please try your best to honestly assess your thoughts and choose a response that best describes your thoughts at the time when we ask." Participants were then randomly assigned to speeded vs. unspeeded groups and given condition-specific instructions. Those in the speeded group saw a replayable example of the 1s countdown bar to get an idea of the deadline cue.

Next, participants completed a short quiz to ensure that they understood the instructions. The quiz repeated if participants got any question wrong. The first question was "Before the target dot appeared, I was mentally planning a picnic. I was…" with options "thinking about task-unrelated things" (correct) and "focused on the task." The second question was "When should I click the dot?" with options "I will click it once, right before it appears", "I will click it once, right after it appears" (correct), "I will click it twice, both times after it appears", and "I will

keep clicking, even when it is not there." Participants affirmed that they were in a quiet room, free from distractions, and would not leave the page during the experiment. They completed practice trials of the attention task.

Just before the main task trials, participants were reminded of all instructions. They were given the opportunity to take an oath to adhere to their condition-specific probe response instructions (see Materials). Participants performed the task, filled out the SDB questionnaire, provided their demographic information, and responded to the seriousness check.

Results

In some cases, we z-scored target dot RT within participants before specifying the models; this is referred to in the text as RTz, whereas the non-z-scored RT is called RT. RTs to the thought probes are explicitly labeled as such to avoid confusion with RTs to the target dots. All logistic regression models were fitted using a bobyqa optimizer with a maximum of 100,000 function evaluations. For all mixed-effect models, we assigned subjectID as a random intercept to account for within-participant variability in the dependent variable; this random effects structure was predetermined to avoid increasing the complexity of the models.

Task checks

To verify that the paradigm was measuring sustained attention, we performed a series of unregistered task validity checks. First, we expected evidence that attention wanes over time (due to the vigilance decrement; Mackworth, 1948; Hemmerich et al., 2025; Skinner & Giesbrecht, 2025; Schwartzman et al., 2025), which should manifest as progressively slower RTs and greater TUT over time (Zanesco et al., 2024). A linear mixed-effects regression confirmed that RTs grew slower over trials ($b = 1.54$, $SE = 0.55$, $p = .0049$), and a logistic mixed-effects regression confirmed that the likelihood of reporting TUT also increased over trials ($b = 0.031$, $SE = 0.0096$, $p = .0011$).

A second prediction was that RTs should be slower during periods of mind-wandering and faster during on-task states. An unregistered logistic mixed-effects regression that modeled TUT as a function of RTz confirmed that slower RTs co-occurred with TUT reports ($b = 0.16$, $SE = 0.056$, $p = .0033$). This association is a broad indicator of objective–subjective coupling between the measures of attention.

Finally, we performed a preregistered between-groups comparison (speeded vs. unspeeded) on the time until response for thought probes to check that the time pressure manipulation worked. A deadline cue drastically reduced the probe RTs (M ± SD; speeded group M = 2393.88 ± 954.30ms; unspeeded group M = 3575.42 ± 2039.86ms; $t(119.58) = -4.78$, $p < .0001$, $d = -0.73$).

Social desirability bias, time pressure, and TUT

The first hypothesis was that SDB (M = 4.69 ± 2.63) and time pressure would affect TUT. A preregistered logistic mixed-effects regression that modeled TUT by SDB and speeded group assignment showed that higher SDB scores were uniquely associated with a lower likelihood of reporting TUT (b = -0.18, SE = 0.091, p = .044; Figure 2). Time pressure, on the other hand, did not significantly affect TUT (b = -0.90, SE = 0.49, p = .066), though the estimate's direction suggested that it reduced it.

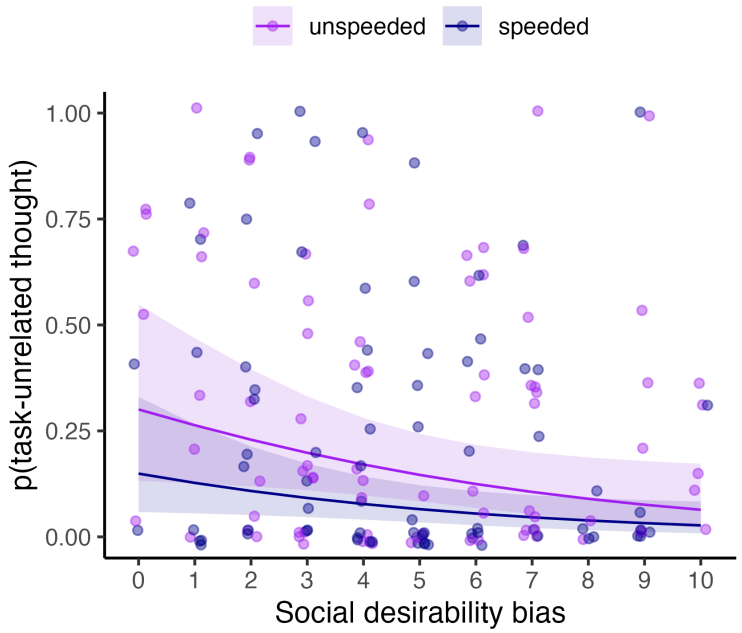

Figure 2. Model predictions from a mixed-effects logistic regression: TUT ~ speeded group + SDB + (1 | subjectID). Jittered points represent individual participants' proportion of TUT by SDB score. SDB was significantly associated with a lower likelihood of reporting TUT. Error shading represents the 95% confidence interval.

An alternative explanation for this pattern of results is that SDB and time pressure were truly associated with fewer attentional lapses rather than just the reporting of those lapses. Thus, we specified an analogous control analysis using an unregistered linear mixed-effects regression with target RT as the dependent variable and SDB and time pressure as unique predictors. SDB was not significantly associated with objective performance (b = 8.81, SE = 5.46, p = .11). Time pressure on the thought probes improved performance on the task (b = -64.51, SE = 28.71, p = .026).

The second hypothesis was that SDB would moderate the effect of time pressure on TUT reports. A preregistered logistic mixed-effects regression was specified with TUT as the dependent variable and the interaction between speeded group and SDB scores as the predictor. The model showed no significant interaction ($b = -0.06$, $SE = 0.19$, $p = .73$) nor significant main effects of SDB ($b = -0.16$, $SE = 0.12$, $p = .17$) and time pressure ($b = -0.60$, $SE = 0.98$, $p = .54$).

The third hypothesis was that SDB would moderate time pressure's decoupling of TUT reports from objective performance (target RTs). We first sought to establish a baseline effect of time pressure on objective–subjective coupling with a preregistered logistic mixed-effects regression that modeled TUT by the interaction between speeded group and RTz. The model again showed an association between RTz and TUT (here indicative of objective–subjective coupling; $b = 0.31$, $SE = 0.075$, $p < .0001$) and no significant main effect of time pressure ($b = -0.86$, $SE = 0.50$, $p = .086$). There was an interaction such that time pressure decoupled RTz and TUT ($b = -0.35$, $SE = 0.11$, $p = .0019$). A simple slopes follow-up confirmed that the speeded group had no significant association between RTz and TUT ($b = -0.037$, $SE = 0.086$, $p = .67$) whereas the unspeeded group did ($b = 0.32$, $SE = 0.075$, $p < .0001$). This broadly indicates that time pressure decoupled the objective and subjective lapse indices.

As preregistered, we next added SDB as a predictor, now testing TUT as a function of the three-way interaction between speeded group, RTz, and SDB. We anticipated that time pressure's effect on the RTz–TUT association should be moderated by SDB if the decoupling is occurring via a self-presentation mechanism. However, there was no three-way interaction ($b = 0.010$, $SE = 0.045$, $p = .82$).

We conducted a planned test that answered the same question as the models specified above. Here, however, we computed an objective–subjective coupling measure per participant (cf. Adam et al., 2020). For each participant, we calculated the Pearson's r correlation coefficient between their RTz and TUT where a greater value represented higher coupling between objective and subjective lapse indices. This computation dropped participants who had zero TUT variance, leaving us with 97 participants to analyze (57 unspeeded, 40 speeded).

In an unregistered check, we performed a t-test with this coupling score as a function of the speeded group assignment (Figure 3). This corroborated the previous tests that showed that time pressure decoupled RTz and TUT (speeded group $M = -0.027 \pm 0.25$; unspeeded group $M = 0.14 \pm 0.31$; $t(92.48) = 2.91$, $p = .0046$, $d = 0.58$). We next used a preregistered linear regression to model this coupling score as a function of the interaction between SDB and speeded group. Contrary to the hypothesis, there was no evidence that SDB moderated the

effect of time pressure on objective–subjective coupling (b = -0.0011, SE = 0.024, p = .97). There were also no main effects of time pressure (b = -0.16, SE = 0.12, p = .19) or SDB (b = 0.012, SE = 0.013, p = .36) on coupling.

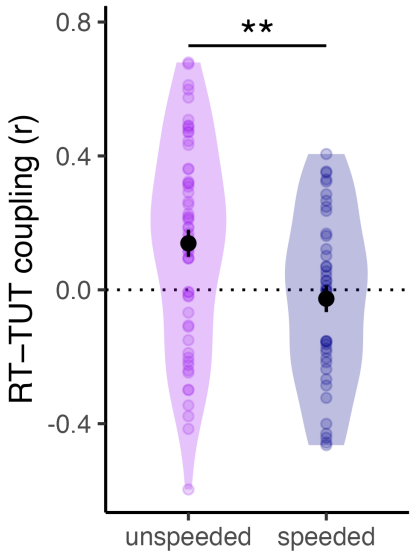

Figure 3. RT–TUT coupling for the speeded vs. unspeeded groups. A deadline cue designed to impose time pressure during the thought probe reduced the intrasubject association between subjective and objective attention measures. Error bars represent standard error. ** p < .01

Interim Discussion

The results of Experiment 1 established, for the first time, a link between scale-measured SDB and state reports of TUT in a sustained attention task. SDB was linked to a lower probability of reporting TUT but not linked to RTs, suggesting that self-presentation biases cannot be explained by true attentional differences and only emerge when the index of lapsing attention is self-report. Although we failed to amplify SDB in TUT reports via time pressure during thought probes, we showed that time pressure broadly reduced introspective accuracy likely through some mechanism unrelated to self-presentation. Interestingly, the time pressure manipulation conferred a benefit to objective performance in the task, suggesting some spillover effect of pressure onto the preparedness to react to targets.

There was no pattern in the results to suggest that accounting for SDB improved the coupling between the objective and subjective indices of lapsing attention. We posited that this was because Experiment 1 probed every trial, which likely included a mix of clearly on-task, clearly off-task, and more ambiguous “in-between” moments. In Experiment 2, however, we

selectively triggered thought probes in real time using a RT-speed-based real-time triggering procedure (Shelat et al., 2024). The broad aims were to replicate the SDB–TUT link in a different sustained attention task (the SART; Robertson et al., 1997; deBettencourt et al., 2019) often used to study mind-wandering (e.g., Martínez-Pérez et al., 2021; Stawarczyk et al., 2011; Shelat & Giesbrecht, 2025), and to see whether statistically accounting for SDB improves predictive utility of triggers in anticipating self-reported mind-wandering (i.e., how SDB relates to objective–subjective coupling). In this way, Experiment 2 extends Experiment 1 by asking whether SDB's effect on subjective reports becomes more pronounced when attention is sampled at its extremes. To formally restate the hypotheses, we predicted that (1) fast (vs. slow) triggers would capture reduced WM encoding and deeper TUT, (2) SDB would affect TUT reports, (3) SDB would moderate the association between TUT reports and objective performance (fast/slow triggers), and (4) other individual differences (trait spontaneous/deliberate mind-wandering, motor impulsivity) would shape the coupling between TUT and objective performance (exploratory).

Experiment 2

Methods

Participants

For Experiment 2, we used a different heuristic to select the intended sample than in Experiment 1. Rather than doing a simulation-based power analysis on pilot data, we inspected sample sizes from the original study we partly aimed to replicate (deBettencourt et al., 2019). Their maximum recruited sample across experiments was $n = 28$. We doubled this given the new planned individual differences analyses and aimed for 56 participants.

A total of 58 pre-exclusion participants from the subject pool at the University of California, Santa Barbara completed the experiment. They participated in exchange for either course credit or US$20. Like Experiment 1, participants were required to be 18 years of age or older and have normal or corrected-to-normal color vision. We applied exclusion criteria in two steps. First, we calculated a measure of behavioral performance (A′) based on SART responses and checked for anyone who exceeded three standard deviations from the sample mean. Frequent trials that had accurate or inaccurate responses were coded as hits or misses, respectively; infrequent trials that had accurate or inaccurate/no responses were coded as correct rejections or false alarms, respectively. A′ was computed as $0.5 + ((HR – FAR) \times (1 + HR – FAR)) / (4 \times HR \times (1 – FAR))$, where HR was hit rate and FAR was false alarm rate. No participants exhibited irregular A′, and thus, none were excluded at this stage. We next checked whether any participants exhibited an omission error rate of over 10% of trials in the SART (e.g.,

Shin et al., 2024). An omission error occurs when a participant makes no response to a frequent trial; high rates would suggest abnormally low engagement with the task. Three participants were excluded at this step, leaving a final sample of 55 participants (M_{age} = 19.9 years ± 1.9; 44 female, 8 male, 3 other). Two of these participants only completed two blocks instead of three due to time constraints, and their data is included in all analyses. A two-tailed sensitivity analysis in G*Power (Faul et al., 2007) at α = .05 confirmed that this sample could detect d_z = 0.54 with .975 power, which is smaller than the effect size (d_z = 0.71) reported by deBettencourt et al. (2019; Experiment 2b). All procedures were approved by the Institutional Review Board at the University of California, Santa Barbara and the Army Research Office/Human Research Protection Office.

Materials

Sustained attention task. We administered a SART in which six colored shapes were presented along an invisible circumference on each trial (deBettencourt et al., 2019; Keene et al., 2022; Figure 4). On most trials (90%), the shapes were all circles (frequent); on the remaining 10% of trials, the shapes were squares (infrequent). Participants pressed “z” for circles and “m” for squares while maintaining fixation on a central dot. Each trial was presented for 1000ms. RTs were recorded from trial onset to keypress. The fixation dot turned white following a keypress to signal that the response was registered. Shape colors were drawn without replacement from nine colors: red, blue, green, yellow, magenta, cyan, white, black, and dark orange. Colors were resampled if any color consecutively repeated at the same spatial location.

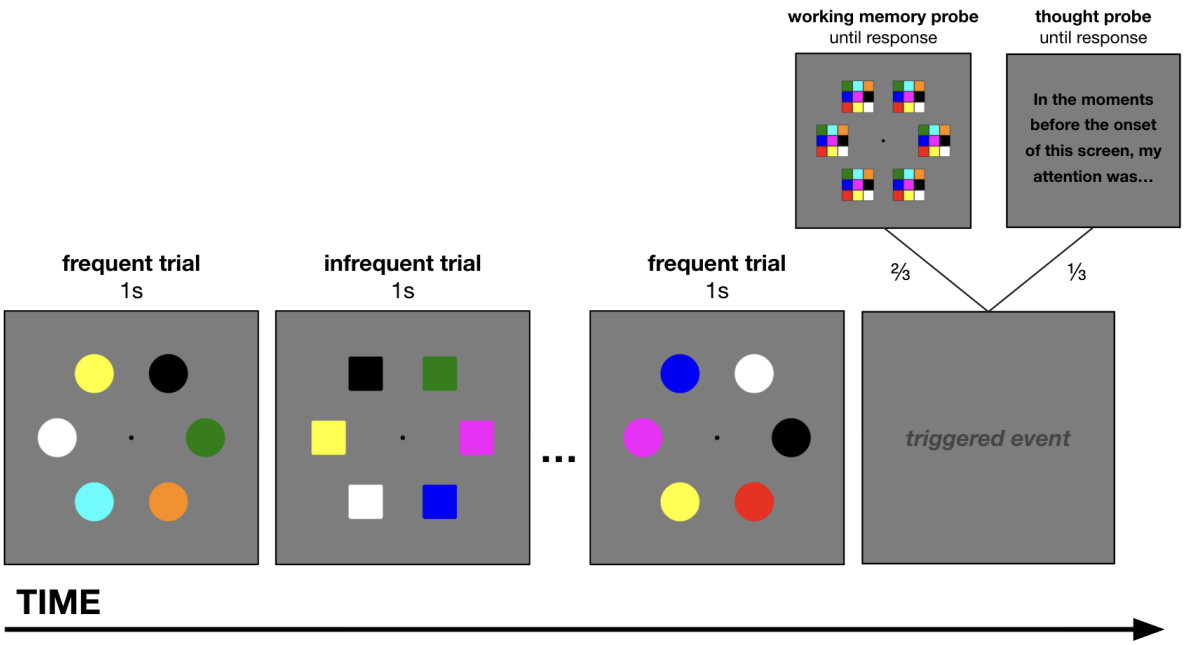

Figure 4. A schematic of a trial sequence in Experiment 2's SART. Participants pressed a key based on whether the shapes were squares or circles. During triggered events in high and low attentional states, participants were interrupted by either a working memory probe or a seven-point Likert scale thought probe.

The task was divided into three blocks of 600 trials each, separated by 30s timed interblock breaks. Prior to the main blocks, participants performed 25 trials of a practice SART block, 7 trials of a practice WM report block, and 50 trials of a combined practice SART with embedded WM and thought probes.

We integrated a triggering procedure to sample periods of high and low sustained attention in real time (Figure 5). Triggers began on frequent trials only after the first 60 trials of each block. If the mean RT across the previous three frequent trials was more than one standard deviation from the participant's cumulative mean RT and no trigger had occurred in the previous five trials, then a triggered event was presented immediately after the current trial. Triggered events were either a WM probe (⅔ chance) or a thought probe (⅓ chance). The WM probe prompted participants to use their cursor to select the color that had been at each spatial location on the previous trial. The thought probe asked participants to rate the depth of their mind-wandering on a seven-point scale from "completely on task" to "completely off task."

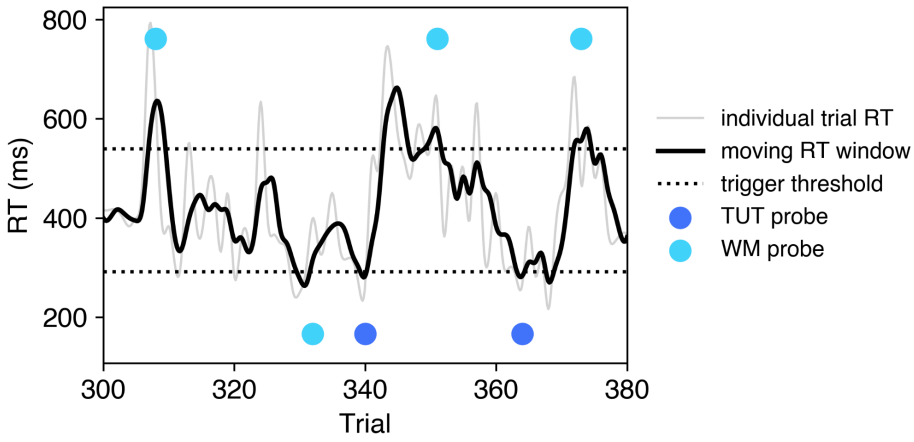

Figure 5. One participant's subset of trials with the real-time speed-based triggering procedure. When the three-trial rolling mean RT exceeded one standard deviation from the cumulative RT, a TUT probe or WM probe was triggered.

Social desirability bias survey. We used the same SDB survey as in Experiment 1.

Other surveys. We administered more surveys than Experiment 1 since one of the aims was to assess the generalizability of the real-time triggering procedure across multiple individual differences (Shelat et al., 2024). Participants filled out trait measures of deliberate (MWD) and spontaneous mind-wandering (MWS; Carriere et al., 2013). MWD items assessed the tendency to intentionally engage in mind-wandering (e.g., "I allow my thoughts to wander on purpose"), whereas MWS items assessed the tendency to unintentionally mind-wander (e.g., "I find my thoughts wandering spontaneously"). Responses were on a seven-point frequency scale (1 = "rarely", 7 = "a lot").

We also administered the seven-item motor impulsivity subscale of the Barratt Impulsiveness Scale (BIS-11; Patton et al., 1995). This scale contains items such as "I act on the spur of the moment", measured on a four-point scale (1 = "rarely/never", 4 = "always/almost always").

Participants filled out the 19-item Effortful Control Scale (ECS; Evans & Rothbart, 2007) with items such as "I often make plans that I do not follow through with." Participants responded on seven-point scales (1 = "extremely untrue of you", 7 = "extremely true of you"). We did not analyze these scores.

Procedure

Participants first provided informed consent and completed the full set of individual differences/demographic questionnaires. They were seated individually in a quiet laboratory room with white noise. Viewing distance was 88 cm from a 48 × 27 cm monitor (1920 × 1080

pixels). The testing station was placed in a corner of the room, with the experimenter facing away in a separate corner across the room.

TUT was defined in the same way as Experiment 1 (Kane et al., 2021), but here measured on a seven-point depth scale instead of binary to capture finer-grained shifts in experience. Practice blocks could be repeated, and the experimenter was available for questions. Participants performed the blocks of the main task and then were dismissed.

Results

Task checks

To verify the task mechanics, we checked whether slow triggers were preceded by significantly slower trial windows than fast triggers. A linear mixed-effects regression confirmed this ($b = 371.80$, $SE = 3.40$, $p < .0001$). We next checked that there were enough instances of measurement for all trigger types and that the ratio of thought probes to WM probes was approximately 1:2 as programmed. For fast triggers, there was an average of 12.82 ± 4.54 TUT probes and 24.96 ± 7.79 WM probes. For slow triggers, there was an average of 12.82 ± 4.16 TUT probes and 25.47 ± 9.13 WM probes.

We tested two baseline predictions. First, the triggering procedure should catch fluctuations in visual WM encoding such that fast triggers precede worse WM accuracy than slow triggers (replicating deBettencourt et al., 2019); a linear mixed-effects regression confirmed this ($b = -0.20$, $SE = 0.046$, $p < .0001$). Second, the triggering procedure should catch fluctuations in the task-relatedness of thought such that fast triggers precede higher TUT reports than slow triggers (Franklin et al., 2011). A linear mixed-effects regression did not support this prediction ($b = 0.081$, $SE = 0.063$, $p = .20$).

Finally, we tested to what degree the predictive utility of the triggering procedure for WM and TUT converged. We computed the average WM accuracy and TUT report across fast and slow triggers per participant and took their difference. A bigger difference indicated that triggers worked particularly well in separating that variable. These predictive utility scores for WM and TUT were moderately correlated ($r = -.31$, $p = .021$; Figure 6), suggesting that participants whose WM was sensitive to the triggering procedure also tended to have TUT that was sensitive to the procedure. This medium effect suggests that triggers track overlapping but not identical attentional dynamics across objective and subjective measures, and that there might be unaccounted-for trait differences that explain variance in trigger utility across individuals.

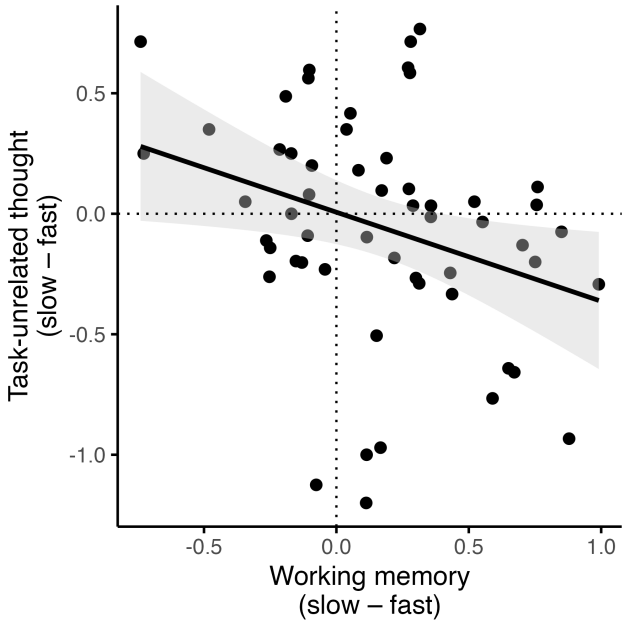

Figure 6. The correlation between the predictive utility of the triggering procedure for WM and TUT. Each point represents a participant's difference score between slow- and fast-triggered trials. Note that triggers' success is represented by a negative "slow – fast" difference in TUT, but a positive "slow – fast" difference in WM. The negative association between the two shows that the metrics capture overlapping dynamics. Error shading represents the 95% confidence interval.

Social desirability bias, real-time triggers, and TUT

The first hypothesis was that SDB (M = 4.52 ± 2.42) would be associated with TUT, like in Experiment 1. This sets up our sequence of tests on whether SDB distorts state reports even when objective performance remains unchanged. A linear mixed-effects model failed to provide evidence for this (b = -0.064, SE = 0.072, p = .38). A separate check confirmed that SDB was not associated with WM accuracy (b = 0.0075, SE = 0.033, p = .82). The second hypothesis was that SDB would moderate the association between TUT reports and objective performance (fast/slow triggers). To test this, we added an interaction term between SDB and trigger type to the model. The model revealed that fast triggers captured deeper TUT relative to slow triggers (b = 0.34, SE = 0.14, p = .016; Figure 7). There was a significant interaction such that the trigger–TUT coupling was weaker for those with high SDB scores (b = -0.055, SE = 0.027, p = .040). Only after accounting for SDB did the triggering procedure reveal its expected relationship with TUT, and this relationship was weaker for high-SDB individuals.

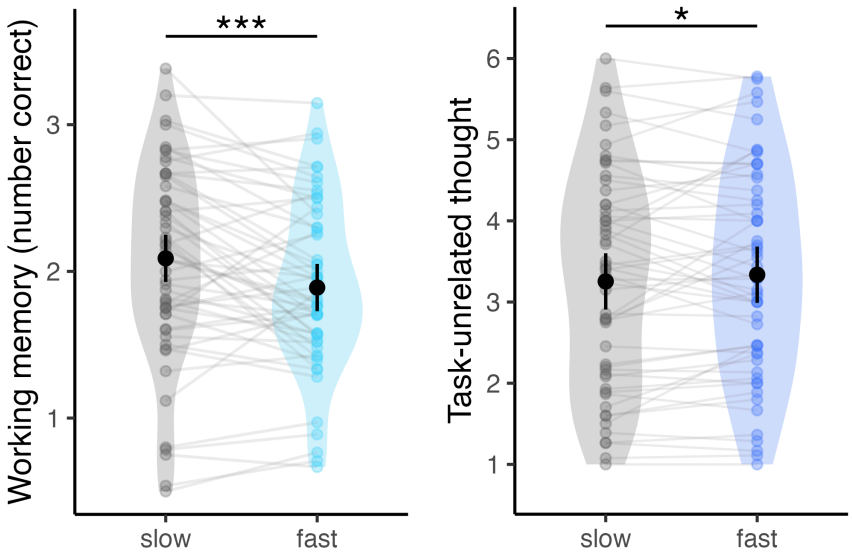

Figure 7. Linear mixed-effects model predictions for WM performance via WM ~ trigger + (1 | subjectID), and TUT via TUT ~ trigger × SDB + (1 | subjectID). Note that modeling TUT by triggers alone did not yield a significant effect of triggering. The effect only emerged when SDB was included as a moderator, and this plot illustrates the main effect of triggers on TUT once accounting for SDB. Error bars represent 95% confidence intervals. * p < .05, *** p < .0001

To confirm that the benefit to statistical precision from accounting for SDB in the model was unique to the self-report measure of lapsing attention, we specified an analogous control analysis with WM accuracy modeled by the interaction between SDB and trigger type. Fast triggers still predicted lower WM accuracy than slow triggers (b = -0.30, SE = 0.10, p = .0036) and, as expected, there was no significant main effect of SDB on WM accuracy (b = -0.0035, SE = 0.034, p = .92) nor a significant interaction (b = 0.022, SE = 0.020, p = .28). The dissociation between this result and the previous one provides the strongest evidence that social desirability distorts introspective reports rather than the actual attentional dynamics themselves.

We expanded the TUT model to now include interactions between trigger type and three additional traits (MWD, MWS, and motor impulsivity). The goal of this analysis was twofold: to test the durability of the moderating SDB effect while controlling for other attention- and behavior-related individual differences, and to explore whether the predictive utility of the triggering procedure varied across such traits. The main effect of trigger type on TUT was no longer significant but remained directionally consistent (b = 0.68, SE = 0.36, p = .066). The interaction between trigger type and SDB remained significant such that triggers were worse at predicting TUT for high SDB participants (b = -0.062, SE = 0.028, p = .025). All other individual differences had no interactions with trigger type (motor impulsivity p = .73; MWS p = .58; MWD

p = .11). However, given that the trigger × MWD interaction (b = -0.076, SE = 0.048, p = .11) was nearest the alpha threshold of .05, we reasoned that periods of TUT in deliberate mind-wanderers might result in different behavioral patterns, weakening the usefulness of speed-based triggering in anticipating lapses.

To extend this, we modeled WM accuracy as a function of trigger type and MWD. The main effect of trigger on WM accuracy remained (b = -0.54, SE = 0.15, p = .00026). There was an interaction where, for high deliberate mind-wanderers, the trigger–WM link substantially weakened (b = 0.07, SE = 0.030, p = .015). In other words, the triggering procedure did not work as well in catching WM lapses for individuals prone to intentional TUT in their daily life.

General Discussion

The field of mind-wandering is built on the assumption that participants, when asked, are introspecting on their thoughts and veridically self-reporting their attentional state. Although thought probes provide a direct readout of phenomenological states, the field's reliance on subjective reports raises the concern of social desirability bias (SDB) in conclusions drawn from existing empirical work. One limitation of prior work (Marcusson-Clavertz & Kjell, 2018) is that it relied only on trait questionnaires, which cannot distinguish whether SDB represents true differences in attention or just differences in self-presentation. To address this gap, we looked at state-level mind-wandering during cognitive tasks to determine if SDB is associated with thought reports independent of attention.

We applied this approach in two sustained attention tasks: a boring target reaction task (Experiment 1; Nedergaard et al., 2023) and a SART (Experiment 2; Robertson et al., 1997; deBettencourt et al., 2018). We interleaved thought probes to measure mind-wandering and recorded trait SDB scores on a 10-item true/false survey. Experiment 1 also tested whether time pressure during probes could amplify SDB (Protzko et al., 2019), while Experiment 2 used real-time triggers to sharpen the mapping between objective lapses and subjective reports (Shelat et al., 2024).

Across experiments, the pattern was consistent with a reporting-bias account. Experiment 1 showed that SDB was linked with reduced TUT but not with RTs. SDB did not moderate the relationship between RTs and TUT, nor did it moderate the decoupling effect of time pressure on RT–TUT. Experiment 2 showed that SDB moderated the link between speed-based triggers and TUT, but not between triggers and WM encoding. Objective–subjective coupling was reduced for those with high bias, and the main effect of triggers on TUT was only significant after statistically accounting for SDB. Our state-level approach isolated the influence of SDB on mind-wandering reports, and moreover revealed

when subjective measures diverge from attention. We have divided the general discussion into takeaways that are methodological and those that are theoretical.

Methodological implications

Methodological implications of this work are critical to the field's advancement. First, a brief SDB survey is a low-cost way to improve the validity of thought probe data. It enables statistical control of a bias that has, until now, gone unaccounted for and possibly muddied inferences about mind-wandering. Experiment 2 demonstrated this clearly: real-time triggers were able to catch the wandering mind in flight, but only once SDB was statistically accounted for. In this way, SDB control both improves estimates of TUT occurrence and increases the strength of conclusions derived from objective–subjective coupling.

Second, introspecting on task-relatedness of thought requires adequate time. In Experiment 1, adding a rapid deadline cue for thought reports disrupted the relationship between subjective reports and objective RTs. This reduction in the value of probes occurred alongside an unexpected improvement in performance on the attention task. Researchers should be wary of incidental sources of time pressure (whether experimenter- or participant-imposed) as they could obscure introspective validity and inadvertently affect behavior.

Finally, while real-time triggers remain a valuable tool for unobtrusively tracking attentional states, their predictive utility varies from person to person. For example, we posited that triggers may be less effective for participants with high motor impulsivity since their RT speeding may represent a transient period of impulsive responding rather than TUT and/or reduced visual processing; however, exploratory analyses yielded no evidence for this. Instead, trait tendencies for deliberate (but not spontaneous) TUT did weaken the triggers' ability to capture flux in WM encoding. This pattern is consistent with existing work showing that MWD is not associated with RT variability in a metronomic response task, while MWS is (Seli et al., 2013). As those authors propose, deliberate TUT might strategically occur only when it is less disruptive to performance. The present observation that RT-based sampling was less informative for high-MWD participants would suggest that these individuals can either maintain controlled responding during visual lapses or maintain visual attention during uncontrolled responding. Our speculation that intentional lapses manifest differently in behavioral patterns is uncertain since we did not collect state-level reports of thought intentionality (Seli et al., 2016).

Theoretical implications

SDB is a heterogeneous construct that can be decomposed into two dimensions: impression management and self-deception (Paulhus, 1984). Impression management is

thought to be deliberate and conscious positive self-presentation, whereas self-deception is unconscious and genuinely believed self-perceptions. If the findings are partly driven by self-deception, SDB might reflect a *translation dissociation* in meta-awareness (Schooler, 2002; Schooler & Schreiber, 2004). Instead of willfully concealing lapses, people may instead be misreporting them because their introspective access is distorted by expectations or self-concepts. Rather than drawing directly on their experiential state, they replace beliefs about themselves when self-reporting. In this interpretation, self-deceptive SDB is a metacognitive failure to accurately re-represent ongoing experience.

An alternative is that the results reflect impression management more than self-deception. This is because the MCSD survey is shown to be especially effective at capturing deliberate distortion (Lambert et al., 2016) rather than unconscious bias. Notably, our version of the MCSD (Vésteinsdóttir et al., 2017) was slightly different from that used by Lambert & colleagues. Given that it was not a goal of this study to tease apart the relative contributions of each type of bias, one way forward is to directly address this by using measures that break SDB down into subdimensions such as the Balanced Inventory of Desirable Responding (BIDR; Paulhus, 1988).

Both forms of bias imply that attention itself has intrinsic value, whether individuals claim attentiveness for social reasons or unconsciously misperceive lapses to align with a positive self-view. This is made clear by a familiar example. Imagine zoning out mid-conversation when a friend probes you: “Are you paying attention?” Whether to protect how they see you or to reaffirm your self-concept as a focused friend, you instinctively reply: “Yes, of course I was!” Within the lab, similar dynamics could lead a participant to underreport mind-wandering to appear competent or maintain an internal sense of being a good subject. This perspective aligns with an account of attention as a symbolic resource in an “attention economy” (Heitmayer, 2025). In this metaphorical economy, inattention lowers one’s value where attentiveness is prized. At the same time, our results leave open the possibility that distortions come from inward self-perceptions as much as from outward pressures. In either case, self-reports of mind-wandering are intertwined with biases beyond experience itself, and this is an important caveat when using subjective approaches to study attention.

More broadly, it may be helpful to reconceptualize TUT reports as the outcome of a multidimensional decision-making process rather than a direct readout of inner states. Polychroni et al. (2025) showed that TUT reports were uniquely associated with random false feedback *and* objective performance measures. The combined influence suggests that participants weighed both recent external performance cues (i.e., in a process of retrospective

inference) and introspection on attentional state when selecting their mind-wandering report. The data here suggests that SDB is another layer in this decision process. We theorize a hydraulic decision-making mechanism during thought probes, where the relative weight of different influences shifts depending on multiple factors. For example, inference from external performance cues may increase in their contribution to TUT reports in tasks where errors are particularly salient; impression management SDB, on the other hand, may exert more pressure when an experimenter is visibly monitoring the participant's behavior. In both cases, the contribution of true introspection might be suppressed by contextual or social factors. If TUTs indeed arise from composite judgments, the field should be cautious of states, traits, or experimental designs that create sources of influence which reduce introspection.

Future work and limitations

We showed that bias can be accounted for by entering SDB scale scores into statistical analyses. Since this risks overcomplicating analytical approaches, it would be useful to identify designs that eradicate the bias without regression. There are two paths forward that we propose. First, we can identify types of TUT probes that are resilient to bias. Kane et al. (2021) recommended content probes where participants categorize their thoughts into the task, everyday things, current state of being, personal worries, daydreams, external environment, or other. The first response is coded as on-task and the rest are TUT. Kane et al. showed that these were responded to with greater confidence and may have been less susceptible to retrospective inference. Given that they prompt the immediate reporting of thought content instead of an evaluative judgment (e.g., on TUT depth), they might be less susceptible to SDB. Second, we can validate interventions that operate on constructs related to SDB. A recent megastudy by Zickfeld et al. (2025) showed what types of honesty oaths best reduce dishonest behaviors. Oaths that encourage truthful reporting may extend previous work linking honesty and TUT validity (Vinski & Watter, 2012), and could serve as a practical way to reduce SDB in mind-wandering studies. Direct tests of content probes and honesty oaths are one clear way forward.

Given that inner experience is deeply heterogeneous, SDB's impact may vary based on the types of thought being measured. Mind-wandering can occur with or without intention (Seli et al., 2016) and with or without awareness (Schooler et al., 2011; Seli et al., 2017). TUTs vary on phenomenological qualities such as temporality (Stawarczyk et al., 2011) or even stickiness (van Vugt & Broers, 2016). An important question for future work is how different types of thought reports are shaped by SDB as the quantitative field of thought phenomenology grows (Marome et al., 2025; Garg et al., 2025b; Ortega et al., 2025; Zhou et al., 2025). Unintentional

TUT, for example, could relate more strongly to SDB given that trait SDB was negatively associated with spontaneous but not deliberate trait mind-wandering in previous psychometric work (Marcusson-Clavertz & Kjell, 2018).

A limitation of this study is that its two experiments had several differences, which complicates their comparison. We argued that SDB had a more pronounced influence on objective–subjective coupling in Experiment 2 because that design provided a more sensitive method for sampling high and low attention states in real time. However, it is important to remember that Experiment 2 differed from Experiment 1 in many other ways: it used a different sustained attention task, recruited from a student pool instead of online volunteers, was conducted in-person instead of online, and administered a seven-point depth probe instead of a binary probe. All of these differences limit the cross-experiment comparisons we draw. Even so, they strengthen the robustness of one basic finding; across two different tasks, samples, settings, and probe formats, SDB was linked with TUT. We caution the reader against overinterpretation due to cross-experiment differences, but the convergence across methods leads to the same conclusion: SDB is present and should be accounted for in studies on the wandering mind.

Conclusion

In two experiments, we asked whether a previous link between trait mind-wandering and trait social desirability reflects genuine differences in attentional lapses or differences in what people report. By pairing state-level TUT reports with objective performance, we found a consistent dissociation: SDB was associated with subjective reports but not behavioral markers of lapsing. This pattern indicates that high-SDB individuals do not appear to mind-wander less; rather, they disclose it less. A practical takeaway follows: a brief SDB scale can be used to identify and control this bias. Accounting for the biases that muddy introspective judgments is one critical step to strengthen the edifice on which the science of mind-wandering is built.

Funding: SS is supported by a National Science Foundation Graduate Research Fellowship under grant 2139319. Experiment 1 was supported by the Gene and Susan Lucas Undergraduate Research Fund to ARC. Experiment 2 was supported by the Institute for Collaborative Biotechnologies through contract W911NF-19-2-0026 from the U.S. Army Research Office. The content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

Adam, K. C., Doss, M. K., Pabon, E., Vogel, E. K., & de Wit, H. (2020). Δ9-Tetrahydrocannabinol (THC) impairs visual working memory performance: a randomized crossover trial. *Neuropsychopharmacology, 45*(11), 1807-1816.

Aust, F., Diedenhofen, B., Ullrich, S., & Musch, J. (2013). Seriousness checks are useful to improve data validity in online research. *Behavior Research Methods, 45*(2), 527–535.

Carriere, J. S., Seli, P., & Smilek, D. (2013). Wandering in both mind and body: individual differences in mind wandering and inattention predict fidgeting. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *67*(1), 19.

Corneille, O., & Gawronski, B. (2024). Self-reports are better measurement instruments than implicit measures. *Nature Reviews Psychology, 3*(12), 835-846.

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, *24*(4), 349–354.

deBettencourt, M. T., Keene, P. A., Awh, E., & Vogel, E. K. (2019). Real-time triggering reveals concurrent lapses of attention and working memory. *Nature human behaviour*, *3*(8), 808-816.

deBettencourt, M. T., Norman, K. A., & Turk-Browne, N. B. (2018). Forgetting from lapses of sustained attention. *Psychonomic Bulletin & Review*, *25*(2), 605-611.

Evans, D. E., & Rothbart, M. K. (2007). Developing a model for adult temperament. *Journal of research in personality*, *41*(4), 868-888.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods, 39*(2), 175-191.

Franklin, M. S., Smallwood, J., & Schooler, J. W. (2011). Catching the mind in flight: Using behavioral indices to detect mindless reading in real time. *Psychonomic bulletin & review, 18*(5), 992-997.

Garg, A., Shelat, S., & Schooler, J. W. (2025a). Now I feel like I'm going to get to it soon: a brief, scalable intervention for state procrastination. *BMC psychology, 13*(1), 1158.

Garg, A., Shelat, S., Gross, M. E., Smallwood, J., Seli, P., Taxali, A., ... & Schooler, J. W. (2025b). Opening the black box: Think Aloud as a method to study the spontaneous stream of consciousness. *Consciousness and Cognition*, *128*, 103815.

Heitmayer, M. (2025). The second wave of attention economics. Attention as a universal symbolic currency on social media and beyond. *Interacting with Computers*, *37*(1), 18-29.

Helton, W. S. (2009). Impulsive responding and the sustained attention to response task. *Journal of Clinical and Experimental Neuropsychology*, *31*(1), 39-47.

Hemmerich, K., Luna, F. G., Martín-Arévalo, E., & Lupiáñez, J. (2025). Understanding vigilance and its decrement: theoretical, contextual, and neural insights. *Frontiers in Cognition*, 4, 1617561

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., Smeekens, B. A., Meier, M. E., Welhaf, M. S., & Phillips, N. E. (2021). Testing the construct validity of competing measurement approaches to probed mind-wandering reports. *Behavior Research Methods*, *53*(6), 2372-2411.

Keene, P. A., DeBettencourt, M. T., Awh, E., & Vogel, E. K. (2022). Pupillometry signatures of sustained attention and working memory. *Attention, Perception, & Psychophysics*, *84*(8), 2472-2482.

Lambert, C. E., Arbuckle, S. A., & Holden, R. R. (2016). The Marlowe–Crowne social desirability scale outperforms the BIDR impression management scale for identifying fakers. *Journal of Research in Personality*, 61, 80-86.

Larson, R. B. (2019). Controlling social desirability bias. *International Journal of Market Research*, *61*(5), 534-547.

Leisge, K., Kaczmarek, C., & Schaefer, S. (2024). How often do you cheat? Dispositional influences and intrapersonal stability of dishonest behavior. *Frontiers in Psychology*, *15*, 1297058.

Mackworth, N. H. (1948). The breakdown of vigilance during prolonged visual search. *Quarterly journal of experimental psychology*, *1*(1), 6-21.

Marcusson-Clavertz, D., & Kjell, O. N. (2018). Psychometric properties of the spontaneous and deliberate mind wandering scales. *European Journal of Psychological Assessment*.

Marome, B., Shelat, S., & Schooler, J. W. (2025). The phenomenology of encoding: Experience sampling reveals thoughts associated with the retention of visual and verbal materials. *Consciousness and Cognition*, *136*, 103958.

Martindale, A. P., Deane, E. M., Peral-Fuster, C. I., Elkelani, O., Qi, Z., Ribeiro-Ali, S. I., ... & Witchel, H. J. (2024). Two lingering delays in a go/no-go task: mind wandering and caution/uncertainty slow down thought probe response times. *Behaviour & Information Technology*, *43*(9), 1729-1739.

Martínez-Pérez, V., Baños, D., Andreu, A., Tortajada, M., Palmero, L. B., Campoy, G., & Fuentes, L. J. (2021). Propensity to intentional and unintentional mind-wandering differs in arousal and executive vigilance tasks. *PLoS one, 16*(10), e0258734.

Mooneyham, B. W., & Schooler, J. W. (2013). The costs and benefits of mind-wandering: a review. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 67*(1), 11.

Nedergaard, J., Skewes, J. C., & Wallentin, M. (2023). “Stay focused!”: The role of inner speech in maintaining attention during a boring task. *Journal of Experimental Psychology: Human Perception and Performance*, *49*(4), 451.

Nederhof, A. J. (1985). Methods of coping with social desirability bias: A review. *European journal of social psychology*, *15*(3), 263-280.

Ortega, J. R., Gross, M. E., & Schooler, J. W. (2025). When life is but a dream: Does transliminality predict continuity of thought across the sleep-wake cycle?. *Philosophy and the Mind Sciences*, *6*.

Pashler, H., Rohrer, D., & Harris, C. R. (2013). Can the goal of honesty be primed?. *Journal of Experimental Social Psychology*, *49*(6), 959-964.

Patton, J., Stanford, M., Barratt, E. (1995). Factor structure of the Barratt impulsiveness scale. *Journal of Clinical Psychology*, *51*(6), 768-774.

Paulhus, D. L. (1988). Balanced inventory of desirable responding (BIDR). Acceptance and Commitment Therapy. *Measures Package, 41*, 79586-79587.

Polychroni, N., Konishi, M., Steinecker, I., & Terhune, D. B. (2025). Introspective Access or Retrospective Inference? Mind-Wandering Reports Are Shaped by Performance Feedback. *Psychological science*, *36*(7), 545-558.

Protzko, J., Zedelius, C. M., & Schooler, J. W. (2019). Rushing to appear virtuous: Time pressure increases socially desirable responding. *Psychological science*, *30*(11), 1584-1591.

Robertson, I. H., Manly, T., Andrade, J., Baddeley, B. T., & Yiend, J. (1997). Oops!': performance correlates of everyday attentional failures in traumatic brain injured and normal subjects. *Neuropsychologia, 35*(6), 747-758.

Schooler, J., & Schreiber, C. A. (2004). Experience, meta-consciousness, and the paradox of introspection. *Journal of consciousness studies*, 11(7-8), 17-39.

Schooler, J. W., Smallwood, J., Christoff, K., Handy, T. C., Reichle, E. D., & Sayette, M. A. (2011). Meta-awareness, perceptual decoupling and the wandering mind. *Trends in cognitive sciences*, 15(7), 319-326.

Schooler, J. W. (2002). Re-representing consciousness: Dissociations between experience and meta-consciousness. *Trends in cognitive sciences*, 6(8), 339-344.

Schwartzman, B., Zanesco, A. P., Denkova, E., Tsukahara, J. S., & Jha, A. P. (2025). Examining the association between vigilance and mind wandering. *Frontiers in Cognition*, *4*, 1577053.

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*, 430.

Seli, P., Ralph, B. C., Risko, E. F., W. Schooler, J., Schacter, D. L., & Smilek, D. (2017). Intentionality and meta-awareness of mind wandering: Are they one and the same, or distinct dimensions?. *Psychonomic bulletin & review*, *24*(6), 1808-1818.

Seli, P., Risko, E. F., & Smilek, D. (2016). On the necessity of distinguishing between unintentional and intentional mind wandering. *Psychological science, 27*(5), 685-691.

Seli, P., Risko, E. F., Smilek, D., & Schacter, D. L. (2016). Mind-wandering with and without intention. *Trends in cognitive sciences*, *20*(8), 605-617.

Shelat, S., & Giesbrecht, B. (2025). Perceptual decoupling in the sustained attention to response task is likely: comment on Bedi, Russell, & Helton (2024). *Experimental Brain Research, 243*(4), 1-3.

Shelat, S., Schooler, J. W., & Giesbrecht, B. (2024). Predicting attentional lapses using response time speed in continuous performance tasks. *Frontiers in Cognition, 3*, 1460349.

Shin, M., Taseski, D., & Murphy, K. (2024). Media multitasking is linked to attentional errors, mind wandering and automatised response to stimuli without full conscious processing. *Behaviour & Information Technology*, *43*(3), 445-457.

Skinner, H. E., & Giesbrecht, B. (2025). Beyond detection rate: understanding the vigilance decrement using signal detection theory. *Frontiers in Cognition*, 3, 1505046.

Smallwood, J., & Schooler, J. W. (2006). The restless mind. *Psychological bulletin*, *132*(6), 946-958.

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.

Smallwood, J., McSpadden, M., Luus, B., & Schooler, J. (2008). Segmenting the stream of consciousness: The psychological correlates of temporal structures in the time series data of a continuous performance task. *Brain and cognition*, *66*(1), 50-56.

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

Stone, C., Mattingley, J. B., & Rangelov, D. (2025). Neural mechanisms of metacognitive improvement under speed pressure. *Communications Biology*, *8*(1), 223.

van Vugt, M. K., & Broers, N. (2016). Self-reported stickiness of mind-wandering affects task performance. *Frontiers in psychology*, *7*, 732.

Vésteinsdóttir, V., Reips, U. D., Joinson, A., & Thorsdottir, F. (2017). An item level evaluation of the Marlowe-Crowne Social Desirability Scale using item response theory on Icelandic Internet panel data and cognitive interviews. *Personality and Individual Differences*, *107*, 164-173.

Vinski, M. T., & Watter, S. (2012). Priming honesty reduces subjective bias in self-report measures of mind wandering. *Consciousness and cognition*, *21*(1), 451-455.

Zanesco, A. P., Denkova, E., & Jha, A. P. (2024). Mind-wandering increases in frequency over time during task performance: An individual-participant meta-analytic review. *Psychological bulletin*.

Zhang, Y., Song, X., Ye, Q., & Wang, Q. (2015). Children with positive attitudes towards mind-wandering provide invalid subjective reports of mind-wandering during an experimental task. *Consciousness and Cognition*, *35*, 136-142.

Zhou, S. S., Rowchan, K., Mckeown, B., Smallwood, J., & Wammes, J. D. (2025). Drawing behaviour influences ongoing thought patterns and subsequent memory. *Consciousness and Cognition, 127*, 103791.

Zickfeld, J. H., Ścigała, K. A., Elbæk, C. T., Michael, J., Tønnesen, M. H., Levy, G., ... & Mitkidis, P. (2025). Effectiveness of ex ante honesty oaths in reducing dishonesty depends on content. *Nature Human Behaviour*, 9(1), 169-187.