Behavior Research Methods
https://doi.org/10.3758/s13428-021-01766-4

A "Goldilocks zone" for mind-wandering reports? A secondary data analysis of how few thought probes are enough for reliable and valid measurement

Matthew S. Welhaf[1] · Matt E. Meier[2] · Bridget A. Smeekens[1] · Paul J. Silvia[1] · Thomas R. Kwapil[3] · Michael J. Kane[1]

Accepted: 3 December 2021

Abstract
Mind-wandering assessment relies heavily on the thought probe technique as a reliable and valid method to assess momentary task-unrelated thought (TUT), but there is little guidance available to help researchers decide how many probes to include within a task. Too few probes may lead to unreliable measurement, but too many probes might artificially disrupt normal thought flow and produce reactive effects. Is there a "Goldilocks zone" for how few thought probes can be used to reliably and validly assess individual differences in mind-wandering propensity? We address this question by reanalyzing two published datasets (Study 1, $n = 541$; Study 2, ns ≈ 260 per condition) in which thought probes were presented in multiple tasks. Our primary analyses randomly sampled probes in increments of two for each subject in each task. A series of confirmatory factor analyses for each probe "bin" size tested whether the latent correlations between TUT rate and theoretically relevant constructs like working memory capacity, attention-control ability, disorganized schizotypy, and retrospective self-reported mind wandering changed as more probes assessed the TUT rate. TUT rates were remarkably similar across increasing probe-bin sizes and zero-order correlations within and between tasks stabilized at 8–10 probes; moreover, TUT-rate correlations with other latent variables stabilized at about 8 thought probes. Our provisional recommendation (with caveats) is that researchers may use as few as 8 thought probes in prototypical cognitive tasks to gain reliable and valid information about individual differences in TUT rate.

Keywords Mind wandering · Thought probe · Measurement · Reliability · Validity

The primary way psychologists assess mind wandering as it occurs, whether in the laboratory or in daily life, is through an experience-sampling technique known as the thought-probe method (for reviews see Smallwood & Schooler, 2006, 2015). Here, subjects engaged in an ongoing activity are presented with periodic visual or auditory signals that ask them to report on their immediately preceding thoughts (for a review of thought-probe variations, see Weinstein, 2018). Researchers typically assess the frequency with which subjects report task-unrelated thoughts (TUTs) to these probes during the ongoing task or activity.

In the laboratory, the thought-probe technique has been successfully implemented in a variety of tasks, including attention-control and working memory tasks (e.g., Kane et al., 2016; McVay & Kane, 2009; Robison et al., 2020; Unsworth & Robison, 2016), passage reading (e.g., Schooler et al., 2004; Smallwood et al., 2008; Unsworth & McMillan, 2013), simulated driving (e.g., Baldwin et al., 2017; Zhang & Kumada, 2018), and video-lecture viewing (e.g., Hollis & Was, 2016; Risko et al., 2012; Szpunar et al., 2013). Probed TUT-report rates appear to be valid individual-differences measures, as they are reliable across different tasks and occasions (e.g., Kane et al., 2016; Unsworth et al., 2020) and they correlate with other measures argued to reflect mind wandering and attentional lapses, such as reaction time (RT) variability (Bastian & Sackur, 2013; McVay & Kane, 2012; Seli et al., 2013b; Unsworth et al., 2010), pupil dilation and

✉ Matthew S. Welhaf
mswelhaf@uncg.edu

✉ Michael J. Kane
mjkane@uncg.edu

[1] University of North Carolina at Greensboro, Greensboro, NC 27412, USA

[2] Western Carolina University, Cullowhee, NC, USA

[3] University of Illinois at Urbana-Champaign, Champaign, IL, USA

Published online: 05 April 2022

eye movements (Reichle et al., 2010; Unsworth & Robison, 2017; Zhang et al., 2020), and retrospective self-reports of mind-wandering propensity (Carriere et al., 2013; Mrazek et al., 2013; Seli et al., 2016; Smeekens & Kane, 2016). Variation in TUT rate is also predicted by measures of theoretically relevant constructs like working memory capacity (WMC) and attention-control ability (Kane et al., 2016, 2017; McVay & Kane, 2012; Robison & Unsworth, 2018; Rummel & Boywitt, 2014), attention-deficit/hyperactivity disorder symptoms (Franklin et al., 2017; Meier, 2021; Seli et al., 2015b), and motivation for and interest in the ongoing activity (Brosowsky et al., 2020; Robison et al., 2020; Seli et al., 2015a).

Although probed TUT reports demonstrate reasonable construct validity (for a review, see Kane et al., 2021), researchers face a challenge in designing mind-wandering studies—deciding on the number and frequency of thought probes to present during a task. Infrequent probing may not provide enough reports to reliably and validly measure the TUT rate, especially in short-duration tasks; infrequent probes may also miss many instances of off-task thought that occur in the time between them. Probing too frequently, in contrast, might disrupt subjects' natural flow of thought too severely and provide insufficient time between probes to drift off-task; frequent probes might also reactively remind subjects to stay mentally on-task (see Konishi & Smallwood, 2016).

Unfortunately, the literature provides little guidance regarding optimal (or minimal) numbers or frequencies of thought probes. Only a few studies have recently examined the impact of probe frequency on observed TUT rates, either by comparing experimental groups that receive a typical versus a more-than-typical number of probes within a task (Robison et al., 2019; Schubert et al., 2019), or by parametrically varying the frequency of probes across subjects (Seli et al., 2013a). Results have been mixed. Robison et al. (2019, Experiment 1) presented subjects with thought probes after either 7% or 13% of trials in the Sustained Attention to Response Task (SART); neither TUT rate nor task performance differed significantly between groups. In contrast, Schubert et al. (2019) found significantly higher TUT rates for subjects seeing probes after only 3% of SART trials than after 6% of trials. Similarly, in the Seli et al. (2013a) study using a continuous metronome response task (MRT), probes could occur following 0.8–4.2% of trials, and TUT rates increased with more time between probes. At the same time, neither SART performance nor MRT performance in these studies was affected by probe rate, suggesting that probe rate artifactually changed subjects' subjective reports but not their underlying attentional states.

Although two of three relevant studies show that average TUT rates vary somewhat with probe rate (perhaps varying most across probe rates of 1–6%), only Schubert et al. (2019) also assessed individual differences. Probe rate did not interact significantly with any other variables in their study to predict the TUT rate (including SART performance, WMC, and questionnaire measures of mind wandering propensity), suggesting good news for mind wandering researchers: Probe rate was unrelated to individual differences assessment. But there is only so much we can conclude from one study's null effects, particularly given the modest range of probe rates tested (3% vs. 6%).

The present study reports two secondary data analyses to address a pragmatic methodological question: *How few thought probes are enough to reliably and validly assess individual differences in TUT rates?* Given the variety of tasks and contexts in which thought probes have been used, a correspondingly wide range of probe numbers and frequencies have been employed, with some studies using tasks that include as few as four probes (e.g., Forster & Lavie, 2009; Levinson et al., 2012; Robison et al., 2020; Rummel & Boywitt, 2014) and others using tasks with as many as 45–120 probes (e.g., Kane et al., 2016; McVay & Kane, 2009, 2012). Ceteris paribus, using fewer probes has the potential advantages of less reactivity, reduced demand characteristics, more natural or representative task experience, and fewer artificial disruptions of thought flow. Therefore, researchers of mind wandering should strive to include as few probes as are needed for reliable and valid measurement of the TUT rate.

To provide provisional guidance to the field, we reanalyze data from two published studies (Kane et al., 2016; 2021) in which we probed subjects in multiple laboratory tasks and assessed a variety of cognitive and non-cognitive correlates of TUT rate. Our approach mirrors that taken by recent investigations into how few task trials of complex span tasks are needed to reliably and validly measure WMC (Foster et al., 2015; Oswald et al., 2015, Study 1); those studies presented all subjects with full-length complex span tasks but analyzed subsets of the tasks' data to determine the fewest trials needed to roughly reproduce the full tasks' correlations with each other and with measures of a related construct (fluid intelligence). In the present study, each analyzed task presented the same number of probes to all subjects, and we analyzed subsets of subjects' probe responses to determine at what number of analyzed probes (i.e., at what probe bin size) do average TUT rates and, of most importance, TUT rates' correlations with other variables, stabilize.

Specifically, regarding reliability, we will focus on: (a) the stability of *M* TUT rates across probe bin sizes (e.g., 2-probe TUT rate vs. 20-probe TUT rate), (b) item-total correlations, considering the correlation of TUT rate from each subset of analyzed probes with that from the largest bin size, within each task (e.g., 2-probe TUT rate × 20-probe TUT rate from the SART), and (c) factor loadings from latent variable models for each task's TUT rate across different probe bin sizes (e.g., the 2-probe TUT rate loadings on a TUT rate factor for

all the indicator tasks vs. the 20-probe TUT rate loadings on a TUT factor for all the indicator tasks). Regarding validity, we focus on the correlations of a TUT rate latent variable with other theoretically relevant constructs across different probe bin sizes (e.g., the TUT rate × WMC latent correlation with TUT rates calculated from 2-probe bins vs. 20-probe bins). In our primary analyses, we selected these analyzed probes randomly for each subject from each task; in secondary analyses that assess the robustness of our findings and conclusions (with details reported in supplemental materials), we analyzed the first *n* probes that appeared within the task for each subject.

Study 1

Study 1 reanalyzes data from Kane et al. (2016), which tested several hundred subjects in three 2-hour lab sessions. Thought probes were presented within five tasks—two in session 1, two in session 2, and one in session 3—and individual-differences constructs included WMC, attention-control inability (higher scores = worse performance), and several dimensions of psychometrically assessed schizotypy.

Methods

The original Kane et al. (2016) study reported how they determined sample sizes, all data exclusions, and all included measures (Simmons et al., 2012). The study received ethics approval from the Institutional Review Board (IRB) at the University of North Carolina at Greensboro (UNCG), a minority-serving comprehensive state university.

Subjects

As reported in Kane et al. (2016), 541 UNCG undergraduates completed the first session, 492 completed the second, and 472 completed the third. Here is the originally reported demographic information:

> Sixty-six percent of our 541 analyzed subjects self-identified as female and 34% as male (5 missing cases), with a mean age of 19 years (sd = 2; 2 missing cases). Also by self-report, the racial composition of the sample was 49% White (European/Middle Eastern descent); 34% Black (African/Caribbean descent); 7% Multiracial; 4% Asian; <1% Native American/Alaskan Native; 0% Native Hawaiian/Pacific Islander; 4% Other (4 missing cases). Finally, self-reported ethnicity, asked separately, was 7% Latino/Hispanic (1 missing case). (Kane et al., 2016, pp. 1026–1027)

Tasks, measures, and procedures

Subjects were each seated at their own workstation and were tested in groups of 1–4. An experimenter remained present throughout the entire session to initiate each task after all subjects had completed the prior one, to read all task instructions aloud, and to monitor subjects' behavior (and record any problems).

For detailed descriptions of all the computer-administered cognitive tasks and schizotypy questionnaires (both analyzed and unanalyzed), as well as their scoring and dependent measures, see Kane et al. (2016). Below we describe the key constructs of interest for the present study—WMC, attention-control inability, and disorganized schizotypy. Across each the three experimental sessions, subjects completed at least one measure of each construct and at least one probed task (except the schizotypy assessments, which were presented only in sessions 1 and 2).

WMC Subjects completed six WMC tasks. Four complex span tasks (operation, reading, symmetry, and rotation span) presented sequences of to-be-remembered items (e.g., letters; spatial locations in a matrix) of varying set sizes for immediate serial recall; prior to each memory item, an unrelated processing task required a yes/no response (e.g., a mathematical equation that was correct or incorrect; an abstract pattern that was vertically symmetrical or not). Two memory-updating tasks (an updating counters task and a running span task) required subjects to maintain an evolving set of stimuli (letters or numbers) of varying set sizes and to abandon no longer relevant stimuli. Across all WMC tasks, higher scores reflected more items accurately recalled in serial order.

Attention control Five tasks required subjects to override a prepotent response in favor of a goal-appropriate one. Subjects completed two antisaccade tasks (requiring identifying stimuli [either arrows or letters] presented to the opposite side of an attention-attracting cue, one task requiring a choice among 3 response options and the other among 4 response options; the dependent variable for each was accuracy rate), a go/no-go SART task (requiring withholding of a key-press response on a minority of semantic-classification trials [animal names appeared on 89% of trials and vegetables appeared on 11%]; dependent variables were d' and intrasubject standard deviation in RT [RTsd]), and two Stroop-like tasks, a number Stroop and a spatial Stroop task (requiring ignoring a salient stimulus dimension in favor of responding to another stimulus dimension; the dependent variable for spatial Stroop was the residual of the incongruent trial error rate regressed on the congruent trial error rate, and for number Stroop was the *M* RT on incongruent trials). Measures for attention control were scored such that higher

scores reflected worse performance (e.g., greater error rate, poorer signal detection, longer or more variable RTs).

Disorganized schizotypy Subjects completed a battery of valid questionnaires assessing multiple dimensions of schizotypy. In latent variable analyses, Kane et al. (2016) found equivalent TUT correlations with disorganized, positive, and paranoid dimensions of schizotypy (all .21–.22) as well as strong correlations among these schizotypy facets (≥.60). For simplicity, then, we investigated only the disorganized dimension here, analyzing data from the following scales: the Schizotypal Personality Questionnaire–Odd Behavior and Odd Speech subscales (Raine, 1991), the Cognitive Slippage Scale (Miers & Raulin, 1987), and the Dimensional Assessment of Personality Pathology-Basic Questionnaire (6 items from the Cognitive Dysregulation subscale; Livesley & Jackson, 2009). Subjects answered "yes" or "no" to each item. Higher scores reflected greater endorsement of behaviors in each dimension.

Probed thought reports Thought probes appeared randomly, with some constraints, within five tasks. In the letter flanker task, 12 probes were presented (following 8.3% of total trials), 4 after congruent trials, 2 after neutral trials, 2 after stimulus-response (S-R) incongruent trials, 2 after stimulus-stimulus (S-S) incongruent trials, and 2 after an unanalyzed trial type. In the SART, 45 probes were presented following no-go target trials (6.6% of total trials). In the number Stroop task, 20 probes were presented in the second block of the task, always following incongruent trials (13% of block 2 trials). In the arrow flanker task, 20 probes were presented across the two blocks; 4 were presented in the first block and 16 in the second (10.4% of total trials). Finally, in the 2-back task, 15 probes were presented, following 6.3% of trials.

Each probe presented the following 8 response options and subjects were told to select the one that most closely aligned with the content of their immediately preceding thoughts by pressing the corresponding number key on the keyboard: (1) "the task" (thought related to the stimuli and goals of the task), (2) "task experience/performance" (evaluative thoughts about one's performance on the task, (3) "everyday things" (thoughts about normal life concerns and activities), (4) "current state of being" (thoughts about one's physical, cognitive, or emotional states), (5) "personal worries" (worried thoughts), (6) "daydreams" (fantastical, unrealistic thoughts), (7) "external environment" (thoughts about environmental stimuli), and (8) "other" (any thoughts not fitting the other categories). As in Kane et al. (2016), we defined TUTs as response options 3–8.

Probe frequency assessment

For our primary analyses, we randomly selected probes for each subject in each task. To use as much data as possible while remaining consistent across the tasks, we randomly selected probes in increments of two, up to 14 probes, for each subject in each task (except for the letter flanker task, which presented only 12 total probes; in the letter flanker task, then, we repeated the bin 12 data for bin 14). Random selection of probes was independent of the previous bin (i.e., the probes for bin size 2 could be completely different from the probes for bin size 4). For each bin of selected probes for each subject, we calculated the TUT rate.

A limitation of this approach is that all subjects responded to the full set of probes for each task, and so their responses to the randomly selected bins of analyzed probes could have been influenced by the appearance of, or their responses to, other probes. Therefore, our secondary analyses selected the first *n* probes that appeared to each subject in each task (in increments of two, up to 14, again except for the letter flanker task, which presented only 12 probes), before other probes could have had any influence on reporting. The secondary analyses yielded similar results to the primary analyses; we thus discuss them below but present details in supplementary materials.

Results and discussion

Data used for these reanalyses, as well as Rmarkdown files for all primary and secondary analyses, are available on the Open Science Framework (https://osf.io/f46e7/). We adopted a .05 α-level throughout. Details regarding data exclusions, task scoring, and outlier treatments can be found in Kane et al. (2016). We modeled the cognitive and schizotypy predictors exactly as in Kane et al. (2016), including any residual correlations among indicators (e.g., between the SART d' score and the SART RTsd score for the attention control construct). We first report descriptive statistics and bivariate correlations among the various probe-bin measures to address questions of reliability, and then address questions about reliability and validity using confirmatory factor analyses (CFAs) for each level of probe bin size (2–14).

Mean TUT rate and TUT-rate correlations across probe bin sizes

As seen in Table 1, mean TUT rates were remarkably consistent across probe bin sizes, including bin size 2, with slight decreases in variation around these estimates as bin size increased. All these TUT-rate estimates are also similar to those reported in Kane et al. (2016) for the full

Table 1 Descriptive statistics for TUT rate by probe bin sizes for each probed task from Study 1

	N	Mean	SD	Median	Skew	Kurtosis
Letter flanker						
Bin size 2	462	0.60	0.38	0.50	−0.35	−1.24
Bin size 4	462	0.59	0.31	0.50	−0.29	−0.90
Bin size 6	462	0.58	0.29	0.67	−0.40	−0.69
Bin size 8	462	0.59	0.28	0.62	−0.45	−0.68
Bin size 10	462	0.59	0.27	0.60	−0.44	−0.69
Bin size 12	462	0.59	0.26	0.67	−0.49	−0.54
SART						
Bin size 2	526	0.52	0.39	0.50	−0.07	−1.35
Bin size 4	526	0.52	0.32	0.50	−0.05	−1.09
Bin size 6	526	0.52	0.30	0.50	0.01	−1.08
Bin size 8	526	0.51	0.28	0.50	−0.08	−0.99
Bin size 10	526	0.51	0.27	0.50	−0.03	−0.91
Bin size 12	526	0.50	0.27	0.50	−0.06	−0.91
Bin size 14	526	0.51	0.26	0.50	−0.09	−0.88
Number Stroop						
Bin size 2	478	0.46	0.41	0.50	0.14	−1.51
Bin size 4	478	0.45	0.35	0.50	0.18	−1.27
Bin size 6	478	0.46	0.33	0.50	0.22	−1.18
Bin size 8	478	0.45	0.33	0.38	0.23	−1.16
Bin size 10	478	0.45	0.32	0.40	0.29	−1.14
Bin size 12	478	0.45	0.32	0.42	0.26	−1.11
Bin size 14	478	0.45	0.32	0.42	0.28	−1.10
Arrow flanker						
Bin size 2	479	0.51	0.41	0.50	−0.05	−1.53
Bin size 4	479	0.50	0.36	0.50	0.04	−1.35
Bin size 6	479	0.50	0.34	0.50	0.02	−1.23
Bin size 8	479	0.49	0.32	0.50	0.06	−1.16
Bin size 10	479	0.49	0.32	0.50	0.11	−1.15
Bin size 12	479	0.49	0.31	0.50	0.10	−1.13
Bin size 14	479	0.48	0.32	0.42	0.11	−1.16
2-Back						
Bin size 2	461	0.41	0.40	0.50	0.32	−1.37
Bin size 4	461	0.42	0.33	0.33	0.28	−1.14
Bin size 6	461	0.42	0.33	0.33	0.28	−1.14
Bin size 8	461	0.42	0.32	0.38	0.31	−1.08
Bin size 10	461	0.42	0.32	0.40	0.30	−1.16
Bin size 12	461	0.42	0.32	0.42	0.29	−1.13
Bin size 14	461	0.42	0.31	0.36	0.29	−1.15

SART Sustained Attention to Response Task.

complement of thought probes in each task. Thus, it appears we can gather reliable and credible point estimates of TUT rates with as few as two thought probes embedded within a task.

Regarding the correlations presented in Table 2, we first consider the *within*-task correlations among probe bin sizes, akin to examining item-total correlations in questionnaire research. Within each task, the TUT-rate correlations between even the smallest bin (2 probes) and the largest bin (14 probes) were strong (rs = .56–.78); correlations with the largest bins generally increased with bin size but became stable at about eight thought probes (rs = .80–.96). These within-task correlations suggest that the TUT-rate variation measured by 14 probes was reliably captured by just 8 probes (and even reasonably captured by only 4 probes).

Table 2 Zero-order correlations among cognitive predictors and TUT rates across tasks and probe bin sizes in Study 1

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1 OPERSPAN	1																			
2 READSPAN	0.58	1																		
3 SYMMSPAN	0.40	0.38	1																	
4 ROTASPAN	0.44	0.32	0.54	1																
5 RUNNSPAN	0.45	0.37	0.27	0.20	1															
6 COUNTERS	0.36	0.23	0.37	0.29	0.39	1														
7 ANTI-LET	−0.21	−0.17	−0.34	−0.21	−0.25	−0.35	1													
8 ANTI-ARO	−0.25	−0.18	−0.30	−0.36	−0.27	−0.33	0.59	1												
9 SART d’	0.15	0.20	0.19	0.14	0.21	0.17	−0.36	−0.27	1											
10 SART RTSD	−0.14	−0.19	−0.21	−0.11	−0.23	−0.21	0.36	0.28	−0.63	1										
11 N-STROOP	−0.17	−0.03	−0.18	−0.18	−0.10	−0.20	0.22	0.26	−0.12	0.21	1									
12 S-STROOP	−0.04	−0.05	−0.08	−0.18	−0.09	−0.07	0.19	0.21	−0.17	0.16	0.08	1								
13 ODBEHAVR	0.01	0.01	0.01	−0.05	−0.07	−0.03	0.09	0.01	−0.03	0.00	0.00	0.07	1							
14 ODSPEECH	−0.02	−0.03	0.01	0.01	−0.10	−0.03	0.05	−0.04	−0.06	0.00	0.01	0.06	0.56	1						
15 COGSLIPG	−0.07	−0.06	0.01	0.06	−0.17	−0.08	0.14	0.05	−0.13	0.09	0.06	0.08	0.46	0.70	1					
16 COGDYSRG	−0.08	−0.10	−0.02	0.02	−0.16	−0.13	0.14	0.02	−0.14	0.09	0.06	0.06	0.43	0.60	0.60	1				
17 LETTER FLANKER 2	0.06	0.07	−0.02	0.00	0.02	−0.03	0.03	0.05	−0.09	0.10	0.00	0.03	0.07	0.05	0.10	0.03	1			
18 LETTER FLANKER 4	0.08	0.04	−0.10	0.02	0.05	−0.08	0.10	0.07	−0.12	0.09	0.06	0.12	0.07	0.08	0.13	0.09	0.55	1		
19 LETTER FLANKER 6	0.08	0.02	−0.09	0.02	0.01	−0.03	0.13	0.09	−0.20	0.17	0.08	0.11	0.11	0.10	0.19	0.12	0.64	0.71	1	
20 LETTER FLANKER 8	0.07	−0.01	−0.08	0.01	0.01	−0.01	0.11	0.09	−0.18	0.14	0.07	0.15	0.12	0.10	0.16	0.10	0.68	0.76	0.86	1
21 LETTER FLANKER 10	0.10	0.02	−0.07	0.00	0.06	−0.01	0.11	0.07	−0.17	0.14	0.07	0.16	0.09	0.10	0.15	0.09	0.68	0.80	0.88	0.93
22 LETTER FLANKER 12	0.08	0.01	−0.09	−0.01	0.04	−0.02	0.11	0.07	−0.19	0.15	0.09	0.16	0.10	0.09	0.16	0.09	0.70	0.81	0.90	0.95
23 SART 2	−0.03	−0.06	−0.09	−0.05	0.00	−0.08	0.18	0.07	−0.23	0.21	0.14	0.09	0.04	0.05	0.14	0.16	0.13	0.21	0.29	0.28
24 SART 4	−0.03	−0.07	−0.10	0.01	−0.02	−0.03	0.09	−0.01	−0.20	0.26	0.03	0.04	0.03	0.06	0.10	0.06	0.24	0.28	0.36	0.40
25 SART 6	−0.06	−0.09	−0.08	−0.03	−0.06	−0.07	0.14	0.07	−0.25	0.23	0.14	0.06	−0.02	0.05	0.12	0.08	0.31	0.34	0.43	0.41
26 SART 8	0.01	−0.12	−0.06	0.04	−0.07	−0.07	0.14	0.04	−0.23	0.27	0.08	0.08	0.04	0.06	0.14	0.11	0.29	0.36	0.42	0.43
27 SART 10	0.01	−0.11	−0.06	−0.01	−0.06	−0.01	0.16	0.09	−0.26	0.28	0.12	0.13	0.07	0.10	0.16	0.15	0.31	0.36	0.43	0.44
28 SART 12	−0.02	−0.15	−0.06	−0.02	−0.03	−0.03	0.13	0.04	−0.24	0.28	0.11	0.08	0.03	0.08	0.16	0.07	0.32	0.35	0.45	0.46
29 SART 14	−0.01	−0.09	−0.08	−0.02	−0.03	−0.07	0.15	0.06	−0.25	0.30	0.12	0.06	0.04	0.07	0.14	0.08	0.30	0.34	0.44	0.46
30 NUMBER STROOP 2	0.04	−0.04	−0.01	0.00	−0.09	0.00	0.08	0.06	−0.15	0.11	0.09	−0.02	0.08	0.13	0.15	0.14	0.15	0.18	0.22	0.25
31 NUMBER STROOP 4	0.00	−0.06	−0.02	−0.01	−0.13	−0.07	0.13	0.11	−0.20	0.17	0.12	0.03	0.06	0.04	0.10	0.08	0.13	0.26	0.26	0.33
32 NUMBER STROOP 6	−0.04	−0.10	−0.07	−0.05	−0.13	−0.02	0.14	0.14	−0.20	0.16	0.15	0.07	0.08	0.10	0.12	0.11	0.08	0.23	0.24	0.30
33 NUMBER STROOP 8	−0.05	−0.12	−0.05	−0.04	−0.10	−0.05	0.14	0.12	−0.19	0.19	0.16	0.07	0.08	0.09	0.12	0.11	0.12	0.26	0.28	0.32
34 NUMBER STROOP 10	−0.01	−0.09	−0.02	−0.03	−0.10	0.00	0.12	0.11	−0.22	0.20	0.14	0.06	0.09	0.08	0.12	0.11	0.15	0.27	0.30	0.35
35 NUMBER STROOP 12	−0.02	−0.10	−0.04	−0.05	−0.12	−0.02	0.14	0.13	−0.21	0.19	0.17	0.06	0.06	0.07	0.10	0.09	0.15	0.26	0.30	0.34
36 NUMBER STROOP 14	−0.02	−0.10	−0.02	−0.04	−0.11	0.00	0.14	0.13	−0.21	0.20	0.15	0.07	0.06	0.07	0.11	0.10	0.14	0.26	0.29	0.32
37 ARROW FLANKER 2	−0.02	−0.08	−0.03	0.02	−0.08	−0.02	0.11	0.03	−0.12	0.10	0.12	−0.01	0.11	0.09	0.16	0.11	0.16	0.26	0.30	0.29
38 ARROW FLANKER 4	−0.01	−0.07	−0.02	0.04	−0.13	−0.04	0.10	0.01	−0.13	0.11	0.15	0.01	0.13	0.12	0.18	0.16	0.15	0.25	0.31	0.30

Table 2 (continued)

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
39 ARROW FLANKER 6	0.00	−0.11	−0.03	0.04	−0.10	−0.03	0.14	0.05	−0.17	0.15	0.13	0.02	0.12	0.08	0.18	0.15	0.18	0.23	0.32	0.33
40 ARROW FLANKER 8	0.03	−0.05	−0.02	0.01	−0.09	−0.02	0.13	0.06	−0.16	0.12	0.16	0.03	0.10	0.08	0.16	0.13	0.20	0.26	0.33	0.33
41 ARROW FLANKER 10	0.01	−0.06	−0.03	0.05	−0.12	−0.06	0.15	0.07	−0.18	0.17	0.16	0.06	0.14	0.10	0.19	0.16	0.21	0.25	0.34	0.34
42 ARROW FLANKER 12	0.02	−0.07	−0.02	0.03	−0.09	−0.04	0.15	0.04	−0.17	0.14	0.14	0.04	0.14	0.10	0.20	0.16	0.21	0.27	0.36	0.35
43 ARROW FLANKER 14	−0.01	−0.09	−0.03	0.02	−0.10	−0.04	0.15	0.06	−0.19	0.16	0.15	0.00	0.12	0.09	0.18	0.17	0.22	0.26	0.36	0.36
44 2-BACK 2	0.04	0.00	−0.07	−0.13	−0.13	−0.09	0.18	0.15	−0.22	0.20	0.08	0.22	0.12	0.09	0.11	0.09	0.14	0.18	0.22	0.24
45 2-BACK 4	−0.03	−0.09	−0.03	−0.08	−0.21	−0.08	0.17	0.19	−0.28	0.24	0.12	0.20	0.05	0.05	0.07	0.06	0.10	0.19	0.25	0.27
46 2-BACK 6	−0.06	−0.11	−0.03	−0.11	−0.20	−0.12	0.16	0.17	−0.24	0.23	0.10	0.25	0.06	0.06	0.09	0.06	0.20	0.25	0.27	0.31
47 2-BACK 8	−0.04	−0.08	−0.06	−0.14	−0.20	−0.13	0.19	0.21	−0.28	0.24	0.12	0.27	0.06	0.06	0.08	0.06	0.17	0.23	0.26	0.29
48 2-BACK 10	−0.06	−0.09	−0.06	−0.14	−0.20	−0.12	0.18	0.20	−0.28	0.26	0.14	0.26	0.05	0.05	0.07	0.05	0.16	0.25	0.26	0.29
49 2-BACK 12	−0.06	−0.10	−0.06	−0.15	−0.21	−0.13	0.17	0.20	−0.28	0.27	0.13	0.25	0.05	0.05	0.07	0.05	0.17	0.24	0.26	0.29
50 2-BACK 14	−0.06	−0.10	−0.07	−0.14	−0.21	−0.13	0.19	0.21	−0.28	0.27	0.13	0.25	0.06	0.05	0.08	0.05	0.17	0.24	0.26	0.30

Table 2 (continued)

Variable	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
1 OPERSPAN																				
2 READSPAN																				
3 SYMMSPAN																				
4 ROTASPAN																				
5 RUNNSPAN																				
6 COUNTERS																				
7 ANTI-LET																				
8 ANTI-ARO																				
9 SART d’																				
10 SART RTSD																				
11 N-STROOP																				
12 S-STROOP																				
13 ODBEHAVR																				
14 ODSPEECH																				
15 COGSLIPG																				
16 COGDYSRG																				
17 LETTER FLANKER 2																				
18 LETTER FLANKER 4																				
19 LETTER FLANKER 6																				
20 LETTER FLANKER 8																				
21 LETTER FLANKER 10	1																			
22 LETTER FLANKER 12	0.98	1																		
23 SART 2	0.30	0.31	1																	
24 SART 4	0.39	0.40	0.43	1																
25 SART 6	0.45	0.44	0.51	0.60	1															
26 SART 8	0.45	0.45	0.52	0.67	0.72	1														
27 SART 10	0.48	0.48	0.54	0.63	0.74	0.76	1													
28 SART 12	0.48	0.49	0.54	0.68	0.77	0.79	0.81	1												
29 SART 14	0.47	0.48	0.56	0.69	0.78	0.80	0.83	0.86	1											
30 NUMBER STROOP 2	0.23	0.24	0.19	0.28	0.30	0.30	0.33	0.28	0.31	1										
31 NUMBER STROOP 4	0.28	0.30	0.25	0.30	0.32	0.35	0.35	0.35	0.36	0.68	1									
32 NUMBER STROOP 6	0.27	0.28	0.21	0.33	0.37	0.34	0.37	0.39	0.38	0.70	0.79	1								
33 NUMBER STROOP 8	0.30	0.31	0.22	0.36	0.37	0.38	0.38	0.40	0.39	0.71	0.81	0.86	1							
34 NUMBER STROOP 10	0.32	0.33	0.23	0.34	0.38	0.39	0.40	0.40	0.40	0.73	0.82	0.88	0.90	1						
35 NUMBER STROOP 12	0.31	0.33	0.26	0.35	0.40	0.40	0.40	0.42	0.42	0.75	0.83	0.89	0.92	0.93	1					
36 NUMBER STROOP 14	0.30	0.31	0.24	0.35	0.37	0.38	0.39	0.41	0.40	0.74	0.85	0.89	0.91	0.93	0.95	1				
37 ARROW FLANKER 2	0.28	0.30	0.18	0.26	0.28	0.27	0.27	0.30	0.30	0.45	0.47	0.50	0.51	0.52	0.53	0.53	1			
38 ARROW FLANKER 4	0.29	0.31	0.21	0.30	0.32	0.33	0.36	0.34	0.33	0.49	0.49	0.54	0.55	0.56	0.57	0.56	0.67	1		

Table 2 (continued)

Variable	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
39 ARROW FLANKER 6	0.31	0.33	0.22	0.33	0.35	0.35	0.39	0.38	0.38	0.50	0.54	0.59	0.60	0.60	0.62	0.62	0.66	0.79	1	
40 ARROW FLANKER 8	0.31	0.33	0.21	0.29	0.33	0.35	0.35	0.35	0.36	0.50	0.54	0.59	0.58	0.61	0.63	0.62	0.71	0.81	0.86	1
41 ARROW FLANKER 10	0.33	0.35	0.22	0.32	0.32	0.36	0.37	0.35	0.38	0.52	0.56	0.59	0.60	0.61	0.63	0.63	0.72	0.83	0.87	0.91
42 ARROW FLANKER 12	0.34	0.36	0.21	0.33	0.34	0.37	0.39	0.37	0.38	0.54	0.57	0.61	0.62	0.63	0.65	0.65	0.73	0.86	0.89	0.91
43 ARROW FLANKER 14	0.33	0.36	0.22	0.32	0.34	0.37	0.38	0.37	0.38	0.53	0.56	0.60	0.63	0.63	0.65	0.65	0.72	0.85	0.88	0.92
44 2-BACK 2	0.25	0.25	0.16	0.26	0.27	0.27	0.25	0.32	0.30	0.23	0.29	0.32	0.34	0.35	0.35	0.34	0.28	0.29	0.28	0.33
45 2-BACK 4	0.25	0.27	0.20	0.24	0.24	0.27	0.29	0.29	0.30	0.22	0.31	0.32	0.35	0.36	0.36	0.36	0.25	0.30	0.32	0.35
46 2-BACK 6	0.30	0.32	0.20	0.25	0.29	0.31	0.31	0.32	0.32	0.24	0.34	0.36	0.39	0.40	0.40	0.40	0.32	0.34	0.34	0.37
47 2-BACK 8	0.29	0.31	0.21	0.28	0.28	0.31	0.31	0.32	0.33	0.26	0.37	0.38	0.41	0.41	0.43	0.42	0.31	0.35	0.36	0.39
48 2-BACK 10	0.30	0.32	0.23	0.28	0.30	0.33	0.34	0.34	0.35	0.26	0.34	0.37	0.39	0.41	0.41	0.40	0.31	0.34	0.33	0.38
49 2-BACK 12	0.29	0.31	0.22	0.28	0.29	0.32	0.32	0.33	0.34	0.27	0.36	0.38	0.40	0.41	0.42	0.42	0.30	0.35	0.35	0.39
50 2-BACK 14	0.30	0.32	0.22	0.29	0.31	0.33	0.33	0.35	0.35	0.27	0.37	0.39	0.42	0.42	0.44	0.42	0.32	0.36	0.37	0.40

Table 2 (continued)

Variable	41	42	43	44	45	46	47	48	49	50
41 ARROW FLANKER 10	1									
42 ARROW FLANKER 12	0.93	1								
43 ARROW FLANKER 14	0.93	0.94	1							
44 2-BACK 2	0.34	0.33	0.31	1						
45 2-BACK 4	0.35	0.33	0.34	0.66	1					
46 2-BACK 6	0.40	0.38	0.37	0.71	0.81	1				
47 2-BACK 8	0.41	0.39	0.38	0.73	0.83	0.89	1			
48 2-BACK 10	0.38	0.37	0.36	0.74	0.85	0.91	0.94	1		
49 2-BACK 12	0.40	0.38	0.38	0.73	0.86	0.92	0.95	0.96	1	
50 2-BACK 14	0.41	0.40	0.39	0.76	0.87	0.93	0.96	0.97	0.98	1

OPERSPAN operation span, *READSPAN* reading span, *SYMMSPAN* symmetry span, *ROTASPAN* rotation span, *RUNNSPAN* running span, *COUNTERS* updating counters, *ANTI-LET* antisaccade with letters, *ANTI-ARO* antisaccade with arrows *SART d'* d' score from semantic SART, *SART rtsd* intrasubject standard deviation in RT from semantic SART, *N-Stroop* number Stroop, *S-Stroop* spatial Stroop, *ODBEHAVR* SPQ odd behavior subscale, *ODSPEECH* SPQ odd speech subscale, *COGSLIPG* cognitive slippage scale, *COGDYSRG* cognitive dysregulation subscale of the Dimensional Assessment of Personality Pathology – Basic Questionnaire

We next consider *between*-task TUT-rate correlations within each probe bin size. The cross-task correlations are weakest in bin size 2 (*Mdn r* = .17) and increase to bin size 4 (*Mdn r* = .29), and to bin size 6 and 8 (*Mdn r*s = .36 and .37, respectively). Correlations change little but are numerically strongest in bin sizes 10–14 (*Mdn rs* = .39–.40). These analyses suggest viable estimates of TUT-rate variation with as few as 6–8 probes per task. Across within- and between-task comparisons, then, measuring TUTs with 8 probes may be an optimal approach.

Confirmatory factor analyses across probe bins

Zero-order correlations among TUT-rate assessments suggest we can more reliably capture individual differences in mind wandering propensity using 8 or more probes per task. Here, we tested how TUT-rate-indicator factor loadings, and TUT-rate correlations with cognitive and schizotypy predictors, changed as we included more random thought-probe responses from each task. To do this we ran a series of CFAs in *lavaan* (Rosseel, 2012), where a TUT-rate factor was modeled at each level of probe bin size (2–14) from each of the 5 probed tasks, and correlated with factors for WMC, attention control, and disorganized schizotypy. As seen in Table 3, all models adequately fit the data (Schermelleh-Engel et al., 2003). Figure 1 displays the overall structural model with path estimates from each model, from probe bin sizes 2 to 14 (for clarity, the factor loadings for each model are presented in Table 4).

As in the simple bivariate correlations, loadings for each task's TUT rate on the TUT rate factor were low but mostly acceptable for the 2–4 probe bins, and most consistent from 6–8 probe bins to 14 probe bins, and most consistent with the factor loadings (see Table 4) from the full complement of probes used by Kane et al. (2016). The path estimates between our predictor constructs and the TUT rate factor were also reasonably similar across all probe bin sizes (see Fig. 1), but they generally stabilize within a .02 window from 8 probes upward. For additional comparison, the TUT-rate correlations from the full Kane et al. (2016) model were −.17 for WMC (vs. −.17 from the 8-probe bin), .37 for attention control (vs. .38 from the 8-probe bin), and .21 for disorganized schizotypy (vs. .20 from the 8-probe bin). We provisionally conclude that researchers can reliably and validly estimate individual differences in TUT rate with as few as 8 thought probes, at least in latent-variable studies that assess TUT rates in many tasks.

Secondary analysis of first-n probes

Supplemental Tables 1 and 2 present *M* TUT rates and TUT-rate correlations, respectively, across probe-bin sizes, with probes drawn consecutively from the beginning of each task,

Table 3 Fit statistics for latent variable models from Study 1

Model	χ^2	df	RMSEA [95% CI]	SRMR	CFI	TLI
2 Probes	224.289	111	.044 [.035–.052]	.049	.930	.914
4 Probes	262.846	111	.050 [.043–.058]	.055	.914	.894
6 Probes	241.114	111	.047 [.039–.055]	.053	.932	.917
8 Probes	258.353	111	.050 [.042–.058]	.055	.925	.908
10 Probes	254.789	111	.049 [.041–.057]	.056	.929	.913
12 Probes	272.583	111	.052 [.044–.060]	.057	.922	.905
14 Probes	261.155	111	.050 [.042–.058]	.055	.928	.912

rather than randomly (e.g., bin size 2 reflects data from the first two probes presented in each task). TUT rates changed more here across probe-bin sizes than they did in the randomly selected probe analyses, as expected from findings that TUT rates increase with time on task (e.g., Lindquist & McLean, 2011; McVay & Kane, 2009, 2012; Risko et al., 2012). Despite this general increase, TUT rates appeared to stabilize by bin size 6–10, depending on the task, and TUT-rate correlations across bins within tasks stabilized (with *r*s ≥ .90 with bin size 14 TUT rate) by bin size 6–8, as with the randomly selected probes. Between-task TUT correlations stabilized in the .30 range for bin sizes 10–14 (*Mdn r*s = .32–.36), also like the randomly selected probes.

CFAs on these data indicated adequate fit for all models (with TUT rate indicators based on 2–14 probe bins per task; see Supplemental Table 3). Supplemental Figure 1 presents the overall structural model with path estimates and Supplemental Table 4 presents the factor loadings. All TUT-rate factor loadings exceeded .45 for models based on bin sizes 8–14, but TUT-rate loadings were most consistent for models with 10 or more probes (vs. bin sizes 6–8 from randomly selected probes). Path estimates for the correlations between TUT rate and all the other constructs (WMC, attention control, disorganized schizotypy) appeared to stabilize with estimates within a .02 window for bin sizes 10–14 (and for bin sizes ≥ 8 for the WMC and disorganized schizotypy correlations).

Overall, then, despite TUT rates increasing over bin sizes within each task, the correlational findings here strongly replicate those from randomly selected probes. TUT rates calculated from bins of 8 thought probes efficiently demonstrate nearly as strong reliability and validity as those calculated from bins of 14 (and even TUT rates calculated from bins as small as 4 or 6 provide reasonable reliability and validity). These findings indicate that the promising results from randomly selected probes aren't driven by the appearance of other unmeasured probe responses in the tasks.

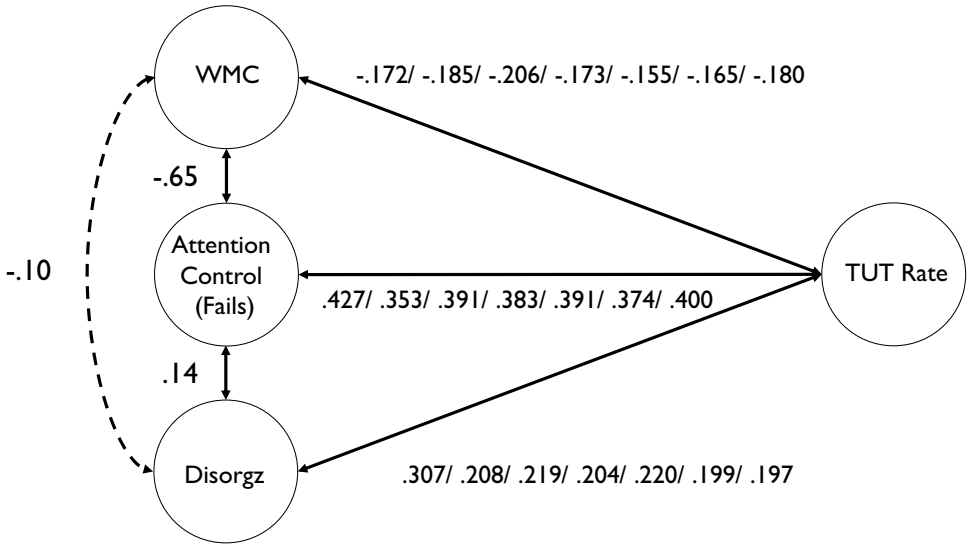

Fig. 1 Standardized path estimates from confirmatory factor analyses between WMC, attention control failures (Fails), disorganized schizotypy (Disorgz), and TUT-rate bins (2–14). Estimates from each model are separated by the slash starting with the probe bin size 2 model and ending with the probe bin size 14 model. Factor loadings for each model are presented in Table 4

Table 4 Standardized factor loadings (and standard errors) for latent variable models for Study 1

Construct and measure	Confirmatory factor analysis models							
	2 Probes	4 Probes	6 Probes	8 Probes	10 Probes	12 Probes	14 Probes	Kane et al. (2016)
Working memory capacity								
OPERSPAN	.64 (.04)	.63 (.04)	.63 (.04)	.63 (.04)	.64 (.04)	.63 (.05)	.64 (.04)	.63 (.05)
READSPAN	.51 (.05)	.51 (.05)	.51 (.05)	.51 (.05)	.51 (.05)	.51 (.05)	.51 (.05)	.51 (.05)
SYMMSPAN	.62 (.04)	.62 (.04)	.62 (.04)	.62 (.04)	.62 (.04)	.62 (.04)	.62 (.04)	.62 (.05)
ROTASPAN	.54 (.05)	.54 (.05)	.54 (.05)	.54 (.05)	.54 (.05)	.54 (.05)	.54 (.05)	.54 (.06)
RUNNSPAN	.59 (.04)	.59 (.04)	.59 (.04)	.59 (.04)	.59 (.04)	.59 (.04)	.59 (.04)	.59 (.05)
COUNTERS	.61 (.04)	.61 (.04)	.61 (.04)	.61 (.04)	.61 (.04)	.61 (.04)	.61 (.04)	.61 (.04)
Attention control (failures)								
ANTI-LET	.77 (.03)	.77 (.03)	.76 (.03)	.76 (.04)	.76 (.03)	.76 (.04)	.76 (.03)	.77 (.03)
ANTI-ARO	.74 (.03)	.74 (.03)	.73 (.03)	.73 (.03)	.73 (.03)	.73 (.04)	.73 (.04)	.73 (.04)
SART d'	−.46 (.04)	−.46 (.04)	−.47 (.04)	−.47 (.04)	−.47 (.04)	−.47 (.04)	−.47 (.04)	−.47 (.05)
SART rtsd	.47 (.04)	.47 (.04)	.47 (.04)	.47 (.04)	.48 (.04)	.48 (.04)	.48 (.04)	.48 (.04)
N-STROOP	.34 (.05)	.34 (.05)	.34 (.05)	.34 (.05)	.34 (.05)	.34 (.05)	.34 (.05)	.33 (.05)
S-STROOP	.26 (.05)	.26 (.05)	.27 (.05)	.27 (.05)	.27 (.05)	.27 (.05)	.27 (.05)	.26 (.05)
Disorganized schizotypy								
ODBEHAVR	.62 (.03)	.62 (.03)	.62 (.03)	.62 (.03)	.62 (.03)	.62 (.03)	.62 (.03)	.63 (.03)
ODSPEECH	.86 (.02)	.86 (.02)	.86 (.02)	.86 (.02)	.86 (.02)	.86 (.02)	.86 (.02)	.83 (.02)
COGSLIPG	.81 (.02)	.81 (.02)	.81 (.02)	.81 (.02)	.81 (.02)	.81 (.02)	.81 (.02)	.83 (.02)
COGDYSRG	.72 (.03)	.72 (.03)	.72 (.03)	.72 (.03)	.72 (.03)	.72 (.03)	.72 (.03)	.74 (.03)
TUT rate								
LETTER FLANKER	.30 (.07)	.58 (.06)	.48 (.06)	.48 (.05)	.49 (.05)	.50 (.05)	.50 (.05)	.50 (.06)
SART	.41 (.06)	.49 (.05)	.57 (.05)	.55 (.05)	.60 (.04)	.59 (.04)	.60 (.04)	.64 (.04)
NUMBER STROOP	.44 (.06)	.59 (.06)	.63 (.05)	.66 (.05)	.64 (.04)	.68 (.04)	.66 (.04)	.68 (.05)
ARROW FLANKER	.48 (.06)	.58 (.06)	.62 (.05)	.63 (.05)	.63 (.05)	.65 (.05)	.64 (.04)	.67 (.05)
2-BACK	.52 (.06)	.52 (.05)	.56 (.05)	.62 (.04)	.61 (.04)	.61 (.04)	.63 (.04)	.64 (.05)

OPERSPAN operation span, *READSPAN* reading span, *SYMMSPAN* symmetry span, *ROTASPAN* rotation span, *RUNNSPAN* running span, *COUNTERS* updating counters, *ANTI-LET* antisaccade with letters, *ANTI-ARO* antisaccade with arrows, *SART d'* d' score from SART, *SART rtsd* intrasubject standard deviation in RT from SART, *N-Stroop* number Stroop, *S-Stroop* spatial Stroop, *ODBEHAVR* SPQ odd behavior subscale, *ODSPEECH* SPQ odd speech subscale, *COGSLIPG* cognitive slippage scale, *COGDYSRG* cognitive dysregulation subscale of the Dimensional Assessment of Personality Pathology – Basic Questionnaire. *SART* Sustained Attention to Response Task

Study 2

Study 2 reanalyzes data from Kane et al. (2021), which tested over 1000 subjects from two public universities in North Carolina in a single lab session. Thought probes were presented within two tasks, and individual-differences constructs included an attention-control ability factor and retrospective self-reports of mind wandering after each probed task. The original study manipulated the type of thought probe that appeared between subjects, creating four experimental groups; here we analyze data from two of the four groups (each with *n*s > 265), in which subjects responded to thought probes asking about thought-content categories, like those used in Study 1 (i.e., the two conditions not analyzed here used different kinds of probes to assess mind wandering).

Methods

This study received ethics approval by the IRBs at UNCG and Western Carolina University (WCU).

Subjects

As reported in Kane et al. (2021), 760 undergraduates from UNCG and 348 from WCU (total $n = 1067$ following exclusions described in Kane et al., 2021), completed a single laboratory session. For the current study, we analyzed data from two of the four experimental conditions (*n*s = 266 and 269 in Conditions 1 and 2, respectively).

Tasks, measures, and procedures

As in Study 1, subjects were each seated at a workstation and were tested in groups of 1–4, and an experimenter remained present for instructions, task pacing, and subject monitoring. See Kane et al. (2021) for a detailed description of all computer-administered measures in the original study. Below we describe the key constructs for the present study—attention-control ability and retrospective mind-wandering reports; the ability measures used here were nearly identical to those from Study 1.

Attention-control tasks Subjects completed two antisaccade tasks (one with letter stimuli and one with arrow stimuli; one requiring a choice among 3 response options and the other among 4 response options) and a go/no-go SART, all identical in structure to those presented in Kane et al. (2016).[1] Subjects also completed an arrow flanker task in which they responded to a centrally presented arrow ("<" or ">") flanked by four distractors. This task served as a secondary source of thought-probe measurement and performance data were not analyzed from this task. The attention control latent variable was modeled identically to Kane et al. (2021): accuracy on the two antisaccade tasks and performance on the SART (d' and RTsd across correct "go" trials). Subjects completed the attention-control tasks in the following order (in each condition): antisaccade-letters, SART, antisaccade-arrows, arrow flanker.

Dundee State Stress Questionnaire (DSSQ) Immediately following completion of the two tasks with thought probes (SART and arrow flanker; see below), subjects answered a set of 12 questions about their conscious experiences in the immediately preceding task. Subjects responded by clicking on their choice along a 1–5 scale labeled, "Never," "Once," "A Few Times," "Often," and "Very Often." The six DSSQ items about TUT experiences were analyzed here (we do not analyze the six items about "task-related interference," i.e., thoughts about task performance). The dependent variable was the mean of the six TUT-item ratings, with higher scores reflecting more frequent TUT experiences during the preceding task.

Probed thought reports Thought probes appeared randomly, with some constraints, within two tasks: In the SART, 45 probes were presented following rare no-go target trials (following 6.6% of total trials); in the arrow flanker task, 4 probes were presented in the first block of 92 trials and 16 probes were presented in the second block of 92 trials (10.4% of total trials). In the arrow flanker task, half the probes followed incongruent trials and half followed neutral trials. As noted above, here we analyze TUT rate data from two of the four between-subject experimental conditions, which presented similar categorical response options to Kane et al. (2016). Specifically, Condition 1 presented the identical probes to Kane et al. (2016), and Condition 2 presented these probes with one thought category removed: task experience/performance (i.e., thoughts about one's task performance; this was response option 2 in Condition 1). TUTs were again defined as response options 3–8 in Condition 1, and as response options 2–7 in Condition 2.

Thought probe frequency assessment for the current study

As in Study 1, we again randomly sampled thought-probe data in increments of two probes from each subject's SART and arrow flanker task. Here, however, we sampled up to 20 probes in each task, as the arrow flanker task presented 20 probes. As in Study 1, we will also report secondary analyses that selected the first *n* probes that appeared to each subject in each task (in increments of 2, from 2–20); again, details of these analyses are available in the supplementary materials.

Results and discussion

As in Study 1, we first report descriptive statistics for TUT rates at each probe bin size, and the bivariate correlations among the bin-size TUT rates. Then we assess CFAs using each bin-size TUT rate (along with factors for attention control ability and retrospective mind wandering reports). We report all analyses separately for Condition 1 and Condition 2.

Mean TUT rate and TUT-rate correlations across probe bin sizes

Table 5 presents descriptive statistics for TUT rate from the SART and arrow flanker tasks, separately for Conditions 1 and 2. For comparison, *M* TUT rates (as proportions) reported in Kane et al. (2021) using all available thought probe data for Condition 1 were .52 and .45 in the SART and arrow flanker tasks, respectively; in Condition 2, *M* TUT rates were .58 and .47 in the SART and arrow flanker tasks, respectively. Here, again, TUT rates were remarkably similar across probe bins, whether estimated from 2 probes or 20; standard deviations around those means tended to narrow from 4 probes upward, but from there remained reasonably stable.

As seen in Table 6, within-task correlations also suggested a similar pattern of results to Study 1. Within both the SART and arrow flanker tasks (in both conditions), correlations across probe bins got stronger with more probes and

[1] As noted in Kane et al. (2021), a programming error in one of the antisaccade tasks led the stimuli to be presented at different distances from central fixation at the different sites; performance data were thus standardized within sites for this task.

Table 5 Descriptive statistics for TUT rate by probe bin sizes for each probed task by experimental condition from Study 2 (Condition 1 $N = 266$; Condition 2 $N = 269$)

	Mean	SD	Median	Skew	Kurtosis
Condition 1					
SART					
Bin size 2	0.53	0.38	0.50	−0.11	−1.27
Bin size 4	0.53	0.31	0.50	−0.22	−0.89
Bin size 6	0.52	0.29	0.50	−0.07	−1.10
Bin size 8	0.50	0.29	0.50	−0.05	−1.01
Bin size 10	0.53	0.27	0.60	−0.18	−0.88
Bin size 12	0.52	0.26	0.58	−0.15	−0.81
Bin size 14	0.52	0.26	0.50	−0.11	−0.87
Bin size 16	0.52	0.26	0.53	−0.13	−0.92
Bin size 18	0.52	0.25	0.56	−0.24	−0.78
Bin size 20	0.52	0.26	0.55	−0.17	−0.95
Flanker					
Bin size 2	0.45	0.40	0.50	0.20	−1.45
Bin size 4	0.45	0.35	0.50	0.23	−1.26
Bin size 6	0.45	0.33	0.50	0.13	−1.26
Bin size 8	0.44	0.31	0.38	0.22	−1.12
Bin size 10	0.46	0.31	0.40	0.14	−1.12
Bin size 12	0.45	0.31	0.42	0.16	−1.17
Bin size 14	0.45	0.31	0.43	0.16	−1.16
Bin size 16	0.45	0.30	0.44	0.14	−1.14
Bin size 18	0.45	0.30	0.44	0.16	−1.14
Bin size 20	0.45	0.30	0.45	0.17	−1.11
Condition 2					
SART					
Bin size 2	0.60	0.39	0.50	−0.37	−1.25
Bin size 4	0.63	0.30	0.75	−0.53	−0.67
Bin size 6	0.61	0.28	0.67	−0.63	−0.41
Bin size 8	0.60	0.27	0.62	−0.42	−0.56
Bin size 10	0.60	0.26	0.60	−0.41	−0.51
Bin size 12	0.60	0.25	0.67	−0.54	−0.45
Bin size 14	0.61	0.25	0.64	−0.44	−0.48
Bin size 16	0.62	0.24	0.62	−0.64	−0.11
Bin size 18	0.61	0.24	0.61	−0.51	−0.25
Bin size 20	0.60	0.24	0.65	−0.57	−0.27
Flanker					
Bin size 2	0.48	0.40	0.50	0.07	−1.44
Bin size 4	0.46	0.32	0.50	0.19	−1.04
Bin size 6	0.48	0.31	0.50	0.09	−1.12
Bin size 8	0.47	0.30	0.50	0.08	−1.03
Bin size 10	0.48	0.29	0.50	0.11	−1.02
Bin size 12	0.47	0.29	0.42	0.17	−0.96
Bin size 14	0.48	0.28	0.50	0.05	−0.98
Bin size 16	0.48	0.28	0.50	0.12	−0.94
Bin size 18	0.48	0.27	0.50	0.09	−0.92
Bin size 20	0.48	0.27	0.45	0.10	−0.94

SART Sustained Attention to Response Task, *Flanker* arrow flanker

this pattern appeared to level out around 8 thought probes; for example, for the SART in both conditions, TUT-rate correlations with the 20-probe bin were at .80 or higher from bin sizes ≥ 8, and for the arrow flanker task in both conditions, TUT-rate correlations with the 20-probe bin were at .90 or higher from bin size ≥ 8. We note again, however, that even the correlations between the smallest and largest probe bins were quite strong (rs ≈ .55 in both SARTs and rs ≈ .75 in both arrow flanker tasks). As in Study 1, these findings suggest that we can reliably capture individual differences in the TUT rate from as few as 8 thought probes (at least, about as reliably as we can measure them from 20 thought probes).

Examining TUT-rate correlations between the SART and flanker tasks, we again found increasing magnitudes with increasing probe bin size. In Condition 1, there was an increase from the 2-probe bin ($r = .25$) to the 4-probe bin ($r = .44$) and then a further increase at the 10-probe bin ($r = .50$), but additional probes beyond 10 did not increase the correlation substantially. In Condition 2, there was a jump in the correlations from the 6-probe bin ($r = .36$) to the 8-probe bin ($r = .44$) and a larger jump once 18 probes were assessed ($r = .57$). Here, then, 8–10 probes appeared to efficiently capture shared cross-task variance in the TUT rate.

Confirmatory factor analyses across probe bins

In parallel to Study 1, here we tested how correlations between TUT rates and attention-control ability and retrospective mind-wandering ratings (from the DSSQ) change as we include more randomly selected thought probes into our TUT-rate measurement from each task. We again ran a series of CFAs in *lavaan* where TUT rates were modeled at each level of probe bin (i.e., 2–20 probes) and examined the TUT rate correlations with attention-control and retrospective mind-wandering ratings. In all models, the unstandardized factor loadings for the two TUT indicators and the two DSSQ indicators were set to be equal.

Table 7 presents the results fit statistics for each model, in each condition. Overall, model fit was generally acceptable by traditional standards (except for the RMSEA indices for some Condition 1 models). However, for Condition 1, model fit tended to decrease with increasing probe-number bins, and two models for Condition 2 (2 Probes and 4 Probes) produced CFIs = 1 and TLIs > 1, suggesting some degree of overfitting. Figure 2 presents the general structural model with path estimates for all the models derived from the Condition 1 data across probe-number bins (for clarity, standardized factor loadings are presented separately in Table 8).

Factor loadings for each task's TUT rate in Condition 1 were reasonably stable when based on 4 or more probes, but they most closely matched the 20-probe loadings with 10 or more probes. Correlations between

Table 6 Zero-order correlations among cognitive and self-report predictors and TUT rates across tasks and probe bin sizes in Study 2, for Condition 1 (n = 266; below diagonal) and Condition 2 (n = 269; above diagonal)

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
1 ANTI-LET	1	0.69	0.49	0.41	−0.10	−0.10	−0.16	−0.11	−0.20	−0.15	−0.17	−0.15	−0.17	−0.19	−0.18	−0.19	−0.08	−0.05	−0.03	−0.09	-0.03	-0.03	-0.03	-0.05	-0.05	-0.05
2 ANTI-ARO	0.61	1	0.43	0.36	−0.05	−0.06	−0.05	−0.03	−0.06	−0.02	−0.07	−0.10	−0.08	−0.08	−0.09	−0.11	−0.05	−0.05	−0.01	−0.07	-0.02	0.00	-0.03	-0.01	-0.03	-0.02
3 SART RTSD	0.38	0.30	1	0.57	−0.09	−0.09	−0.12	−0.17	−0.24	−0.20	−0.23	−0.25	−0.21	−0.21	−0.20	−0.23	−0.07	−0.07	−0.11	−0.15	-0.15	-0.11	-0.13	-0.12	-0.13	-0.13
4 SART d'	0.33	0.21	0.51	1	0.00	−0.05	−0.11	−0.04	−0.16	−0.07	−0.14	−0.14	−0.13	−0.10	−0.13	−0.15	−0.10	−0.02	−0.06	−0.08	-0.08	-0.02	-0.04	-0.03	-0.03	-0.04
5 SART DSSQ	−0.11	−0.07	−0.03	0.04	1	0.70	0.18	0.28	0.27	0.20	0.30	0.22	0.24	0.28	0.25	0.26	0.18	0.18	0.19	0.19	0.23	0.22	0.24	0.24	0.23	0.24
6 FLANKER DSSQ	−0.18	−0.10	−0.08	−0.08	0.70	1	0.20	0.33	0.29	0.23	0.31	0.28	0.30	0.34	0.29	0.31	0.19	0.24	0.27	0.29	0.31	0.31	0.32	0.33	0.32	0.32
7 SART 2	−0.17	−0.14	−0.21	−0.16	0.21	0.24	1	0.52	0.51	0.51	0.52	0.57	0.55	0.52	0.61	0.59	0.37	0.35	0.38	0.33	0.37	0.39	0.38	0.42	0.40	0.41
8 SART 4	−0.25	−0.11	−0.24	−0.22	0.25	0.31	0.47	1	0.59	0.63	0.64	0.63	0.70	0.69	0.69	0.72	0.34	0.33	0.42	0.40	0.41	0.43	0.43	0.45	0.45	0.45
9 SART 6	−0.22	−0.17	−0.35	−0.30	0.24	0.29	0.46	0.65	1	0.62	0.72	0.74	0.74	0.70	0.76	0.78	0.30	0.31	0.36	0.40	0.42	0.39	0.39	0.42	0.42	0.41
10 SART 8	−0.23	−0.11	−0.29	−0.26	0.31	0.33	0.46	0.69	0.69	1	0.71	0.78	0.76	0.75	0.77	0.80	0.31	0.37	0.44	0.44	0.46	0.48	0.45	0.47	0.49	0.49
11 SART 10	−0.24	−0.16	−0.34	−0.26	0.34	0.35	0.52	0.70	0.72	0.77	1	0.79	0.82	0.80	0.81	0.84	0.33	0.35	0.40	0.39	0.43	0.43	0.42	0.45	0.46	0.45
12 SART 12	−0.31	−0.17	−0.39	−0.28	0.29	0.33	0.53	0.71	0.76	0.78	0.81	1	0.85	0.83	0.86	0.87	0.32	0.39	0.45	0.46	0.46	0.48	0.46	0.49	0.50	0.50
13 SART 14	−0.29	−0.20	−0.37	−0.31	0.28	0.34	0.54	0.72	0.75	0.81	0.82	0.86	1	0.86	0.87	0.87	0.32	0.40	0.46	0.46	0.47	0.49	0.49	0.50	0.51	0.51
14 SART 16	−0.27	−0.14	−0.36	−0.28	0.33	0.36	0.53	0.73	0.75	0.82	0.82	0.84	0.88	1	0.86	0.88	0.34	0.39	0.46	0.48	0.48	0.49	0.48	0.51	0.52	0.52
15 SART 18	−0.24	−0.14	−0.34	−0.26	0.33	0.35	0.54	0.73	0.78	0.81	0.86	0.86	0.86	0.88	1	0.90	0.38	0.45	0.50	0.52	0.52	0.54	0.53	0.56	0.57	0.56
16 SART 20	−0.27	−0.14	−0.34	−0.30	0.33	0.37	0.56	0.75	0.78	0.84	0.85	0.86	0.89	0.90	0.91	1	0.38	0.44	0.48	0.49	0.52	0.52	0.51	0.54	0.55	0.55
17 FLANKER 2	−0.20	−0.11	−0.18	−0.10	0.15	0.33	0.25	0.36	0.32	0.31	0.44	0.41	0.38	0.45	0.45	0.43	1	0.60	0.65	0.68	0.70	0.71	0.73	0.72	0.73	0.74
18 FLANKER 4	−0.20	−0.10	−0.16	−0.15	0.20	0.33	0.28	0.44	0.35	0.35	0.44	0.45	0.44	0.48	0.50	0.46	0.61	1	0.70	0.74	0.77	0.79	0.81	0.81	0.82	0.83
19 FLANKER 6	−0.18	−0.11	−0.22	−0.18	0.20	0.36	0.32	0.48	0.42	0.41	0.51	0.53	0.51	0.53	0.55	0.53	0.66	0.77	1	0.80	0.81	0.86	0.86	0.86	0.87	0.88
20 FLANKER 8	−0.19	−0.12	−0.22	−0.19	0.16	0.34	0.31	0.48	0.42	0.43	0.51	0.50	0.52	0.54	0.57	0.54	0.70	0.80	0.84	1	0.86	0.88	0.90	0.89	0.91	0.91
21 FLANKER 10	−0.23	−0.14	−0.26	−0.22	0.21	0.38	0.37	0.52	0.46	0.47	0.54	0.53	0.55	0.56	0.61	0.57	0.72	0.83	0.84	0.90	1	0.91	0.91	0.92	0.93	0.94
22 FLANKER 12	−0.19	−0.11	−0.25	−0.19	0.20	0.39	0.35	0.48	0.43	0.45	0.51	0.50	0.52	0.55	0.59	0.55	0.71	0.82	0.85	0.91	0.93	1	0.94	0.95	0.96	0.96
23 FLANKER 14	−0.21	−0.11	−0.23	−0.19	0.21	0.39	0.37	0.51	0.43	0.44	0.54	0.51	0.54	0.54	0.58	0.55	0.72	0.85	0.88	0.92	0.94	0.96	1	0.96	0.97	0.98
24 FLANKER 16	−0.19	−0.11	−0.25	−0.20	0.21	0.41	0.38	0.53	0.47	0.48	0.56	0.54	0.57	0.58	0.62	0.59	0.72	0.84	0.87	0.93	0.95	0.96	0.97	1	0.98	0.98
25 FLANKER 18	−0.20	−0.11	−0.26	−0.20	0.20	0.40	0.39	0.53	0.47	0.47	0.55	0.54	0.57	0.58	0.61	0.58	0.72	0.86	0.89	0.93	0.95	0.96	0.97	0.98	1	0.99
26 FLANKER 20	−0.21	−0.12	−0.26	−0.21	0.20	0.40	0.39	0.53	0.47	0.47	0.55	0.54	0.57	0.58	0.61	0.58	0.73	0.86	0.89	0.94	0.96	0.97	0.98	0.99	0.99	1

ANTI-LET antisaccade with letters, *ANTI-ARO* antisaccade with arrows, *SART d'* d' score from SART, *SART rtsd* intrasubject standard deviation in RT from SART, *DSSQ* Dundee State Stress Questionnaire – TUT subscale. *SART* Sustained Attention to Response Task. *Flanker* arrow flanker

attention control and TUT rate were a bit less stable and consistent than those between DSSQ and TUT rate. Generally, however, both stabilized (using a .02 window around the 20-bin-size estimate) for bin sizes ≥ 6–8, with some occasional bins' correlations outside that window.

In Condition 2, factor loadings for each task's TUT rate were reasonably stable when based on 8 or more probes (see Table 8). Figure 3 presents the latent variable models for Condition 2. Here, the correlations between attention control and TUT rate factors were numerically weaker, and one model (4 Probes) yielded a nonsignificant correlation, but all correlations were within a range of .08 of the 20-bin value; the relations between retrospective mind-wandering ratings on the DSSQ and TUT rates were a bit more variable, within a range of .11 of the 20-bin value. These path estimates appeared to stabilize for models based on data from 6 or more probes (for attention control) and 12 or more probes (for DSSQ). Considering factor loadings and path estimates across models for Condition 1 and Condition 2, it is more difficult to pin down a single ideal number of probes than it was for Study 1, but analyses based on 6–10 probes per task appear to efficiently provide reasonably strong evidence for TUT rate reliability and validity.

Secondary analyses of first-n probes

As with Study 1, we also examined *M* TUT rates and TUT-rate correlations across probe-bin sizes for each task in each condition, with probes drawn consecutively from the beginning of each task (see Supplemental Tables 5 and 6, respectively). TUT rates again increased with the probe bin size in both conditions, likely reflecting a time-on-task effect, but stabilized by bin sizes 6–10, depending on the task and condition. TUT-rate correlations across bins within tasks stabilized (with *r*s ≥ .90 with bin size 20 TUT rate) by bin size 8–12 (for randomly selected probes, these stabilized by bin size 8), and TUT-rate correlations across tasks stabilized in the .40 range for bin sizes 12 and 8 in Conditions 1 and 2, respectively (for randomly selected probes, these were stabilized by bin sizes 8–10). Overall, then, these findings correspond well to those from the randomly selected probes.

CFAs on the data for Conditions 1 and 2 indicated adequate fit for all models, except for RMSEA fit indices for Condition 1 (see Supplemental Table 7) and again with some overfitting in some Condition 2 models. Supplemental Figures 2 and 3 present the overall structural models with path estimates for Conditions 1 and 2, respectively, and Supplemental Table 8 presents the factor loadings for both conditions. TUT-rate factor

Table 7 Fit statistics for latent variable models from Study 2, by experimental condition

Model	χ^2	df	RMSEA [95% CI]	SRMR	CFI	TLI
Condition 1						
2 Probes	29.671	18	.049 [.010–.080]	.039	.975	.961
4 Probes	24.026	18	.035 [.000–.069]	.041	.988	.982
6 Probes	45.452	18	.076 [.049–.103]	.057	.948	.919
8 Probes	42.005	18	.071 [.043–.099]	.054	.955	.930
10 Probes	47.549	18	.079 [.052–.106]	.055	.950	.921
12 Probes	56.300	18	.089 [.064–.116]	.060	.934	.897
14 Probes	50.238	18	.082 [.056–.109]	.059	.945	.914
16 Probes	57.445	18	.091 [.065–.118]	.061	.935	.900
18 Probes	55.974	18	.089 [.063–.116]	.059	.939	.905
20 Probes	54.343	18	.087 [.061–.114]	.060	.940	.907
Condition 2						
2 Probes	9.106	18	.000 [.000–.000]	.020	1.000	1.025
4 Probes	14.914	18	.000 [.000–.044]	.029	1.000	1.008
6 Probes	23.947	18	.035 [.000–.069]	.040	.990	.984
8 Probes	21.841	18	.028 [.000–.064]	.033	.994	.990
10 Probes	22.251	18	.030 [.000–.065]	.038	.993	.989
12 Probes	21.834	18	.028 [.000–.064]	.040	.994	.990
14 Probes	20.744	18	.024 [.000–.061]	.036	.996	.993
16 Probes	20.225	18	.021 [.000–.060]	.035	.996	.995
18 Probes	18.739	18	.012 [.000–.056]	.033	.999	.998
20 Probes	19.285	18	.016 [.000–.058]	.036	.998	.997

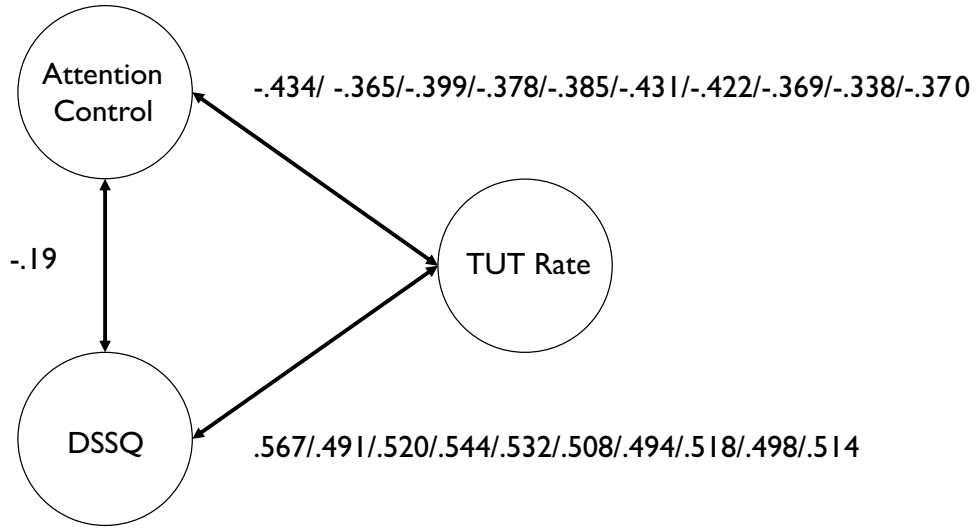

Fig. 2 Standardized path estimates from confirmatory factor analyses among attention-control ability (Attn Control), retrospective mind-wandering ratings (DSSQ), and TUT rate across different probe-number bins, from Study 2, Condition 1. Estimates from each model are separated by the slash starting with the 2 Probes model and ending with the 20 Probes model. For clarity, factor loadings are presented separately for each model in Table 8

Table 8 Standardized factor loadings (and standard errors) for latent variable models for Study 2, by condition

Construct and measure	Confirmatory factor analysis models									
	2 Probes	4 Probes	6 Probes	8 Probes	10 Probes	12 Probes	14 Probes	16 Probes	18 Probes	20 Probes
Condition 1										
Attention control										
ANTI-LET	.89 (.06)	.92 (.06)	.87 (.05)	.89 (.05)	.87 (.05)	.88 (.05)	.87 (.05)	.89 (.05)	.88 (.05)	.88 (.05)
ANTI-ARO	.68 (.05)	.66 (.05)	.69 (.05)	.68 (.05)	.69 (.05)	.69 (.05)	.69 (.05)	.68 (.05)	.68 (.05)	.68 (.05)
SART d’	.37 (.06)	.36 (.06)	.38 (.06)	.37 (.06)	.38 (.06)	.38 (.06)	.38 (.06)	.38 (.06)	.38 (.06)	.38 (.06)
SART rtsd	.44 (.06)	.42 (.06)	.45 (.06)	.44 (.06)	.45 (.06)	.45 (.06)	.45 (.06)	.45 (.06)	.45 (.06)	.45 (.06)
DSSQ										
SART	.82 (.03)	.82 (.03)	.82 (.03)	.82 (.03)	.82 (.03)	.82 (.03)	.82 (.03)	.82 (.03)	.82 (.03)	.82 (.03)
FLANKER	.86 (.03)	.86 (.03)	.86 (.03)	.86 (.03)	.86 (.03)	.86 (.03)	.86 (.03)	.86 (.03)	.86 (.03)	.86 (.03)
TUTs										
SART	.51 (.06)	.71 (.04)	.68 (.04)	.69 (.04)	.79 (.04)	.77 (.04)	.80 (.04)	.82 (.03)	.85 (.03)	.82 (.03)
FLANKER	.48 (.06)	.62 (.04)	.61 (.04)	.63 (.04)	.69 (.03)	.66 (.03)	.68 (.03)	.71 (.03)	.72 (.03)	.72 (.03)
Condition 2										
Attention control										
ANTI-LET	.90 (.04)	.89 (.04)	.90 (.04)	.90 (.04)	.89 (.04)	.89 (.04)	.89 (.04)	.90 (.04)	.89 (.04)	.89 (.04)
ANTI-ARO	.77 (.04)	.77 (.04)	.77 (.04)	.77 (.04)	.77 (.04)	.77 (.04)	.77 (.04)	.76 (.04)	.77 (.04)	.77 (.04)
SART d’	.46 (.06)	.46 (.06)	.46 (.06)	.46 (.06)	.46 (.06)	.46 (.06)	.46 (.06)	.45 (.06)	.46 (.06)	.46 (.06)
SART rtsd	.55 (.05)	.56 (.05)	.56 (.05)	.56 (.05)	.56 (.05)	.56 (.05)	.56 (.05)	.56 (.05)	.56 (.05)	.56 (.05)
DSSQ										
SART	.86 (.03)	.85 (.03)	.85 (.03)	.85 (.03)	.85 (.03)	.85 (.03)	.85 (.03)	.84 (.03)	.85 (.03)	.85 (.03)
FLANKER	.82 (.03)	.83 (.03)	.83 (.03)	.83 (.03)	.83 (.03)	.83 (.03)	.83 (.03)	.83 (.03)	.83 (.03)	.83 (.03)
TUTs										
SART	.62 (.05)	.60 (.05)	.64 (.05)	.70 (.04)	.70 (.04)	.74 (.04)	.75 (.04)	.78 (.04)	.81 (.04)	.80 (.04)
FLANKER	.60 (.05)	.55 (.05)	.56 (.04)	.63 (.04)	.61 (.04)	.65 (.04)	.65 (.04)	.66 (.04)	.70 (.04)	.68 (.04)

ANTI-LET antisaccade with letters, *ANTI-ARO* antisaccade with arrows, *SART d'* d' score from SART, *SART rtsd* intrasubject standard deviation in RT from SART, *DSSQ* Dundee State Stress Questionnaire – TUT subscale. *SART* Sustained Attention to Response Task, *Flanker* Arrow flanker

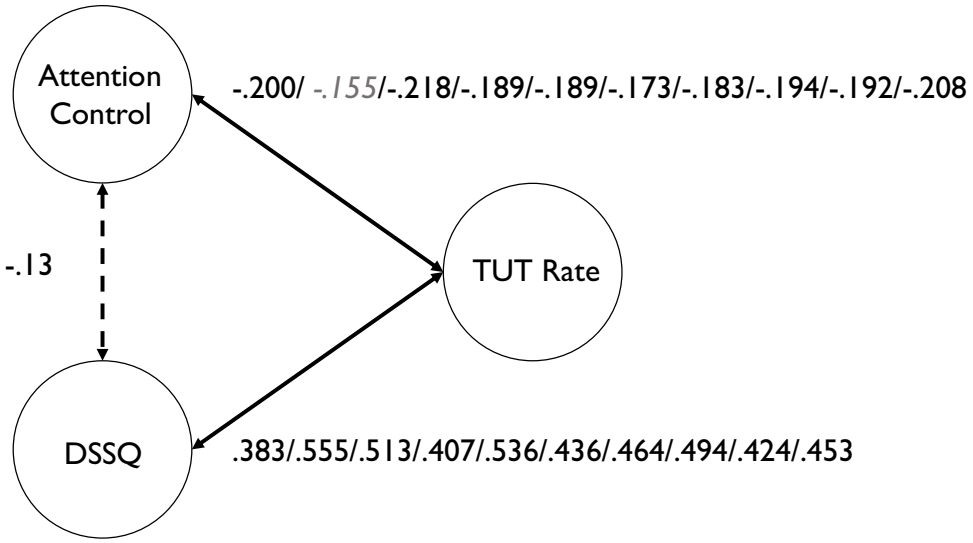

Fig. 3 Standardized path estimates from confirmatory factor analyses among attention-control ability (Attn Control), retrospective mind-wandering ratings (DSSQ), and TUT rate across different probe-number bins, from Study 2, Condition 2. Estimates from each model are separated by the slash starting with the 2 Probes model and ending with the 20 Probes model. For clarity, factor loadings are presented separately for each model in Table 8

loadings for Condition 1 exceeded .50 for models based on bin sizes ≥ 6 and exceeded .60 for models based on bin sizes ≥ 12; loadings for Condition 2 exceeded .50 for all models (even for bin size 2) and exceeded .60 for models based on bin sizes ≥ 6. Path estimates for correlations between TUT rate and the other constructs (attention control, DSSQ mind-wandering ratings) appeared to stabilize with estimates within a .02 window for bin sizes ≥ 6 for attention control and ≥ 10 for DSSQ in Condition 1, and for bin sizes ≥ 8 for attention control and ≥ 6 for DSSQ in Condition 2.

As in Study 1, then, the correlational findings here generally replicate those from randomly selected probes. TUT rates calculated from bins of 6–10 thought probes efficiently demonstrate nearly as strong reliability and validity as those calculated from bins of 20 (and even TUT rates calculated from bins as small as 2 or 4 provide reasonable reliability and validity). Again, the parallels here to those from the randomly selected probes indicate that the findings from randomly selected probes are not driven by subjects' experiences with additional, non-analyzed probes. Overall, then, a recommendation of 8 probes per task appears to fit the findings well from both Studies 1 and 2.

General discussion

Most recent mind wandering research presents thought probes within ongoing tasks and activities to measure rates of TUTs (Weinstein, 2018). Mind wandering researchers therefore face the question of how many thought probes to present. Too few probes may yield unreliable TUT-rate estimates and too many probes may provide invalid assessments if probes reactively change subjects' ongoing conscious experiences via frequent interruption and reminders of the potential for TUTs (e.g., Konishi & Smallwood, 2016; Seli et al., 2013a). Is there a "Goldilocks zone" of probe numbers that maximizes the reliability and validity of TUT measurement while minimizing the interrupting and reactive effects of probing?

In the present exploratory reanalyses, we examined how the number of thought probes analyzed from a task might elicit differences in the reliability or validity of TUT-rate individual differences. We reanalyzed two large datasets where U.S. undergraduates completed 2–5 computerized laboratory tasks with embedded thought probes (from Kane et al., 2016, 2021), and we calculated each subject's TUT rates based on 2–14 randomly selected probes in Study 1 and based on 2–20 randomly selected probes in Study 2 (all in "bin" increments of two); all tasks in Study 1 had originally presented 12–45 thought probes, and all tasks in Study 2 had had presented 20 or 45 probes.

Our reanalyses for both Study 1 and Study 2 examined bin-size changes in mean TUT rates, within-task correlations of TUT rates, between-task correlations of TUT rates, TUT-rate factor loadings in latent-variable models, and TUT-rate factor correlations with other constructs in latent variable models. Generally, the results indicated that TUT rates calculated from 8 randomly selected probes adequately captured similar patterns to those from the largest set of analyzed probes (and to the original correlations that used

all available probe information). Note, however, that in some contexts, using 10–12 probes seemed to improve measurement beyond 8 probes, and that in others, using as few as 4–6 probes would suffice. We were surprised to find that TUT rates calculated even from only *two* randomly selected probes per subject provided some meaningful individual-differences information, but we would not recommend that researchers rely on such a limited TUT-rate assessment of each subject, especially combined with modest sample sizes.

Recommendations

Pending independent replication, our provisional recommendation is that eight thought probes will typically provide efficient assessment of normal variation in TUT rate, at least in laboratory task contexts like those examined here and, perhaps, especially in studies with well-powered designs that use multiple tasks to assess TUT rates. Across the computerized attention and memory tasks we examined here, using eight probes would translate to having probes follow between 0.1% of trials (i.e., in our SARTs) to 5.5% of trials (i.e., in our Study 1 letter flanker task), or one probe every 3.2 min (in our SART) to every 1.5 min (in our Study 1 letter flanker). Studies using much longer or shorter tasks than those represented here may wish to approximate these percentages or inter-probe intervals rather than focusing on the raw number of probes.

Using an economical number of probes has multiple benefits. First, researchers may shorten (reliable) tasks used in the mind wandering literature to accommodate eight thought probes while presenting them after 1–5% of task trials. In doing so, more tasks could be used within a single study, allowing for more varied contexts for estimating TUT rate and enabling the use of latent variable models. Second, as already mentioned, using as few probes as possible minimizes task interruptions, as well as possible demand effects and reactivity to frequent probing.

Limitations and caveats

As a post hoc secondary analysis, this study could not parametrically vary the experience of different probe frequencies for subjects. We could only, instead, vary the number of randomly selected thought probes that contributed to data analyses for each subject, from a task context in which these subjects had responded to many more probes than those analyzed. It is possible, then, that tasks presenting only eight probes to every subject would not yield as reliable or valid TUT rates as indicated by our post hoc analyses here. However, confidence in our conclusions should increase based on the supplemental analyses we reported for Studies 1 and 2 that selected the first *n* probes that subjects encountered in each task (i.e., the first 2 probes, the first 4 probes, etc.). These analyses also indicated that reliable and valid TUT-rate measurement could be gained from as few as eight probes per subject (if not fewer), despite no other probes having been yet encountered by subjects. Together, our findings encourage future studies to assess whether presenting only eight (or fewer) probes per task allows for reliable and valid measurement of individual differences in TUT rate.

With that said, it is possible that our random selection of probes for each bin allowed some probes to be selected across multiple bins within a task, for at least some tasks for some subjects (e.g., a SART probe from bin 4 could have contributed to SART bin 8). This was necessarily true of our supplemental first-*n* analyses, where the TUT rate from the first 2 probes also contributed to the TUT rate for the first 4 probes, and it was increasingly true for larger bins that encroached on the maximum bin sizes (e.g., bins 12 and 14 in Study 1; bins 18 and 20 in Study 2). Such dependencies may have artificially inflated *within-task* probe-bin correlations for both the random-probe and first-*n*-probe analyses (although bin size 8 yielded similar results across studies even though maximum bin size—and thereby probe-selection overlap—varied across studies). However, they should not have affected *between-task* correlations at each bin size (and corresponding factor loadings) that contributed to our latent variable analyses and the conclusions we drew from them, namely that construct-valid TUT assessments can be consistently derived from 6–8 thought probes.

Future research should explore whether minimum probe numbers for reliable and valid TUT measurement vary across different task or activity types. Our assessments were limited to attention and memory tasks presenting simple stimuli in a discrete-trial format, and results might differ in more continuous or engaging tasks, or tasks that better mirror typical daily-life activities. For example, during tasks or activities that have more inherent variation in attentional demand, or that evolve from being more to less interesting with time (or vice versa), more probes might be necessary to faithfully capture the dynamics of mind wandering throughout the task. The laboratory tasks we analyzed here presented relatively stable demands over time, with the same stimuli throughout, which might increase the reliability and validity of TUT reports across a small number of probes.

Moreover, we assessed TUT rates from only one kind of thought probe, which asked subjects to categorize their immediately preceding thoughts into one of several content categories (e.g., worries; fantastical daydreams). Other types of probes, which may ask about the temporal orientation of thought content (i.e., future, present, versus past orientation), or the affective valence of thought content (i.e., positive, neutral, negative), might yield different results, as might probes that take a still more different approach, such as those asking subjects to report on the intentionality of their mind wandering, or on the extent to which their thoughts were

flowing freely, or on the depth of their mind wandering on a rating scale (see Kane et al., 2021).

Finally, we must emphasize that the current findings—and our recommendations for future work—may only apply to instances where researchers are interested in overall TUT rates, and not in specific forms (i.e., sub-types) of TUTs. Researchers interested in examining differences in types of TUTs (e.g., intentional vs. unintentional; past- vs. present- vs. future-oriented; or negative vs. positive vs. neutral thoughts) should aim to use more thought probes (and perhaps subjects) to ensure that there is an adequate number of responses for each thought type. We have found that subject samples often show zero-inflated distributions of specific thought-report types (e.g., externally driven distractions), which will only become more problematic as the number of probes is reduced (Welhaf et al., 2020). Thus, although our findings suggest that it may be safe to use as few as 8 probes in studies of overall TUT rates, they do not suggest similar reliability or validity for studies of TUT sub-types.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13428-021-01766-4.

Author Notes The original data collection for Study 1 was funded by award number R15MH093771 from the National Institute of Mental Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Mental Health or the National Institutes of Health.

References

Baldwin, C. L., Roberts, D. M., Barragan, D., Lee, J. D., Lerner, N., & Higgins, J. S. (2017). Detecting and Quantifying Mind Wandering during Simulated Driving. *Frontiers in Human Neuroscience*, *11*, 406. https://doi.org/10.3389/fnhum.2017.00406

Bastian, M., & Sackur, J. (2013). Mind-wandering at the tips of the fingers: Automatic parsing of subjective states based on response time variability. *Frontiers in Psychology, 4*, 573. https://doi.org/10.3389/fpsyg.2013.00573

Brosowsky, N. P., DeGutis, J., Esterman, M., Smilek, D., & Seli, P. (2020). Mind wandering, motivation, and task performance over time: Evidence that motivation insulates people from the negative effects of mind wandering. *Psychology of Consciousness: Theory, Research, and Practice*.

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, 67*(1), 19.

Forster, S., & Lavie, N. (2009). Harnessing the wandering mind: The role of perceptual load. *Cognition*, *111*(3), 345-355.

Foster, J. L., Shipstead, Z., Harrison, T. L., Hicks, K. L., Redick, T. S., & Engle, R. W. (2015). Shortened complex span tasks can reliably measure working memory capacity. *Memory & Cognition*, *43*(2), 226-236.

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

Hollis, R. B., & Was, C. A. (2016). Mind wandering, control failures, and social media distraction in online learning. *Learning and Instruction, 42*, 104-112.

Kane, M. J, Gross, G. M.., Chun, C. A., Smeekens, B. S., Meier, M. E., Silvia, P. J., & Kwapil, T. R. (2017). For whom the mind wanders, and when, varies across laboratory and daily-life settings. *Psychological Science, 28*, 1271-1289.

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*, 1017-1048.

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*, 2372–2411.

Konishi, M., & Smallwood, J. (2016). Shadowing the wandering mind: How understanding the mind-wandering state can inform our appreciation of conscious experience. *WIREs Cognitive Science, 7*, 233–246

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

Lindquist, S. I., & McLean, J. P. (2011). Daydreaming and its correlates in an education environment. *Learning and Individual Differences, 21*, 158-167.

Livesley, W. J., & Jackson, D. N. (2009). *Dimensional Assessment of Personality Pathology—Basic Questionnaire (DAPP-BQ): Technical Manual*. Sigma Assessment Systems, Inc.

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

McVay J. C., & Kane, M. J. (2012). Drifting from slow to "d'oh!": Working memory capacity and mind wandering predict extreme reaction time and executive control errors. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*, 525-549.

Meier, M. E. (2021). Testing the attention-distractibility trait. Memory & Cognition, 49(7), 1490–1504.

Miers, T. C., & Raulin, M. L. (1987). Cognitive Slippage Scale. In K. Corcoran & J. Fischer (Eds.), *Measures for clinical practice: A sourcebook* (pp. 125-127). Free Press.

Mrazek, M. D., Phillips, D. T., Franklin, M. S., Broadway, J. M., & Schooler, J. W. (2013). Young and restless: validation of the Mind-Wandering Questionnaire (MWQ) reveals disruptive impact of mind-wandering for youth. *Frontiers in psychology*, *4*, 560. https://doi.org/10.3389/fpsyg.2013.00560

Oswald, F. L., McAbee, S. T., Redick, T. S., & Hambrick, D. Z. (2015). The development of a short domain-general measure of working memory capacity. *Behavior Research Methods*, *47*(4), 1343-1355.

Raine, A., (1991). The SPQ-A scale for the assessment of schizotypal personality based on DSM-III-R criteria. *Schizophrenia Bulletin, 17*, 555-564.

Reichle, E. D., Reineberg, A. E., & Schooler, J. W. (2010). Eye movements during mindless reading. *Psychological Science*, *21*(9), 1300-1310.

Risko, E. F., Anderson, N., Sawal, A., Engelhardt, M., & Kingstone, A. (2012). Everyday attention: Variation in mind wandering and memory in a lecture. *Applied Cognitive Psychology, 26*, 234-242.

Robison, M.K., Miller, A.L. & Unsworth, N. (2019). Examining the effects of probe frequency, response options, and framing within the thought-probe method. *Behavior Research Methods, 51***,** 398–408.

Robison, M.K., Miller, A.L. & Unsworth, N. (2020). A multi-faceted approach to understanding individual difference in mind-wandering. *Cognition, 198***,** 104078.

Robison, M. K., & Unsworth, N. (2018). Cognitive and contextual correlates of spontaneous and deliberate mind-wandering. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 44,* 85 - 98.

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

Rummel, J., & Boywitt, C. D. (2014). Controlling the stream of thought: Working memory capacity predicts adjustment of mind-wandering to situational demands. *Psychonomic Bulletin & Review*, *21*, 1309-1315.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online, 8*, 23-74.

Schooler, J. W., Reichle, E. D., & Halpern, D. V. (2004). Zoning out while reading: Evidence for dissociations between experience and metaconsciousness. In D. Levin (Ed.), *Thinking and seeing: Visual metacognition in adults and children* (pp. 203-226). MIT Press.

Schubert, A. L., Frischkorn, G. T., & Rummel, J. (2019). The validity of the online thought-probing procedure of mind wandering is not threatened by variations of probe rate and probe framing. *Psychological Research, 84,* 1846-1856.

Seli, P., Carriere, J. S., Levene, M., & Smilek, D. (2013a). How few and far between? Examining the effects of probe rate on self-reported mind wandering. *Frontiers in Psychology*, *4*, 430.

Seli, P., Cheyne, J. A., & Smilek, D. (2013b). Wandering minds and wavering rhythms: Linking mind wandering and behavioral variability. *Journal of Experimental Psychology: Human Perception and Performance, 39*, 1-15.

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

Seli, P., Smallwood, J., Cheyne, J. A., & Smilek, D. (2015b). On the relation of mind wandering and ADHD symptomatology. *Psychonomic Bulletin & Review*, *22*(3), 629-636.

Seli, P., Risko, E. F., & Smilek, D. (2016). Assessing the associations among trait and state levels of deliberate and spontaneous mind wandering. Consciousness and Cognition, 41, 50-56.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). A 21 word solution. Dialogue: *The Official Newsletter of the Society for Personality and Social Psychology, 26*, 4–7.

Smallwood, J., McSpadden, M., & Schooler, J. W. (2008). When attention matters: The curious incident of the wandering mind. *Memory & Cognition, 36*, 1144-1150.

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*, 487-518.

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*, 389-415.

Szpunar, K. K., Khan, N. Y., & Schacter, D. L. (2013). Interpolated memory tests reduce mind wandering and improve learning of online lectures. *Proceedings of the National Academy of Sciences, USA of the United States of America, 110,* 6313–6317.

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

Unsworth, N., Redick, T.S., Lakey, C.E., Young, D.L. (2010). Lapses in sustained attention and their relation to executive and fluid abilities: An individual differences investigation. *Intelligence*, 38, 111-122.

Unsworth, N., & Robison, M. K. (2016). The influence of lapses of attention on working memory capacity. *Memory & Cognition*, *44*(2), 188-196.

Unsworth, N., & Robison, M. K. (2017). The importance of arousal for variation in working memory capacity and attention control: A latent variable pupillometry study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(12), 1962.

Unsworth, N., Robison, M. K., & Miller, A. L. (2021). Individual differences in lapses of attention: A latent variable analysis. Journal of Experimental Psychology: General, 150(7), 1303–1331.

Weinstein, Y. (2018). Mind-wandering, how do I measure thee with probes? Let me count the ways. *Behavior Research Methods, 50*, 642–661.

Welhaf, M.S., Smeekens, B.A., Gazzia, N.C., Perkins, J.B., Silvia, P.J., Meier, M.E., Kwapil, T.R., & Kane, M.J. (2020). An exploratory analysis of individual differences in mind wandering content and consistency. *Psychology of Consciousness: Theory, Research, and Practice, 7,* 103-125.

Zhang, Y., & Kumada, T. (2018). Automatic detection of mind wandering in a simulated driving task with behavioral measures. *PLoS ONE, 13*(11), Article e0207092.

Zhang, H., Miller, K. F., Sun, X., & Cortina, K. S. (2020). Wandering eyes: Eye movements during mind wandering in video lectures. *Applied Cognitive Psychology*, *34*(2), 449-464.