Validity of the think-aloud procedure in comparison to other methods for studying the phenomenological features and memory of spontaneous thought

Arya Gilles1*, Gaëlle Panneels1, Arnaud D'Argembeau1,2, and David Stawarczyk1,2*

1 Psychology and Neuroscience of Cognition Research Unit, University of Liège, Place des Orateurs 1 (B33), 4000 Liège, Belgium.

2 GIGA – CRC (Cyclotron Research Center) Human Imaging, University of Liège, Allée du 6 Août 8 (B30) , 4000 Liège, Belgium.

* Corresponding authors at: Psychology and Neuroscience of Cognition Research Unit, University of Liège, Place des Orateurs 1 (B33), 4000 Liège, Belgium.

E-mail addresses: arya.gilles@uliege.be (A. Gilles), d.stawarczyk@uliege.be (D. Stawarczyk).

Abstract

A hallmark of the human mind is its tendency to generate spontaneous thoughts, whether during tasks or in idle moments. This phenomenon is typically studied in the laboratory using the Thought-Probe Protocol (TPP), in which participants report the content of their thoughts when prompted at various intervals. Although well validated, the TPP nonetheless suffers from several limitations, such as its inability to track the flow of thoughts between probes. To address these issues, researchers have recently revisited the Think-Aloud Protocol (TAP), which involves the continuous verbalization of spontaneous thoughts. While the TAP offers access to the ongoing flow of thoughts, its validity relative to other methods has not yet been fully established. In this study, we compared four methods for assessing spontaneous thoughts: the TAP, TPP, Daily Life Experience Sampling Protocol (DLESP), and retrospective thought listing. We focused on the phenomenological characteristics of thoughts and features that predicted their recall after a one- day delay. Our results revealed minimal differences between the TAP and TPP in terms of thought characteristics and memory predictors. However, thoughts reported with these two methods differed from those assessed more ecologically with the DLESP, and certain thought features were overrepresented in retrospective thought listing. Overall, our findings suggest that the TAP is as valid as the TPP for investigating spontaneous thought, although thought characteristics may differ between laboratory and real-world settings. They also suggest that concurrent reporting methods, such as the TAP and TPP, provide a more representative view of spontaneous thought features than retrospective assessments.

Keywords

Spontaneous thought; Think-aloud protocol; Experience sampling; Mind-wandering; Memory

1. Introduction

Even when our attention is not engaged in a particular task, our mind is far from idle. It continuously generates thoughts that move from one topic to another in a constant flow, a phenomenon that William James (1890) described in his seminal work as the 'stream of thought.' For example, while sitting in a waiting room before an appointment, our thoughts may move from reflecting on our surroundings to recalling a pleasant memory, then pondering a complex personal problem, imagining how the appointment might unfold, daydreaming about a pleasant imaginary scenario, and so on. The units that compose this continuous flow are referred to as *spontaneous thoughts,* a broad term that encompasses concepts such as mind-wandering, creative thinking, and dreaming (Christoff et al., 2016). Among these, mind-wandering can be defined as a shift of attention from the immediate task at hand to thoughts whose content is decoupled from the here and now, such as thinking about a future meeting or reminiscing about a past event while listening to a lecture (Smallwood & Schooler, 2015; Stawarczyk et al., 2011). Mind-wandering may represent, on average, up to 30 % of daily thoughts (Kawashima et al., 2023). Although generally associated with decreased performance on the task at hand, mind-wandering may possess various benefits (Mooneyham & Schooler, 2013; Randall et al., 2014). For example, a notable finding is that mind-wandering shows a prospective bias, with a substantial portion of thoughts consisting in future planning, which may help guide upcoming decisions and actions (Baird et al., 2011; Stawarczyk et al., 2011, 2013).

Currently, researchers rely on a variety of concurrent and retrospective self-reporting methods to assess spontaneous thought content and occurrence. Thus far, the predominant concurrent method is the experience sampling technique, which can be divided into two sub-methods: the probe-caught method and the self-caught method (Smallwood & Schooler, 2006, 2015). In the probe-caught method, participants are interrupted at various intervals during task performance or at rest and asked about the content of their mental experience, with questions such as "Where was your attention focused just before the probe?" (Weinstein, 2018). A variation of this method involves probing participants using a smartphone app during everyday life to capture more ecological occurrences of spontaneous thoughts (Kawashima et al., 2023). On the other hand, the self-caught method requires participants to identify, without being probed, when a spontaneous thought occurs and to report it to the experimenter (Chu et al., 2023). Besides experience sampling, more retrospective methods have also been used, in which participants are asked to report their thoughts after completing a task or at the end of a specified time period. For instance, besides self-report scales and questionnaires (e.g., Carriere et al., 2013; Matthews, 2021; Stawarczyk et al., 2012; Andrews-Hanna et al., 2013, 2022) recently proposed an Autobiographical Thought Sampling Task in which participants have to list spontaneous thoughts they experienced in the past seven days of their daily lives.

These different methods of assessing spontaneous thoughts each have their pros and cons. Important limitations of the self-caught method include the need for participants to first become aware of their spontaneous thoughts (e.g., by explicitly noticing that they were mind-wandering) and then to remember to report these thoughts (Chu et al., 2023). Retrospective methods rely entirely on participants' memory to provide accurate and exhaustive accounts of their spontaneous thoughts—an assumption that has been scarcely tested (but see Brewer, 1988; Stawarczyk & D'Argembeau, 2019). Impacted to a much lower extent by these issues, the probe-caught method is currently considered the standard method for studying spontaneous thoughts (Kane et al., 2021; Weinstein, 2018). Findings from behavioral, physiological, lesion, and neuroimaging studies support the idea that participants' reports to the probes accurately reflect the underlying cognitive processes giving rise to the reported thoughts or experiences (Kam et al., 2022; Smallwood & Schooler, 2006, 2015). However, although the probe-caught method is now widely accepted, it nonetheless suffers from methodological limitations of its own. First, because participants only report their thoughts when probed, a limited sample of mental experiences occurring during a given period are assessed; the thoughts that occur between probes are not sampled. Second, by probing the flow of thoughts at single time points, the probe-caught method cannot easily be used to assess how the content of spontaneous thought changes over time, which is a key aspect of recent models such as the Dynamic Framework of Thought (Christoff et al., 2016). According to this model, excessive or insufficient dynamism in spontaneous thinking is a key factor determining whether such thoughts are maladaptive. Excessive dynamism may be related to disorders characterized by disorganized thought patterns, such as ADHD or schizophrenia, whereas insufficient dynamism may be related to depressive or anxiety disorders through increased rumination and other forms of negative repetitive thinking (see also Watkins, 2008).

To overcome the limitations of the probe-caught method, researchers have begun to use the *Think-Aloud Protocol* (TAP) to assess spontaneous thoughts. Originally introduced by Watson (1920), the TAP requires participants to verbalize their thoughts continuously over a specified period of time or task. Initially used to study the cognitive processes underlying complex problem solving, this method has been successfully applied in many disciplines over the last decades (e.g., education, Kesler et al., 2016; text comprehension, Wang, 2016; sport, Samson et al., 2017). The TAP offers several potential advantages. First, by requiring participants to verbalize all ongoing mental processes, it theoretically provides access to the entirety of the experiential content over the defined time period. In doing so, the TAP is particularly suited for studying the dynamics of spontaneous thought, as it allows following the stream of thoughts directly as it occurs without interrupting it. Second, it can capture numerous consecutive thoughts within a brief timeframe, thus reducing the time required to gather mental experiences. Finally, commonly used methods to investigate spontaneous thoughts may be susceptible to reporting bias due to the delay between thought occurrences and their subsequent report (Mathews et al., 2024; Smallwood & Schooler, 2006). As the TAP requires participants to directly voice the content of their thoughts aloud, this method may provide a more immediate and accurate insight into participants' mental experiences by minimizing retrospective memory bias.

Cognitive researchers have recently begun using the TAP to explore the dynamic nature, content, and correlates of spontaneous thoughts. Notably, Sripada and Taxali (2020) examined the structure of spontaneous thoughts, revealing that the flow of thought might follow a "clump-and-jump" structure characterized by clusters of semantically related thoughts (clumps) interspersed with abrupt transitions (jumps) to a new topic. In line with the Dynamic Framework of Thought (Christoff et al., 2016), Raffaelli and colleagues (2021) found that trait brooding, a form of

maladaptive rumination, was associated with longer negative and shorter positive spontaneous thoughts, a tendency to move away from positive topics, and a decreased variability in the content of negative thoughts (for similar TAP findings regarding rumination and spontaneous thoughts in healthy and clinically depressed individuals, see Li et al., 2022, 2024). Conversely, participants with higher levels of hyperactive symptoms on an ADHD self-rating scale showed higher variability in thought content (Raffaelli et al., 2025). Raffaelli and colleagues (2024) also reported that people with higher originality scores on a divergent thinking task experienced more freely moving spontaneous thoughts and looser associative transitions between thoughts. Finally, Li and colleagues (2023) used the TAP during functional magnetic resonance imaging scanning and found that variability in thought content was linked to changes in neural activity across a wide variety of cortical regions.

While the TAP conceptually offers numerous advantages over commonly used methods for assessing spontaneous thought, several questions about the validity of this method remain (Fox et al., 2011; Jordano & Touron, 2018; Yang & Zhang, 2023). First, since the TAP involves continuous verbalization, one might wonder whether this process interferes with the natural flow of thought. In other words, is there a risk of altering the cognitive processes that these verbal reports are intended to capture, thereby introducing a *reactivity* issue. A second concern is whether the verbal reports made during the TAP fully and accurately reflect the cognitive processes experienced by the participant, raising the topic of *veridicality*. In other words, the TAP would be considered unreactive if no changes occur in the flow of thoughts due to verbalization, and veridical if it accurately reflects the content of thoughts. Despite theoretical concerns about reactivity and veridicality, empirical findings from domains other than spontaneous thought seem to indicate that the TAP is a valid method. For example, in their *meta*-analysis of 92 studies, Fox and colleagues (2011) showed that, even if the use of concurrent verbal reports increases the time necessary to achieve a task, it does not affect participants' performance compared to silent conditions, provided that participants simply report their thoughts as they occur and do not have to explain the reasoning behind them.

Although no study has directly examined the validity of the TAP for investigating spontaneous thoughts in comparison to other methods, recent findings offer preliminary insight into this question. Sripada and Taxali (2020) found significantly more frequent topic shifts during rest while participants were silently thinking than during the TAP. However, subsequent studies did not observe such a difference (Garg et al., 2025; Li et al., 2022). Using a retrospective self-report questionnaire to assess the overall features of thoughts at the end of a resting period, Li and colleagues (2022) did not find any difference between thinking aloud and silent thinking on various content dimensions, including temporal orientation, affective valence, representational format, or thinking about oneself and other people. They also replicated the prospective bias of spontaneous thoughts with participants reporting more future than past thoughts, but only for the TAP. Using a similar method, Garg et al. (2025) found that the TAP differed from silent thinking on only 7 of 36 items assessing thought features, topics, and task experience. Participants retrospectively reported that the TAP was overall more difficult, distressing, and involved more mind-blanking than the silent thinking condition. Thoughts in the TAP were also rated as more repetitive, controlled, private, and less likely to involve the "partner, intimacy, love, and sexual matters" topic category. Finally, participants reported that, although thinking aloud did not feel natural, they were rarely unable to vocalize their thoughts and only slightly filtered in their verbalization, suggesting low reactivity issues. Raffaelli and colleagues (2021, 2024) also asked their participants to complete self-report scales after the TAP to assess their experience during the task. These ratings revealed that participants reported only slight self-censorship while verbalizing their thoughts, and that these thoughts were fairly similar to those of their daily life. These authors also assessed the temporal orientation of spontaneous thoughts and found evidence of a prospective bias when focusing on thoughts unrelated to the present environment (Raffaelli et al., 2021, Study 2).

In summary, previous findings provide preliminary evidence supporting the validity of the TAP as a method for assessing spontaneous thoughts, although some differences compared to silent thinking have also been observed. However, whether thoughts reported using the TAP differ from those reported with other methods—including the more widely used and validated probe-caught method—has not yet been directly investigated.

1.1. Present study

The main aim of the present study was to compare the phenomenological features of spontaneous thoughts reported using the TAP to those reported in response to thought probes under similar laboratory conditions. To include a broader and more ecological range of methods, we also asked participants to complete the Autobiographical Thought Sampling Task, a retrospective thought listing procedure (Andrews-Hanna et al., 2013, 2022) and to rate the features of their spontaneous thoughts in daily life using probes sent through a smartphone app (Kawashima et al., 2023). Although it is generally assumed in the literature that all these methods similarly investigate spontaneous thought, to the best of our knowledge, no study has directly compared the features of thoughts reported during resting periods in the laboratory using the TAP and TPP with those experienced in daily life. In addition, we also examined whether we could replicate the prospective bias of spontaneous thoughts and mind-wandering (Baird et al., 2011; Baumeister et al., 2020; Stawarczyk et al., 2011, 2013), and whether participants rated their subjective experience as being similar across all four methods (Raffaelli et al., 2021, 2024). These comparisons between methods were exploratory, and we did not formulate specific a priori hypotheses about which thought characteristics would differ.

As mentioned above, no study has formally assessed whether retrospective thought reports are subject to a memory bias (e.g., whether thoughts with specific features are more easily recalled, leading to their overrepresentation in retrospective reports). To address this question as a secondary aim of the present study, we asked participants to perform a surprise free recall task for thoughts previously reported during the TAP and in response to thought probes. This allowed us to examine whether specific phenomenological features can predict which thoughts are later remembered, and whether these features differ between methods. As with the comparison of thought features, we did not formulate any specific a priori hypotheses regarding whether some features would differentially predict recall across methods. If the TAP was associated with different patterns of thought retention compared to thoughts provided in response to probes, it could reflect differences in the types of thoughts elicited or

in how those thoughts are encoded and retrieved. This would in turn question the validity of the TAP for assessing memory for thoughts relative to other methods. Conversely, if the same phenomenological features predict whether thoughts are remembered or forgotten in the TAP and in response to thought probes, this would further support the use of the TAP as a suitable method for assessing spontaneous thoughts. Preliminary studies have revealed that features such as pleasantness, personal significance, repetition, unusualness, attempts at suppression, and the involvement of a planning function can predict how easily a thought can be remembered (Brewer, 1988; Stawarczyk & D'Argembeau, 2019). In the present study, we examined whether we could replicate these findings across methods.

To address these questions, we implemented a 6-day experimental design (see Fig. 1) combining four methods for collecting mental experiences: (1) the Think-Aloud Protocol (TAP), (2) a laboratory-based thought-probe method (Thought-Probe Protocol; TPP), (3) a daily life thought-probe method via a smartphone app (Daily Life Experience Sampling Protocol; DLESP), and (4) the Autobiographical Thought Sampling Protocol (ATSP; Andrews-Hanna et al., 2013, 2022). On the first day, participants completed the TPP and the TAP in a counterbalanced order, followed by a rating of their experiences using a Thought Characteristics Questionnaire (TCQ; Stawarczyk et al., 2011, 2013). On the second day, participants freely recalled the experiences rated on Day 1, completed the ATSP, and received instructions for the DLESP. Over the next three days, participants reported and rated their daily life mental experiences when probed by smartphone notifications. Finally, on the sixth day, participants freely recalled their DLESP experiences. This design allowed us to systematically compare the features of spontaneous thoughts across methods and settings.

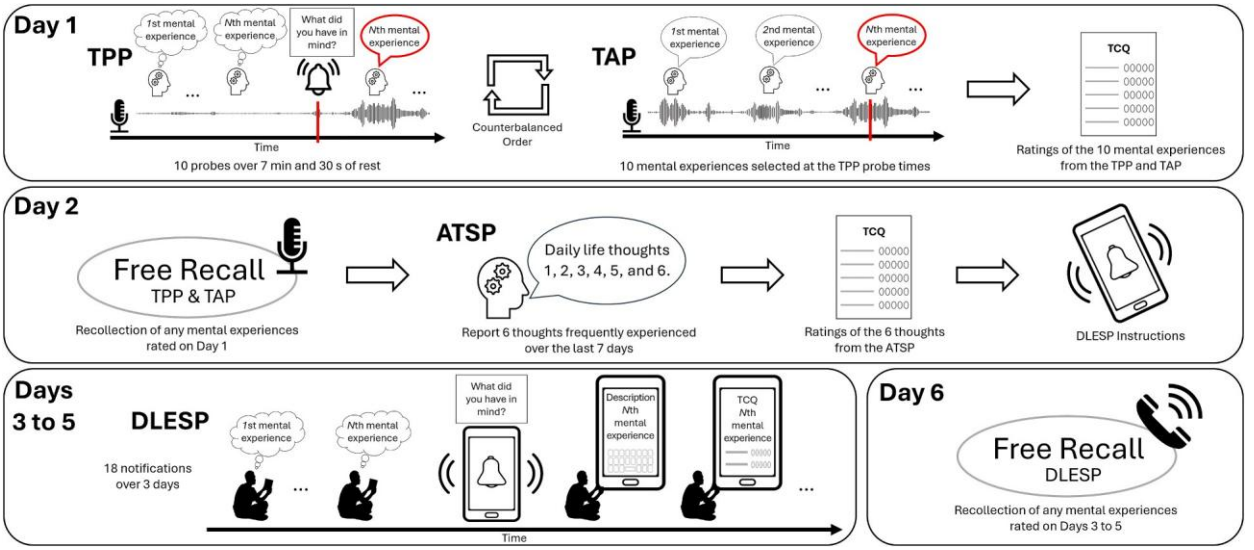

Fig. 1. Timeline of the experimental procedure. For the first session on Day 1, participants completed the Thought-Probe Protocol (TPP) and the Think-Aloud Protocol (TAP) in a counterbalanced order, followed by the rating of their mental experiences on a Thought Characteristics Questionnaire (TCQ). For the second session on Day 2, participants first recalled the mental experiences they had rated in the first session, completed the Autobiographical Thought Sampling Protocol (ATSP), and received instructions for the Daily Life Experience Sampling Protocol (DLESP). From Days 3 to 5, participants reported and rated their daily mental experiences via the smartphone app. On the final day, they recalled their DLESP experiences during a phone call.

2. Method

2.1. Transparency and Openness

We report how we determined our sample size and describe all data exclusions, manipulations, and all measures used in the study. The link to the anonymized data and analysis scripts appears in the Author Note. We did not preregister this study.

2.2. Participants

Forty-three young adults from the French-speaking community of Belgium volunteered to participate in the study and were recruited through word-of-mouth by the experimenters. None of them reported a history of neurological or psychiatric disorders or the use of medications that could affect their cognitive functioning. One participant voluntarily withdrew from the study due to difficulties in performing the TAP. The final sample included 42 participants (32 women) aged 18 to 27 years old (M_{age} = 22.3, *SD* = 2.3 years) with a mean of 14.61 (*SD* = 1.77) completed years of education; 34 participants were still students at the time data collection. Participants received no compensation for their participation. Written informed consent was obtained from all participants, and all procedures were approved by the Ethical Committee of the Faculty of Psychology, Speech Therapy, and Educational Sciences of the University of Liège. Data were collected between March 13 and May 10, 2023. We determined our sample size a priori based on a power analysis using G*Power 3.1.9.7 (Faul et al., 2007). A minimum of 34 participants was required to achieve 80 % power (alpha = 0.05) to detect a medium within-participant effect size (*d* = 0.5). Note that this sample size estimate should be considered an approximation because we analyzed the data using mixed effects models.

2.3. Materials

2.3.1. Thought-probe protocol (TPP)

Participants performed an adapted version of the TPP during a 7.5-minute rest period in the laboratory, using the online experiment builder Gorilla (www.gorilla.sc) while being recorded with a microphone. A white fixation cross was displayed on a black background in the center of the screen and participants were instructed to remain silent until prompted by sound probes. Upon hearing the sound, they were instructed to verbalize their most recent mental experience prior to the probe. During verbalization, the text 'What did you have in mind?' replaced the fixation cross. As with the TAP (see below), participants were instructed to say what they had in mind without trying to analyze or explain it, and without worrying about grammar or forming complete sentences (Raffaelli et al., 2021). If participants did not want to disclose the content of an experience because they judged it too personal, they could say 'private' without giving a detailed description. Participants had no time limit for describing their mental experiences to the probes and clicked a red 'Next' button at the bottom of the screen to resume the fixation cross after each description. The probe sound occurred 10 times at predetermined semi-random intervals (35, 40, 45, 50, or 55 s), with two occurrences of each duration. The order of these intervals was determined using a Latin square design, such that each participant was randomly assigned to one of five predetermined orders for the 10 intervals. Prior to the actual TPP, participants completed a short training session with two probes (one every 40 s), which they could repeat if necessary. The experimenter remained in the room but out of sight of the participant during training to provide direct feedback after the two probes. After the training, the experimenter left the room, leaving the participant alone during the probed period. Participants were assured that the room was soundproof and that they could not be overheard by people outside.

2.3.2. Think-aloud protocol (TAP)

The setup, instructions, and overall procedure of the TAP were identical to those of the TPP, except that no probes were presented. Instead, participants had to continuously verbalize whatever came to their mind while being recorded. If the same mental experience came to their mind multiple times during the task, they were instructed to simply repeat it. The TAP lasted 7 min and 50 s-20 s longer than the TPP—to allow participants to finish verbalizing the experience they had at the 7-minute and 30-second mark for subsequent ratings on the TCQ (see below). A sound indicated the beginning and end of the think-aloud period. Participants completed a 2-minute training session that they could repeat before performing the actual TAP.

2.3.3. Autobiographical thought sampling protocol (ATSP)

We adapted the ATSP developed by Andrews-Hanna and colleagues (2013, 2022) to retrospectively assess daily life spontaneous thoughts. The ATSP was similar to the TAP and TPP, except that participants were asked to verbally describe six thoughts that they had frequently experienced in daily life over the last seven days. Participants generated their thoughts aloud while being recorded, one at a time, initiating each description by clicking a red 'Next' button at the bottom of the screen, which triggered a sound. Participants had no time limit for describing their thoughts, and the number of each thought ('First thought', 'Second thought', etc.) appeared at the top of the screen throughout the task.

2.3.4. Daily life experience sampling protocol (DLESP)

The experience sampling method was used to assess participants' mental experiences in daily life using the smartphone app m-Path (https://m-path.io). Eighteen notifications were sent over a 3-day period, with six notifications per day. Notifications were semi-randomly scheduled between 9:30 AM and 9:30 PM, in two-hour slots. Ten minutes were removed from the beginning and end of each slot (except for the first and last slots of the day) to prevent the occurrence of overlapping notifications. Participants had five minutes to respond to the notifications before a reminder was sent, and an additional five minutes to begin completing the questionnaire after the reminder. The questionnaire became unavailable after these ten minutes. For each notification, participants were instructed to write a brief description of the last mental experience they had just prior to receiving the notification, and then to rate the features of that experience on the Thought Characteristics Questionnaire (TCQ; see below). On average, participants responded to 13.05 out of the 18 notifications (*SD* = 2.9, range: 6–18), corresponding to a compliance rate of 72.5 %, 95 % CI [67.44, 77.50]. This rate is in the range of a recent *meta*-analysis of 25 studies that used experience sampling in daily life to assess mind-wandering and reported an average compliance rate of 76.73 %, 95 % CI [69.83, 83.62] (Kawashima et al., 2023). The distribution of notification responses across days and two-hour slots is reported in the Supplementary materials.

2.3.5. Thought characteristics questionnaire (TCQ)

For each of the four methods, participants rated their mental experiences on a Thought Characteristics Questionnaire (TCQ; Stawarczyk et al., 2011, 2013). Participants first identified the type of experience they reported by answering the following item: 'This experience corresponds to: (a) something I directly perceived around me, (b) something I directly felt, or (c) a thought (anything not directly perceived or felt)?' Option (a) was defined as perceiving an external stimulus through one or more senses without any additional thought (e.g., 'I look at the fixation cross'). Option (b) referred to an internal feeling or a bodily sensation without any additional thought (e.g., 'I feel hungry'). Option (c) defined any mental experience that was not solely something directly felt or perceived. If participants categorized their mental experience as a thought, they were then asked to rate it on several phenomenological dimensions (see Table 1). All methods used the same questions except three that were adapted for the DLESP. Prior to the first rating, the experimenter reviewed all the questions with the participant to ensure that each was understood correctly.

2.3.6. Free recall task

Participants completed surprise free recall tasks on the second and on the final day of the experiment (see Fig. 1). For the recall performed on the second day, participants were instructed to verbally recollect all the mental experiences they had rated the day before following the TAP and TPP. The room setup mirrored that of the previous day. Participants sat in front of a black screen with a white fixation cross and a microphone recorded their recall. No time constraint or specific recall order were imposed during the task. The recall of the experiences reported during DLESP took place on the last day of the experiment. Participants received a phone call from the experimenter during which they were asked to verbally recollect all mental experiences that they had reported to the notifications during the experience sampling period. Instructions were identical to those of the first recall task; no time limit or specific order was imposed during the recall process. Two independent raters (the first and the second author) assessed recall accuracy for each reported thought in the TAP, TPP, and DLESP. Correct recalls were defined as instances where a reported thought specifically matched one of the recall responses. The raters agreed on 979 out of 1,045 thoughts, resulting in a Cohen's $\kappa = 0.87$, 95 % CI [0.84, 0.90], indicating very good agreement (Landis & Koch, 1977). In cases of disagreement, the raters discussed the discrepancies until a consensus was reached, and these final ratings were used in all subsequent analyses.

Table 1
TCQ items.

Content rating	Question	Scale
Visual format	This thought was in the form of visual images (such as people, objects, etc.).	1 to 7, from 'Not at all' to 'Totally'
Inner Speech format	This thought was in the form of inner speech (that is, as if I were mentally speaking).	1 to 7, from 'Not at all' to 'Totally'
Task-relatedness	This thought is about the task I was carrying out (that is, verbalizing my thoughts). *This thought is about the task I was carrying out.	1 to 7, from 'Not at all' to 'Totally'
Stimulus-dependence	This thought is about something I perceived or felt during the task. *This thought is about something I perceived or felt directly.	1 to 7, from 'Not at all' to 'Totally'
Temporality**	This thought refers to something happening in the... (several picks are allowed)	(1) Past, (2) Present, (3) Future, (4) No precise temporal orientation
Concrete	The content of this thought is about something concrete and well-defined (e.g., a specific situation or a particular action).	1 to 7, from 'Not at all' to 'Totally'
Affective Valence	The affective content of this thought is ...	-3 to +3, from 'Very negative' to 'Very positive'
Self-relatedness	This thought is about me.	1 to 7, from 'Not at all' to 'Totally'
Other-relatedness	This thought is about other people than me.	1 to 7, from 'Not at all' to 'Totally'
RepetXP	This thought came back to me at various points during the experiment. *This thought came back to me at various times since I started receiving the notifications.	1 to 7, from 'Not at all' to 'Totally'
RepetDL	This thought often comes to me in everyday life.	1 to 7, from 'Not at all' to 'Totally'
Importance	This thought is about something important to me.	1 to 7, from 'Not at all' to 'Totally'
Goal-relatedness	The content of this thought is related to my personal goals.	1 to 7, from 'Not at all' to 'Totally'
Mundane	This thought is about something mundane.	1 to 7, from 'Not at all' to 'Totally'
Structure	This thought belongs to a structured sequence of thoughts (such as in reasoning, reflection or argumentation).	1 to 7, from 'Not at all' to 'Totally'
Deliberateness	I deliberately decided to think about this topic.	1 to 7, from 'Not at all' to 'Totally'
Suppression	I've tried to remove this thought from my mind.	1 to 7, from 'Not at all' to 'Totally'
Similarity	This thought is similar to those I experience in daily life.	1 to 7, from 'Not at all' to 'Totally'
Function	The main function or utility of this thought is... (several picks are possible)	(1) to make a decision, (2) to solve a problem, (3) to plan something, (4) to reappraise a situation, (5) to keep me alert and awake, (6) to make me feel better, to relax, or to entertain myself, (7) other, (8) no apparent function.

Note. *Questions used in the Daily Life Experience Sampling Protocol; **For past and future thoughts, participants were additionally asked to indicate in a text box approximately how far in time the content of their thoughts referred to. These data are not reported in the present paper. RepetXP: Repetition in the experiment; RepetDL: Repetition in daily life.

2.3.7. Retrospective self-ratings of subjective experimental and recall experience

Following the TAP, TPP, and ATSP, participants answered two Likert scale questions regarding similarity and self-censorship: 'The experiences I have reported are similar to those of daily life' and 'To what extent did you censor or restrain yourself during the task?' The scales ranged from 1 ('Not at all') to 7 ('Totally'). At the end of the first and the second session, participants were also asked whether they had experienced any difficulties in rating the features of their thoughts: 'Did you encounter difficulties in assessing the characteristics of your thoughts?' on a 7-point Likert scale ranging from 1 ('Not at all') to 7 ('Totally'), and were asked to rate the overall quality of their retrospective evaluations: 'How would you assess the overall quality of your ratings?' using a 7-point Likert scale ranging from - 3 ('Very poor') to + 3 ('Very good').

Upon completing the experiment, participants were asked to retrospectively rate on a scale ranging from 1 ('Not at all') to 7 ('Totally') the extent to which they had guessed earlier in the study that they would be asked to recall their mental experiences. Participants provided two such retrospective ratings at the end of the experiment: once for the end of the first day and once for the end of the experience sampling period. Participants who gave ratings higher than one were additionally asked to indicate whether they had used any strategies to help them memorize their thoughts.

2.4. Experimental procedure

Data were collected in three sessions spread over six days (see Fig. 1). On the first day, participants gave their informed consent and completed the demographic questionnaire before completing the TAP and TPP, the order of which was counterbalanced across participants. Participants sat alone in front of a computer in a quiet, well-lit room with minimal stimulation: no shelves, no items on the wall, limited foot traffic near the room, and basic furnishings including a table, a chair, and two laptops. The primary laptop presented the Gorilla experiment to the participant and the second laptop recorded the participants via a microphone using Audacity (https:// www.audacityteam.org/). Once the TAP and TPP were finished, participants rated their mental experiences on the TCQ while listening to their recording. For the TPP, participants rated the 10 mental experiences they reported in response to the probes. For the TAP, they also rated 10 mental experiences, which were selected to match the timing of their occurrence with the timing of the thought-probes in the TPP. For instance, if the probes occurred at 35 s, 1 min 30 s, 2 min 15 s, etc., participants rated the experiences they had at those times during the TAP. This subsampling procedure allowed us to compare the same number of experiences separated by similar time intervals between the two methods—the only difference being whether the participants were silent or continuously verbalizing their experiences. No mention was made of the recall task during this session.

To ensure that the TAP and TPP were matched as closely as possible, we used the same procedure in both methods to determine the onset of the reported mental experiences that participants rated on the TCQ. Specifically, participants determined on their own the onset of the last experience they had in mind preceding each thought probe in the TPP, which is the usual procedure with this method (Smallwood & Schooler, 2015). We adopted a similar procedure in the TAP, where participants also identified on their own the onset of the last experience they had in mind before each of the time points corresponding to the occurrence of thought probes in the TPP. For each of these moments, the experimenter played back the audio recording starting a few seconds before the corresponding probe time to let the participant identify the mental experience to be rated. Participants then went back in the recording to the point they judged to be the onset of that mental experience. As with the TPP, this allowed participants to listen to the entirety of their verbal description before rating the corresponding experience on the TCQ.

The second laboratory session took place the following day in the same testing room as the first session. Participants were first asked to freely recall the experiences they had rated in the first session. They then completed the ATSP and rated their six reported thoughts on the TCQ while listening to their recording. After these ratings, participants installed m-Path on their phones and were given instructions regarding the DLESP. A training notification was sent to the participants, and they filled in the TCQ on their phone with the assistance of the experimenter to ensure that everything was clear about the procedure. The daily life experience sampling period lasted for the next three days. No mention was made of the final recall task during this session.

On the sixth and final day, the experimenter called the participants on the phone for the third session and asked them to freely recall the experiences they had reported in response to the notifications. After this recall task, participants were given a short debriefing on the purpose of the study and answered the retrospective questions regarding whether they had guessed the purpose of the experiment at the end of the first and second days.

2.5. Statistical analyses

All analyses were conducted in R version 4.3.2 (R Core Team, 2023). Unless stated otherwise, statistical analyses consisted of mixed models to account for the nested structure of the data: each participant reported several mental experiences in each of the four methods. The variables used as fixed effects are detailed in each section below. We used cumulative link mixed models (CLMMs), computed with the clmm function from the ordinal package (Christensen, 2023), for ordinal outcome variables (i.e., ratings made by participants the Likert scale items). Significance tests for these models were performed using Type II Wald Chi-square tests with the Anova.clmm function from the RVAideMemoire package (Hervé, 2023). For binary outcome variables (i.e., correct vs. incorrect recall, as well as the temporality and function items of the TCQ), we used generalized linear mixed models (GLMMs), computed with the glmer function from the lme4 package (Bates et al., 2015). When the outcome was continuous, we used linear mixed models (LMMs) with the lmer function of the same package. Significance tests for both GLMMs and LMMs were performed using Type II Wald Chi-square tests with the Anova function from the car package (Fox & Weisberg, 2019). For all models, post-hoc tests were performed using the emmeans function from the corresponding package (Lenth, 2023), with Tukey correction for multiple comparisons. We also used this package to obtain the model-estimated probabilities that are reported in the figures. The significance level was set at $\alpha = 0.05$. No multiple test correction was applied for analyses of TCQ items, as we examined a series of independent hypotheses rather than a union hypothesis (Rubin, 2024).

Regarding the random effect structure of the models, we initially attempted to fit a maximal model with a random intercept for participants as well as random slopes for each relevant fixed effect and their interactions. In cases of convergence or singularity issues, we gradually reduced the complexity of the random effect structure until the model successfully converged. Subsequently, we applied a backward-selection method to simplify the models further, continuing until any additional reduction in the random effect structure resulted in a significant decrease in goodness-of-fit. We tested this using a likelihood ratio test with a liberal threshold of $\alpha = 0.30$ (Matuschek et al., 2017).

3. Results

3.1. Self-rated quality and difficulty of TCQ ratings, self-censorship, and similarity to daily life experiences

Before analyzing the features and distribution of spontaneous thoughts across methods, we first examined participants' ratings of their overall subjective experience with the different methods and TCQ. Regarding the ratings collected at the end of each session, we first examined the self-reported difficulty in making ratings on the TCQ that participants assessed on a scale ranging from 1 ('Not at all') to 7 ('Totally'). At the end of Session 1, which included the TAP and TPP, participants had a mean rating score of 3.14 (*SD* = 1.59; 64.3 % of ratings below the midpoint of 4). The corresponding mean rating at the end of Session 2 which included the ATSP was 2.60 (*SD* = 1.47; 73.8 % of ratings below 4). Next, regarding the self-reported overall quality of ratings on the TCQ on a scale ranging from - 3 ('Very poor') to + 3 ('Very good'), participants had a mean score of 1.79 (*SD* = 0.57; 81 % of ratings above 1) at the end of Session 1. The corresponding mean score at the end of Session 2 was 1.90 (*SD* = 0.82; 78.6 % of ratings above 1). Taken together, these results suggest that participants generally felt confident in their ratings on the TCQ.

We next examined the ratings given after each laboratory method. First, for the similarity of the reported experiences to those of daily life, which participants rated on a scale ranging from 1 ('Not at all') to 7 ('Totally'), the TAP had a mean score of 6.02 (*SD* = 1.05; 92.9 % of ratings above 4), the TPP had a mean score of 5.69 (*SD* = 1.20; 88.1 % of ratings above 4), and the ATSP had a mean score of 6.45 (*SD* = 0.94; 95.2 % above 4). CLMMs with Method (TAP vs. TPP vs. ATSP) as categorical predictor revealed a significant difference between methods [$\chi^2(2) = 24.30$, $p < 0.001$], with pairwise comparisons indicating that the TAP showed a significantly higher similarity to daily life experiences than the TPP ($Z = 390.07$, $p < 0.001$), while the ATSP exhibited significantly higher ratings than both the TAP ($Z = 647.64$, $p < 0.001$) and the TPP ($Z = 730.68$, $p < 0.001$). Regarding self-censorship, which participants also rated on a 7-point Likert scale ranging from 1 ('Not at all') to 7 ('Totally'), the TAP had a mean score of 2.10 (*SD* = 1.12; 90.5 % of ratings below 4), the TPP a mean score of 1.88 (*SD* = 1.27; 88.1 % of ratings below 4), and the ATSP a mean score of 2.02 (*SD* = 1.52; 83.3 % of ratings below 4), with no significant difference between the three methods [$\chi^2(2) = 3.29$, $p = 0.19$]. These reports of high levels of similarity to daily life experiences and low levels of self-censorship are consistent with those obtained by Raffaelli and colleagues (2021, 2024), who used similar questions but for the TAP only.

3.2. Distribution of mental experience types across methods

Having shown that participants reported no difficulty in using the TCQ or in reporting their mental experiences, we next examined how the different types of experiences (i.e., perceptions, feelings, and thoughts; see section 2.3.5) were distributed across methods. Note that we excluded the ATSP from these analyses because this method included only spontaneous thoughts (in accordance with instructions). Of the 1,386 experiences reported in the TAP, TPP, and DLESP, perceptions accounted for 13.85 % of the total number of reported mental experiences, feelings for 10.68 %, and thoughts for 75.47 %. Mind-wandering episodes were determined as spontaneous thoughts whose task-relatedness and stimulus-dependence ratings made by the participants both fell below 4 on the 7-point Likert scales (for similar dichotomization procedures, see Christoff et al., 2009; Kam et al., 2025). They represented 27.06 % of the total number of mental experiences and 35.85 % of the total number of thoughts.

We conducted GLMMs to investigate differences in the frequency of types of mental experiences across the three methods (TAP, TPP, and DLESP; see Fig. 2 for model estimated probabilities). First, for the model with perceptions (Yes vs. No) as the binary outcome, results revealed a significant main effect of methods [$\chi^2(2) = 46.99$, $p < 0.001$]. Pairwise comparisons indicated no significant differences between the TAP and TPP [$Z = -0.58$, $p = 0.83$], but participants were more likely to rate an experience as a perception in the DLESP than in the TAP [TAP-DLESP: $Z = -5.93$, $p < 0.001$] and TPP [TPP-DLESP: $Z = -5.53$, $p < 0.001$]. Second, the likelihood of an experience being rated as a feeling did not differ between methods [$\chi^2(2) = 5.43$, $p = 0.07$]. Third, we found a significant main effect of methods on thought ratings [$\chi^2(2) = 49.22$, $p < 0.001$]. Pairwise comparisons showed no significant differences between the TAP and TPP [$Z = -0.25$, $p = 0.97$] but participants were less likely to rate experiences as thoughts in the DLESP than in the TAP [TAP- DLESP: $Z = 6.12$, $p < 0.001$] and TPP [TPP-DLESP: $Z = 5.84$, $p < 0.001$]. Lastly, we found a significant main effect of methods [$\chi^2(2) = 10.25$, $p = 0.006$] regarding the rate of mind-wandering episodes among the reported thoughts. There were no differences between the TAP and TPP [$Z = 1.65$, $p = 0.22$] but the likelihood of a thought being classified as a mind-wandering episode was lower in the DLESP than in the TAP [TAP-DLESP: $Z = 3.18$, $p = 0.004$] and TPP [TPP-DLESP: $Z = 2.46$, $p = 0.037$]. Importantly, applying different thresholds to determine mind-wandering episodes–either more stringent (<3) or more lenient (<5)–to the task-relatedness and stimulus-dependence dimensions led to similar results (see Supplementary materials for detailed analyses). In summary, these analyses did not reveal any differences between the TAP and TPP for any type of experience, but experiences reported during the DLESP were more likely to be perceptions and less likely to be thoughts or mind-wandering episodes compared to the TAP and TPP.

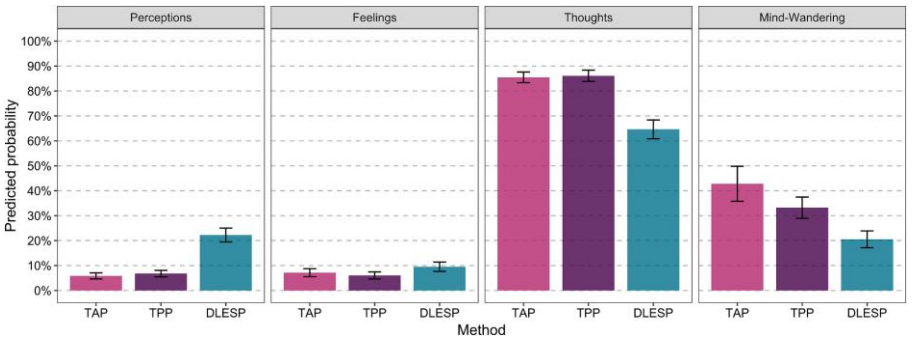

Fig. 2. Predicted probabilities of the four types of mental experiences (Perceptions, Feelings, Thoughts, and Mind-wandering episodes) across methods. Models examining Perceptions, Feelings, and Thoughts included all reported experiences. The model focusing on Mind-wandering episodes was restricted to experiences classified as thoughts.

3.3. Comparisons of spontaneous thought characteristics across methods

After examining the overall subjective experience of the participants and how the different types of mental experiences were distributed across methods, we further narrowed the focus of our analyses down to the main goal of the study: whether the phenomenological features of spontaneous thoughts differed between methods (TAP, TPP, DLESP, and ATSP). We used CLMMs when Likert scale items were the dependent variables and GLMMs for binary choice items (e.g., future orientation, see Table 1). Table 2 summarizes the results for the dimensions assessed with Likert scales and Table 3 summarizes the results for the binary dimensions (see Fig. 3 for model estimated probabilities regarding the Likert-scale items and Fig. 4 for the binary choice items).

First, the TAP and TPP differed on only 2 of the 17 dimensions assessed with Likert scales in the TCQ, with thoughts from the TAP being rated as less stimulus-dependent and more focused on important topics than those from the TPP. With regards to the binary items, we found no differences between the TAP and TPP in terms of temporal orientation (i.e., past, present, future, or no precise temporal orientation). Reappraisal of a situation was more often attributed to the thoughts reported during the TAP than TPP, but this was the only function for which the likelihood of being selected differed between the two methods. Note that participants attributed the 'other' and 'awake' functions to only eight and five percent of the total number of thoughts, respectively, which led to convergence issues in our statistical models. We therefore decided to exclude these functions for all subsequent analyses.

Second, there was no significant difference between the TAP and DLESP for 18 of the 27 TCQ dimensions and this ratio was similar for the TPP-DLESP comparison. Six of the nine dimensions that differed from the DLESP overlapped between the TAP and TPP. More specifically, compared to the DLESP, thoughts reported during the TAP and TPP were rated as less related to the current task, less deliberate, more focused on the past, and less likely to involve decision-making, problem-solving, or planning functions. Thoughts rated during the TPP additionally differed from those in the DLESP on three dimensions: they were less about other people, important topics, and personal goals. In contrast, thoughts reported during the TAP specifically differed from those in the DLESP by being rated as less mundane, more often without any temporal orientation, and more likely to involve a reappraisal function.

Finally, TCQ ratings for the ATSP showed more differences with the other methods. They differed from the TAP and TPP on 13 out of 27 dimensions, with 12 of these differences shared between the two methods. Additionally, they differed from the DLESP on 10 dimensions, all of which overlapped with the differences observed in the TAP-TPP vs. ATSP comparisons. More specifically, thoughts reported during the ATSP were rated lower in task-relatedness, stimulus-dependence, and were less likely to be rated as having no function than those reported in the TAP, TPP, and DLESP. Additionally, they were rated higher in self-relatedness, importance, goal-relatedness, repetition in daily life, similarity to daily life thoughts, and were more likely to be rated as future-oriented and as having a planning function compared to the three other methods. Thoughts from the ATSP were also more likely to involve decision-making or problem-solving than those from the TAP and TPP. Lastly, they were rated as more mundane than those from the TAP and more other-related than those from the TPP.

Table 2

Comparisons of TCQ dimensions assessed with 7-point Likert scales between the four methods.

Dimensions	Main effect X^2, $df = 3$	Pairwise comparisons *Z*-ratio					
		TAP vs. TPP	TAP vs. DLESP	TAP vs. ATSP	TPP vs. DLESP	TPP vs. ATSP	DLESP vs. ATSP
Likert Scales							
Visual Format	6.53 ($p = .09$)	NA	NA	NA	NA	NA	NA
Inner Speech Format	5.17 ($p = .16$)	NA	NA	NA	NA	NA	NA
Task-relatedness	39.07 ($p < .001$)***	-0.81 ($p = .85$)	-5.16 ($p < .001$)***	4.16 ($p < .001$)***	-6.03 ($p < .001$)***	4.76 ($p < .001$)***	6.72 ($p < .001$)***
Stimulus-dependence	54.58 ($p < .001$)***	-3.17 ($p = .008$)**	-2.14 ($p = .14$)	5.88 ($p < .001$)***	1.02 ($p = .73$)	8.55 ($p < .001$)***	7.21 ($p < .001$)***
Concreteness	4.39 ($p = .22$)	NA	NA	NA	NA	NA	NA
Affective Valence	5.19 ($p = .16$)	NA	NA	NA	NA	NA	NA
Self-relatedness	24.75 ($p < .001$)***	2.40 ($p = .08$)	1.78 ($p = .28$)	-3.05 ($p = .01$)*	-0.13 ($p > .99$)	-4.67 ($p < .001$)***	-4.85 ($p < .001$)***
Other-relatedness	14.32 ($p = .002$)**	1.92 ($p = .22$)	-1.58 ($p = .39$)	-1.14 ($p = .67$)	-3.48 ($p = .003$)**	-2.90 ($p = .02$)*	0.32 ($p = .99$)
RepetXP	5.62 ($p = .13$)	NA	NA	NA	NA	NA	NA
RepetDL	72.20 ($p < .001$)***	1.99 ($p = .19$)	0.23 ($p > .99$)	-11.41 ($p < .001$)***	-1.52 ($p = .43$)	-11.35 ($p < .001$)***	-10.89 ($p < .001$)***
Importance	62.08 ($p < .001$)***	2.81 ($p = .03$)*	-0.02 ($p > .99$)	-8.09 ($p < .001$)***	-2.69 ($p = .04$)*	-10.59 ($p < .001$)***	-8.34 ($p < .001$)***
Goal-relatedness	66.17 ($p < .001$)***	2.15 ($p = .14$)	-2.26 ($p = .11$)	-8.63 ($p < .001$)***	-4.27 ($p < .001$)***	-11.23 ($p < .001$)***	-5.92 ($p < .001$)***
Mundaneness	15.67 ($p = .001$)***	-0.94 ($p = .78$)	-3.77 ($p < .001$)***	-3.73 ($p = .001$)**	-2.47 ($p = .06$)	-2.47 ($p = .06$)	-0.39 ($p = .98$)
Structure	6.76 ($p = .08$)	NA	NA	NA	NA	NA	NA
Deliberateness	10.48 ($p = .01$)*	-0.31 ($p = .99$)	-3.62 ($p = .002$)**	-1.80 ($p = .27$)	-2.59 ($p = .047$)*	-1.39 ($p = .51$)	0.83 ($p = .84$)
Suppression	6.19 ($p = .10$)	NA	NA	NA	NA	NA	NA
Similarity	38.46 ($p < .001$)***	0.72 ($p = .89$)	-0.37 ($p = .98$)	-6.34 ($p < .001$)***	-0.93 ($p = .79$)	-6.08 ($p < .001$)***	-5.72 ($p < .001$)***

*$p < .05$; **$p < .01$; ***$p < .001$. RepetXP: Repetition in the experiment; RepetDL: Repetition in daily life.

Table 3

Comparisons of binary TCQ dimensions between the four methods.

Dimensions	Main effect X^2, $df = 3$	Pairwise comparisons *Z*-ratio					
		TAP vs. TPP	TAP vs. DLESP	TAP vs. ATSP	TPP vs. DLESP	TPP vs. ATSP	DLESP vs. ATSP
Temporal Orientation							
Past	25.79 ($p < .001$)***	1.89 ($p = .23$)	4.75 ($p < .001$)***	2.09 ($p = .15$)	3.36 ($p = .004$)**	0.92 ($p = .79$)	-1.94 ($p = .21$)
Present	7.01 ($p = .07$)	NA	NA	NA	NA	NA	NA
Future	55.38 ($p < .001$)***	0.32 ($p = .99$)	-0.82 ($p = .84$)	-6.79 ($p < .001$)***	-0.94 ($p = .78$)	-5.63 ($p < .001$)***	-5.29 ($p < .001$)***
None	8.76 ($p = .03$)*	1.16 ($p = .65$)	2.76 ($p = .03$)*	1.26 ($p = .59$)	1.67 ($p = .34$)	0.26 ($p = .99$)	-1.34 ($p = .54$)
Function							
Decision Making	33.06 ($p < .001$)***	0.63 ($p = .92$)	-3.07 ($p = .01$)*	-5.09 ($p < .001$)***	-2.60 ($p = .046$)*	-3.89 ($p < .001$)***	-2.36 ($p = .08$)
Problem Solving	33.42 ($p < .001$)***	-0.31 ($p = .99$)	-4.20 ($p < .001$)***	-4.29 ($p < .001$)***	-3.90 ($p < .001$)***	-4.00 ($p < .001$)***	-0.32 ($p = .99$)
Planning	100.48 ($p < .001$)***	0.79 ($p = .86$)	-3.65 ($p = .001$)**	-8.85 ($p < .001$)***	-3.75 ($p = .001$)**	-8.08 ($p < .001$)***	-4.72 ($p < .001$)***
Reappraisal	21.02 ($p < .001$)***	3.22 ($p = .007$)**	4.00 ($p < .001$)***	1.70 ($p = .32$)	0.05 ($p > .99$)	-1.48 ($p = .45$)	-1.77 ($p = .29$)
Self-entertainment	1.74 ($p = .63$)	NA	NA	NA	NA	NA	NA
No Function	22.02 ($p < .001$)***	-0.10 ($p > .99$)	2.57 ($p = .05$)	4.10 ($p < .001$)***	2.17 ($p = .13$)	3.87 ($p < .001$)***	2.59 ($p = .047$)*

*$p < .05$; **$p < .01$; ***$p < .001$.

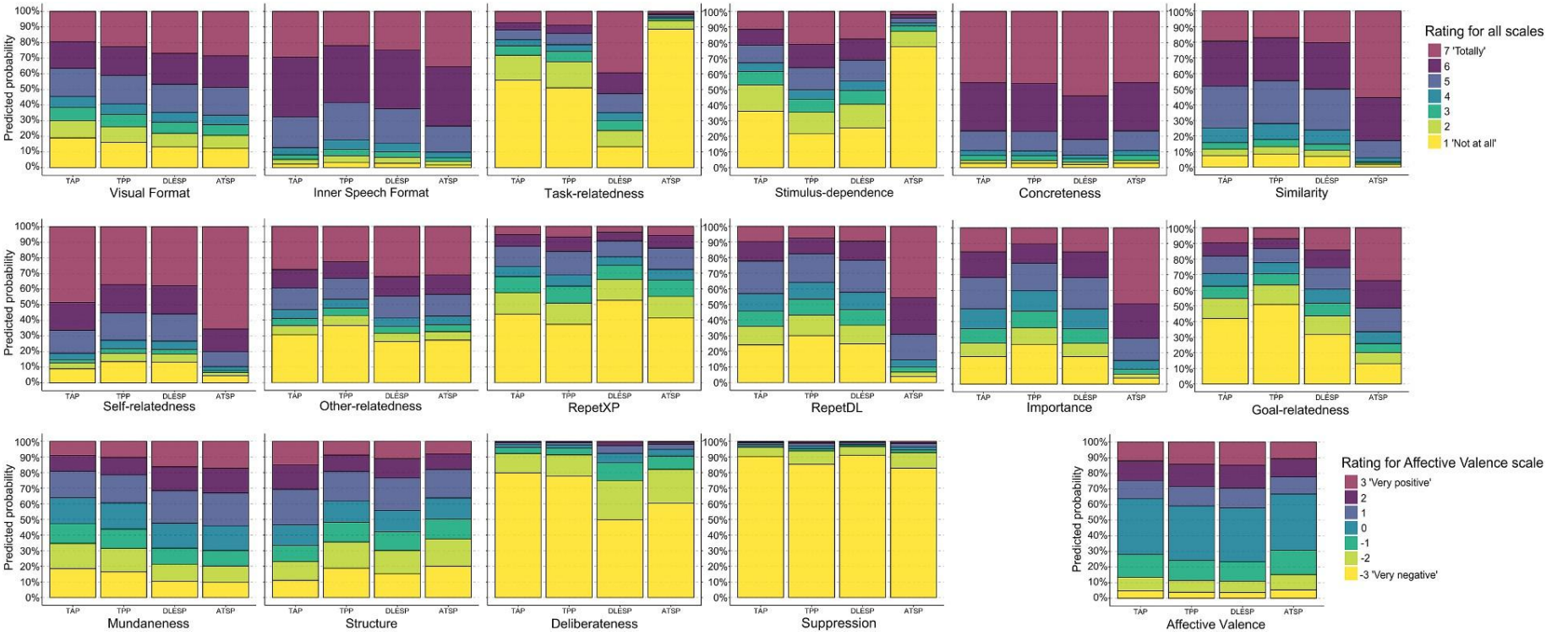

Fig. 3. Predicted probabilities of TCQ ratings across the four methods. *$p < .05$; **$p < .01$; ***$p < .001$. RepetXP: Repetition in the experiment; RepetDL: Repetition in daily life.

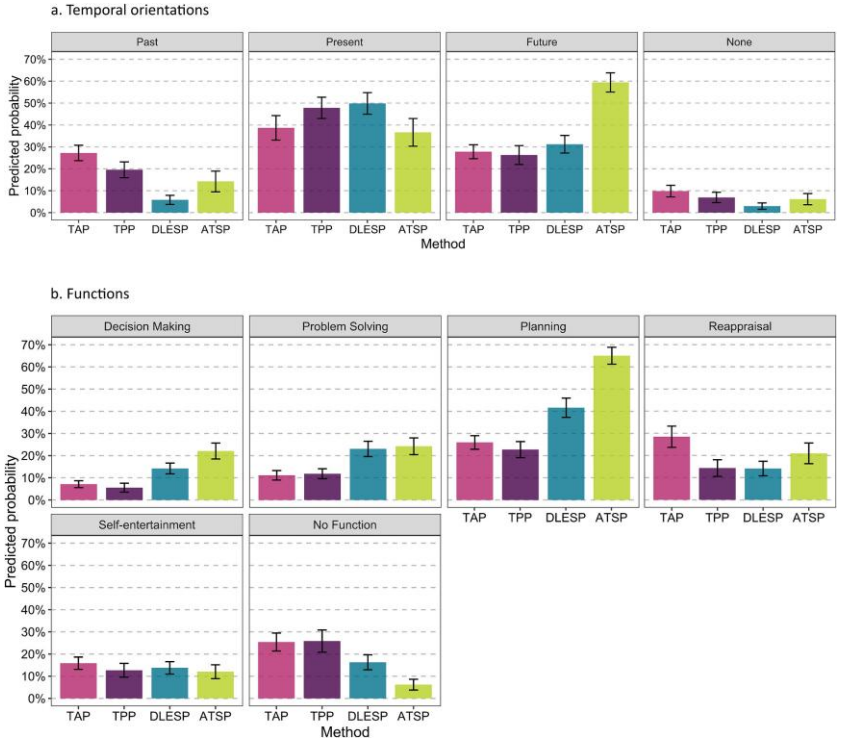

Fig. 4. Predicted probabilities of thought Temporal orientations and Functions across methods.

3.4. Bayesian analysis

Since most of the above analyses did not show significant difference between the TAP and TPP, we conducted Bayesian analyses to determine the extent to which the data supports the null hypothesis (H_0; an absence of differences between the two methods). To do so, we first computed CLMMs with and without the fixed effect for the TAP vs. TPP contrast. We then compared the Bayesian Information Criterion of these two models using the BIC function from the base stats package of R to compute the Bayes Factor (BF) and the logBF (see Fig. 5). In line with the significant differences found in the frequentist analyses, results from the Bayesian analyses revealed moderate evidence in favor of the alternate hypothesis (H_1) for stimulus-dependence and the reappraisal of a situation function (both LogBF > 1.10). However, results regarding importance were inconclusive. As for the remaining 24 dimensions, we found either moderate or strong evidence in favor of H_0 for all dimensions (LogBF < -1.10 and < -2.30, respectively), except goal-relatedness, self-relatedness, and structure, for which the evidence was inconclusive. These results support the view that the TAP is not associated with major reactivity issues compared to the TPP. Among the 27 dimensions assessed with the TCQ to compare these two methods, we found strong evidence in favor of H_0 for 14 dimensions, moderate evidence for H_0 for 7 dimensions, inconclusive evidence for 4 dimensions, and moderate evidence for H_1 for only 2 dimensions.

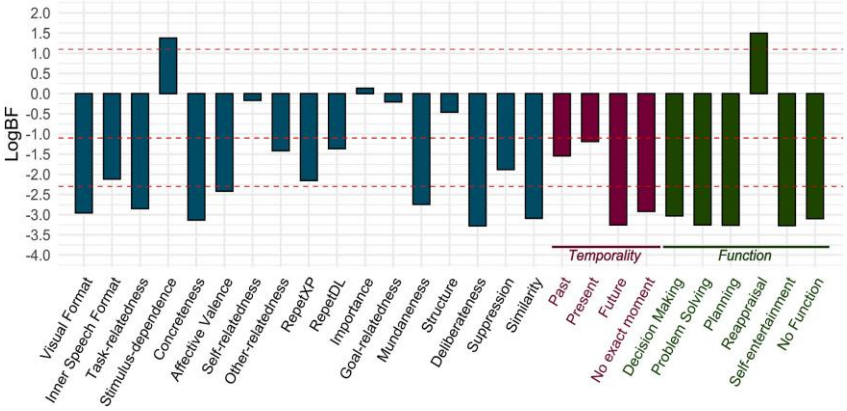

Fig. 5. Logarithmic Bayes factors (logBF) between the TAP and TPP for all thought dimensions. Red dashed lines at ±1.10 and -2.30 represent thresholds for moderate and strong evidence, respectively. Positive values indicate support for the alternative hypothesis (H_1), while negative values indicate support for the null hypothesis (H_0). RepetXP: Repetition in the experiment; RepetDL: Repetition in daily life.

3.5. Prospective bias in spontaneous thoughts

A common feature of spontaneous thoughts, especially mind-wandering, is a prospective bias with future-oriented thoughts being more frequent than those with other temporal orientations or no temporal orientation. To further investigate the validity of the TAP relative to other methods, we examined whether we could replicate this prospective bias for spontaneous thoughts and mind-wandering episodes in the present study, and whether this bias varied depending on the method used.

We computed a 4 Method (TAP, TPP, DLESP, and ATSP) by 4 Temporal Orientation (Future, Past, Present, and None) LMM, using the proportion of thoughts in each cell for each participant as the dependent variable. Results revealed a significant effect of Temporal Orientation [$\chi^2(3) = 114.30$, $p < 0.001$] and Method by Temporal Orientation Interaction [$\chi^2(9) = 69.63$, $p < 0.001$]. However, contrast analyses failed to reveal that future-oriented thoughts were more frequent than the other temporal orientations for any of the methods except for the ATSP (see Table 4). A similar analysis focusing on mind-wandering episodes only for the TAP, TPP, and DLESP also revealed a main effect of Temporal Orientation [$\chi^2(3) = 94.49$, $p < 0.001$] but no Method by Temporal Orientation Interaction [$\chi^2(6) = 8.81$, $p = 0.18$]. Subsequent post-hoc contrasts revealed that the prospective bias of mind-wandering was present across the three methods, with the proportion of future-oriented episodes (model-estimated probability = 54.7 %, SE = 3.04 %) being higher than the one of past-oriented (model-estimated probability = 20.6 %, SE = 3.38 %; $t(124) = 7.48$, $p < 0.001$), present-oriented (model-estimated probability = 21.5 %, SE = 3.43 %; $t(122) = 7.24$, $p < 0.001$), or no temporal orientation (model-estimated probability = 16.3 %, SE = 3.76 %; $t(108) = 7.94$, $p < 0.001$) episodes. Together, these results show that we replicated the prospective bias of mind-wandering and that this bias did not differ across the TAP, TPP, and DLESP. Using different thresholds on the task-relatedness and stimulus-dependence dimensions of the TCQ to determine mind-wandering episodes led to similar results (see Supplementary materials for detailed analyses).

Table 4
Contrasts between future-oriented thoughts and each of the other three temporal orientations (Past, Present, and None) within each method.

	TAP	TPP	DLESP	ATSP
Contrast	t.ratio (df), p-value			
Future vs. Past	-0.15(430), $p = 0.88$	1.17(430), $p = 0.24$	5.12(430), $p < 0.001$***	7.03(430), $p < 0.001$***
Future vs. Present	-2.43(265), $p = 0.02$*	-3.34(265), $p < 0.001$***	-2.55(265), $p = 0.01$*	3.32(265), $p = 0.001$**
Future vs. None	2.66(432), $p = 0.008$**	3.26(432), $p = 0.001$**	5.59(432), $p < 0.001$***	9.06(432), $p < 0.001$***

Note. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

3.6. Spontaneous thought recall

A secondary aim of the present study was to assess whether retrospectively recalling mental experiences induces a memory bias toward certain types of thoughts, and whether this bias varies across methods. Observing similar recall rate between the TAP and TPP, with the same features predicting subsequent thought memory, would indicate that the TAP is as valid as the TPP for investigating memory for spontaneous thought.

As a manipulation check, we first examined whether participants retrospectively reported, at the end of the experiment, that they had guessed at the end of the first day or of the DLESP that they would be asked to remember their mental experiences. Participants provided these ratings on a scale from 1 ('Not at all') to 7 ('Totally'). For the retrospective rating for the first day, participants had a mean score of 1.90 ($SD = 1.43$, range: 1–6), with 3 participants rating higher than 4. For the DLESP, the mean score was 2.69 ($SD = 1.72$, range: 1–7), with 9 participants rating higher than 4. These ratings were significantly higher for the DLESP than the first day [$\chi^2(1) = 10.16$, $p = 0.001$] but no participant reported having used any strategies to remember their mental experiences. Furthermore, GLMMs revealed no significant association between the rating for the first day and the likelihood of correct thought recall from the TAP and TPP [$\chi^2(1) = 0.17$, $p = 0.68$] or between the rating for the second day and the likelihood of correct thought recall from the DLESP [$\chi^2(1) = 0.85$, $p = 0.36$]. These results indicate that retrospectively reporting having guessed the purpose of the experiment was unrelated to memory performance and we therefore decided to use data from all participants in subsequent analyses.

We examined if the likelihood of correct thought recall differed between the three methods (TAP, TPP, and DLESP). The main effect of Method was significant [$\chi^2(2) = 45.48$, $p < 0.001$] and post-hoc tests revealed that the probability of correct thought recall did not differ between the TAP and TPP ($Z = -1.87$, $p = 0.15$; model-estimated probability for the TAP = 51 %, $SE = 3.2$ %; TPP = 59 %, $SE = 3.2$ %), but the probability of thought recall was significantly lower in the DLESP compared to the TAP and TPP (TAP-DLESP: $Z = 5.42$, $p < 0.001$; TPP-DLESP: $Z = 6.72$, $p < 0.001$; model-estimated probability for the DLESP = 24 %, $SE = 3.4$ %).

Our next aim was to determine whether correct recall probability varied depending on the TCQ ratings, and whether these effects differed by methods. Each Likert-scale variable was centered at the participant level for these analyses and Table 5 summarizes the results (see Fig. S1 for model estimated recall probabilities by Methods regarding the Likert-scale items and Fig. S2 for the Temporal orientations and Functions items). We found that thoughts rated as more visual, more important, more related to others, more similar to daily life thoughts, more affectively positive, and having a self-entertainment function were more likely to be accurately recalled. In contrast, thoughts were less likely to be correctly recalled when their function was to reappraise a situation. None of the interaction effect were significant, except for the present temporal orientation with post-hoc showing that thoughts classified as present oriented in the DLESP had a lower probability of being recalled, whereas this effect was not significant in the TAP and TPP (DLESP: $Z = -2.21$, $p = 0.027$; TAP: $Z = 1.26$, $p = 0.21$; TPP: $Z = 0.58$, $p = 0.56$). Given the absence of significant interaction effects, we conducted Bayesian analyses to determine if the data supported an absence of differences between the TAP and TPP regarding the Dimension by Method Interaction effects (see Table 5), using the same methods as the one presented in section 3.5. Results revealed strong evidence in favor of the null hypothesis for all dimensions except for other-relatedness and planning function.

Table 5
Main effects of dimensions on recall performances and interaction with methods.

Dimension	GLMM Analyses: Dimension Main effect	GLMM Analyses: Dimension by Method Interaction	Bayesian Analyses: Dimension by Method Interaction
	X^2, df = 1	X^2, df = 2	*logBF*
Likert Scales			
Visual Format	9.47 (p = 0.002)**	2.19 (p = 0.33)	−2.72††
Inner Speech Format	0.92 (p = 0.34)	0.66 (p = 0.72)	−2.84††
Task-relatedness	0.2 (p = 0.65)	1.37 (p = 0.50)	−3.13††
Stimulus-Dependence	0.61 (p = 0.44)	1.81 (p = 0.40)	−2.95††
Concreteness	0.01 (p = 0.93)	2.69 (p = 0.26)	−3.21††
Affective Valence	10.06 (p = 0.001)**	1.73 (p = 0.42)	−3.11††
Self-relatedness	0.95 (p = 0.33)	0.34 (p = 0.84)	−3.23††
Other-relatedness	8.11 (p = 0.004)**	3.73 (p = 0.15)	−1.89†
RepetXP	3.31 (p = 0.069)	1.52 (p = 0.47)	−3.21††
RepetDL	1.66 (p = 0.198)	1.59 (p = 0.45)	−3.11††
Importance	7.45 (p = 0.006)**	0.38 (p = 0.83)	−3.12††
Goal-relatedness	0.00 (p = 0.998)	1.91 (p = 0.38)	−2.86††
Mundaneness	0.35 (p = 0.55)	2.30 (p = 0.32)	−3.28††
Structure	0.63 (p = 0.43)	1.15 (p = 0.56)	−2.56††
Deliberateness	0.07 (p = 0.798)	2.14 (p = 0.34)	−3.23††
Suppression	2.93 (p = 0.087)	1.30 (p = 0.52)	−2.84††
Similarity	5.34 (p = 0.021)*	0.63 (p = 0.73)	−3.15††
Temporal Orientation			
Past	0.57 (p = 0.45)	5.17 (p = 0.076)	−2.33††
Present	0.01 (p = 0.94)	6.65 (p = 0.036)*	−3.15††
Future	0.20 (p = 0.66)	4.08 (p = 0.13)	−2.64††
None	1.02 (p = 0.31)	0.99 (p = 0.61)	−3.19††
Function			
Decision Making	0.01 (p = 0.91)	3.33 (p = 0.19)	−2.76††
Problem Solving	1.13 (p = 0.29)	1.09 (p = 0.58)	−2.82††
Planning	0.19 (p = 0.66)	4.92 (p = 0.085)	−0.94
Reappraisal	4.99 (p = 0.026)*	1.13 (p = 0.57)	−2.73††
Self-entertainment	6.99 (p = 0.008)**	0.16 (p = 0.93)	−3.27††
No Function	0.46 (p = 0.497)	1.69 (p = 0.43)	−2.81††

Note. **p < 0.05; ****p < 0.01; *****p < 0.001; † Moderate evidence in favor of the null effect (logBF between -1.10 and -2.30); †† Strong evidence in favor of the null effect (logBF < -2.30). GLMMs were performed on all thoughts from TAP, TPP, and DLESP. Bayesian analyses were limited to the TAP and TPP. See Fig. S1 for model estimated recall probabilities by Methods regarding the Likert-scale items and Fig. S2 for the Temporal orientations and Functions items. RepetXP: Repetition in the experiment; RepetDL: Repetition in daily life.

4. Discussion

The Think-Aloud Protocol (TAP) has recently regained attention among cognitive researchers as a tool for exploring the dynamic nature, content, and correlates of spontaneous thought (Li et al., 2022, 2023, 2024; Raffaelli et al., 2021, 2024, 2025; Sripada & Taxali, 2020). Despite this renewed interest, it remains debated whether continuously thinking aloud alters thought content in comparison to other established investigation methods that involve silent thinking. These include the thought-probe method, which is currently the most commonly used, and retrospective methods, in which participants must report and rate their thoughts after a delay. These latter methods raise the question of whether recalling previous mental experiences induces a memory bias toward thoughts with particular features (e.g., Brewer, 1988; Stawarczyk & D'Argembeau, 2019), and whether this recall bias differs between methods.

To address these questions, we conducted a systematic comparison of the TAP with the Thought-Probe Protocol (TPP) during a 7.5- minute resting period. By subsampling what participants had in mind at 10 time points during the TAP—coinciding with the probes in the TPP—we ensured that the same number of mental experiences, separated by similar time intervals, were compared between the two methods. The only difference was that participants either spoke their thoughts aloud (TAP) or thought silently (TPP) in the intervals between the assessed mental experiences. Additionally, we included the Daily Life Experience Sampling Protocol (DLESP) and the Autobiographic Thought Sampling Protocol (ATSP). The DLESP allowed for thought collection in a more naturalistic context, while the ATSP provided retrospective insights into participants' frequently occurring thoughts in daily life. Finally, participants performed surprise free recall tasks after a one-day delay of their reported mental experiences from the TAP, TPP, and DLESP.

With regard to the comparison between the TAP and TPP, results showed that participants reported low levels of self-censorship in both methods (for similar findings regarding the TAP, see Raffaelli et al., 2021, 2024; Garg et al., 2025). The TAP and TPP did not differ in the frequency of spontaneous thoughts and mind-wandering episodes. When examining specific thought features rated on the Thought Characteristics

Questionnaire (TCQ), Bayesian analyses supported the null hypothesis (H_0) for 21 out of the 27 assessed dimensions (see Fig. 5). First, both methods yielded comparable ratings with strong evidence in favor of H_0 for visual thought format, task-relatedness, concreteness, deliberateness, mundaneness, affective valence, and similarity to daily life experiences. Evidence in favor of H_0 was moderate for inner speech thought format, repetition during the experiment and in daily life, suppression attempts, and whether the thoughts were about other people. Second, there were no differences in temporal orientation: both the TAP and TPP exhibited the expected prospective bias (Baird et al., 2011; Stawarczyk et al., 2011, 2013) for mind-wandering episodes, which did not differ from that of daily life thoughts assessed with the DLESP. Bayesian analyses revealed strong evidence in favor of H_0 for the future and no temporal orientation dimensions, while evidence was moderate for the past and present dimensions. Third, the TAP and TPP were characterized by similar proportions of thoughts involving decision making, problem solving, planning, self-entertainment, or with no apparent function, with strong evidence in favor of H_0 for all these dimensions. Finally, regarding memory for thoughts, we found no differences in the proportion of correctly recalled thoughts between the TAP and TPP after the one-day delay. Bayesian analyses showed that the same thought features predicted thought recall for both methods with strong evidence in favor of H_0, except for other-relatedness for which evidence was moderate and the planning function for which results were inconclusive (see Table 5). Finally, although results of the frequentist analysis were not significant, Bayesian analyses revealed inconclusive evidence in favor of either H_0 or H_1 for self-relatedness, goal-relatedness, and structure. Overall, these findings align with recent studies supporting the validity of the TAP as a method for assessing spontaneous thought (e.g., Garg et al., 2025; Li et al., 2023; Raffaelli et al., 2021; Sripada & Taxali, 2020), and extend this prior work by showing that thinking aloud introduces minimal reactivity compared to the widely used thought-probe method.

While largely similar, the experiences reported during the TAP and TPP nonetheless differed on some characteristics. First, participants reported, at the end of each task, that their mental experiences from the TAP were overall more similar to those of daily life compared to their mental experiences from the TPP (but note that ratings for individual thoughts did not differ on this dimension in the TCQ). Second, Bayesian analyses supported the alternative hypothesis with moderate evidence for two of the dimensions assessed with the TCQ (see Fig. 5). Specifically, stimulus-dependency ratings were lower, and participants more often attributed a reappraisal function to thoughts from the TAP than to those from the TPP. Results from frequentist analyses also showed that thoughts from the TAP were rated higher in personal importance, but this effect was inconclusive in the Bayesian analyses. These differences may stem from the fact that remaining silent until prompted in the TPP freed participants' attentional resources from the perceptual load imposed by hearing one's own voice in the TAP. As a result, more thoughts in the TPP may have been triggered by perceptions of the surrounding environment, whereas in the TAP, thoughts may be more often triggered by one's own previously verbalized thoughts. This, in turn, may have led participants in the TAP to focus on more personally relevant content, while the neutral testing room environment in the TPP was less likely to elicit meaningful thoughts involving the reappraisal of situations. Supporting this interpretation, there is evidence from studies on involuntary memories and future thinking indicating that both the external environment and previously experienced thoughts can serve as cues that trigger new thought content (for reviews, see Berntsen, 2021; Cole & Kvavilashvili, 2021; Kvavilashvili & Rummel, 2020). This research suggests that thoughts are more frequently triggered by the external environment than internal mentation. However, it remains unclear how thought triggers are affected when the sensory stream includes hearing one's own verbalized thoughts. Interestingly, Garg and colleagues (2025, Study 2) found no differences in the *meta*-awareness of topic shifts between silent thinking and the TAP, suggesting that verbalizing one's thoughts may not make participants more aware of what triggered them. Future studies should include items that specifically assess thought triggers and the awareness of these triggers in the TCQ to further investigate this question.

While the TAP and TPP demonstrated high similarity in capturing spontaneous thought content, both laboratory methods differed from the DLESP. First, experiences sampled in daily life consisted more often in perception and less often in spontaneous thought or mind-wandering episodes. Furthermore, both the TAP and TPP differed from the DLESP in 9 out of the 27 assessed thought dimensions. Compared to thoughts sampled in the laboratory, those assessed in daily life via the smartphone app were rated as more related to the current task, more deliberate, less focused on the past, and more likely to serve decision-making, problem-solving, or planning functions. Additionally, thoughts from the DLESP were more likely than those from the TPP to be about other people, important topics, and personal goals. Compared to thoughts from the TAP, those from the DLESP were rated as more mundane, less likely to involve the reappraisal of a situation, and less frequently described as lacking temporal orientation. These differences likely reflect the fact that, in daily life, participants spend more time interacting with their environment while engaged in meaningful tasks and activities, in contrast to the artificial resting period in a minimally furnished room during the TAP and TPP. More generally, to the best of our knowledge, this is the first study to directly compare the features of individual thoughts sampled in daily life versus in the lab with the TAP and TPP. The influence of context on thought features observed here aligns with recent findings showing that the characteristics of spontaneous thoughts sampled in the laboratory vary depending on the task participants are performing (Konu et al., 2021). Our results also extend the correlational findings of Linz and colleagues (2019), who examined individual differences and found no relationship between the frequency of task-unrelated, positively valenced, and past-oriented thoughts between the lab and daily life, whereas ratings on other dimensions, such as future orientation and thinking about oneself or others, were strongly correlated between these two settings. Together, these findings suggest that researchers should be cautious when generalizing findings on thought content and features from laboratory-based to daily life experience sampling procedures (for a related discussion regarding the correlates of mind-wandering reports across contexts, see Kane et al., 2017).

The last aim of this study was to investigate memory for spontaneous thoughts—specifically, whether the process of recalling prior mental experiences systematically favors thoughts with particular characteristics, and whether any such bias differs between the TAP and other methods. As mentioned above, we found no differences between the TAP and TPP in the features that predicted subsequent thought recall. Additionally, no differences were found between the DLESP and the two laboratory methods, except that present- oriented thoughts were less likely to be recalled in the DLESP, but not in the TAP and TPP. Taken together, these results suggest that the TAP is as valid as the TPP for investigating memory for spontaneous thoughts. Importantly, regardless of the assessment method, thoughts that were rated higher in visual format, personal

importance, being about other people, positive affective valence, similarity to daily life thoughts, and self-entertainment function were more likely to be correctly recalled. In contrast, thoughts with a reappraisal function were less likely to be recalled. The few preliminary studies that have investigated memory for spontaneous thoughts found better memory for thoughts that were more pleasant, personally significant, and exciting (Brewer, 1988), as well as for those about more important or unusual topics, serving a planning function, coming more frequently to mind, and that participants attempted to suppress (Stawarczyk & D'Argembeau, 2019). Another kind of research examined memory for future-oriented thoughts and found that future-oriented thoughts were better recalled when rated as more personally important, involved more intense emotions, clearer representations of other people, and were goal-related (Jeunehomme & D'Argembeau, 2017, 2021). These previous results were generally replicated in the present study and were also consistent with the characteristics of the daily life thoughts retrospectively reported with the ATSP (Andrews-Hanna et al., 2013, 2022). Compared to thoughts reported in the other methods, those from the ATSP were rated higher in self-relatedness, importance, goal-relatedness, repetition in daily life, and similarity to daily life thoughts, and were more often future-oriented and associated with a planning function. Together, these findings illustrate that spontaneous thoughts recalled through retrospective assessment methods are likely biased toward certain features—such as personal importance—which may be overrepresented in participants' reports.

More generally, these results underscore the paucity of research on memory for spontaneous thoughts, which may be critical for evaluating the validity of retrospective reports and understanding how such thoughts are recalled to inform decisions and actions in daily life. Interestingly, several dimensions associated with thought recall, such as personal significance or affective valence (Brewer, 1988; Jeunehomme & D'Argembeau, 2017, 2021; Stawarczyk & D'Argembeau, 2019), have also been shown to influence memory for perceived stimuli and daily life events. More precisely, Jeunehomme & D'Argembeau (2017) proposed that visual imagery supports memory consolidation of imagined future scenarios by enhancing the vividness and distinctiveness of mental representations, thereby making them more easily retrievable. It is also possible that verbally describing mental images, as participants were instructed to do in this study, promotes multimodal encoding, which may enhance memory retention in a manner similar to how multisensory stimuli are better encoded than unisensory ones (e.g., Murray & Shams, 2023). Personal importance has likewise been identified as a strong predictor of thought recall (Brewer, 1988; Stawarczyk & D'Argembeau, 2019). According to the self-memory system theory (Conway, 2001; Conway & Holmes, 2004), events that are personally important are more likely to be integrated into an individual's internal life schema, which facilitates their retrieval. Thoughts rated as similar to those typically experienced in daily life may benefit from schema- consistent encoding and repeated mental rehearsal, enhancing their accessibility during retrieval (Brewer, 1988), much like repeated exposure to perceived stimuli improves memory (Ebbinghaus, 1913). Positive emotions reflected by higher ratings on the affective valence and self-entertainment items of the TCQ have also been shown to facilitate thought recall, possibly by promoting greater engagement and elaborative processing during the encoding phase (Szpunar et al., 2012; Brewer, 1988). Future studies should investigate why the well-established memory advantage for negatively valenced perceived events (Kensinger & Ford, 2020) was not observed here. The finding that thoughts involving other people are more likely to be recalled is consistent with the idea that people are central elements of autobiographical memory and future event simulation (Conway & Holmes, 2004; Jeunehomme & D'Argembeau, 2017), and may serve as powerful associative cues at retrieval. These results are also in line with the models proposing that memory is fundamentally socially oriented (Alea & Bluck, 2003; Bluck & Alea, 2009). Finally, the finding that thoughts with a reappraisal function are less well remembered does not easily align with previous research on memory for thoughts or perceived events. One tentative interpretation is that their content may not integrate well with existing self-related mental models or schemas, leading to weaker encoding (Conway, 2001). It should nonetheless be noted that neuroimaging studies indicate that the retrieval of internal thoughts involves distinct brain areas compared to the recall of external stimuli (Stawarczyk et al., 2018). These findings suggest that the cognitive processes involved in remembering spontaneous thoughts may differ, at least partly, from those engaged in the recall of perceived environmental information, although the extent of these differences remains to be fully delineated.

Our results, which suggest that the TAP is a suitable technique for studying spontaneous thoughts, should not be interpreted as a call to abandon other methods. Instead, we propose that method selection should be guided by two complementary considerations: the theoretical aims and the practical constraints of the planned study. By theoretical aims, we mean the specific measures required to assess theory-based hypotheses. For example, the TAP—by capturing a continuous stream of mental experiences—is particularly useful when researchers need fine-grained information about the dynamics of thoughts to investigate phenomena such as topic shifts (Sripada & Taxali, 2020), the prevalence of direct versus generative retrieval of memories (Uzer et al., 2012), or the evolution of ruminative and disorganized thought sequences in depression and ADHD, respectively (Christoff et al., 2016; Raffaelli et al., 2021, 2025). According to the present findings, the choice of method appears to have little impact on the assessment of some dimensions that seem invariant across methods, including visual format, inner speech, concreteness, affective valence, thought repetition during the experiment, structure, suppression, present temporality, and the self-entertainment function. At a practical level, the TAP is particularly well suited for studying the occurrence of infrequent subtypes of spontaneous thoughts—such as atemporal mind-wandering (Stawarczyk, 2018)—as it allows for the continuous recording of each occurrence of such thought during the verbalization period. In contrast, laboratory probe-caught methods are particularly useful for investigating the behavioral or physiological correlates of spontaneous thoughts, as they impose minimal constraints on how participants engage with the task between thought probes (i.e., participants are silently thinking). However, this advantage comes at the cost of reduced precision in determining thought onsets and capturing the exact nature and content of mental experiences that occur between probes. Next, more ecological sampling methods (as the DLESP) are optimal for assessing theories that emphasize the interaction between thought and everyday context. Finally, retrospective listing (as the ATSP) selectively captures thoughts that participants are able to recall and, therefore, that are more likely to influence future behavior (Andrews-Hanna et al., 2013), making it especially useful to examine long-term cognitive themes or goal management.

Several limitations of the present study should be acknowledged. First, all verbal protocols were collected in French from a European sample. The few studies that used thought probes and retrospective methods to assess cultural influences on mind-wandering are mixed, with some suggesting consistency in the core features of spontaneous thoughts across cultures (e.g., Gonçalves et al., 2017), while others suggest a cultural

influence on the correlates of mind-wandering frequency and how it is affected by age (e.g., Martinon et al., 2019). It remains unclear whether these findings extend to the TAP, given its greater reliance on continuous verbalization, which may amplify the influence of language structure and cultural norms on how thoughts are reported. Second, our comparisons were based on thoughts assessed during a resting state. Whether the TAP remains non-reactive and effective for investigating spontaneous thoughts when a demanding primary task is added remains to be tested. Future work should incorporate probe-caught and think-aloud methods within the same activity (e.g., education, Kesler et al., 2016; text comprehension, Wang, 2016; sport, Samson et al., 2017) or dual-task setting to isolate potential interactions between method, cognitive load, context, and content collected. For instance, a vigilance task with embedded cue words has been successfully used with both the self-caught and probe-caught methods to assess the features of involuntary memories and future thoughts (Barzykowski et al., 2024; Barzykowski & Niedzwienska, 2018). It would be interesting to investigate whether this task can also be effectively combined with the TAP, notably to determine the occurrence of directly vs generatively retrieved memories and future thoughts (Cole & Kvavilashvili, 2019; Mace et al., 2021; Uzer et al., 2012). Including items in the TCQ that assess these subtypes of thoughts and their triggers in future studies could help improve our understanding of memory retrieval processes. Third, mind-wandering episodes were identified with a somewhat arbitrary cutoff of < 4 on both task-relatedness and stimulus-relatedness dimensions (similarly to Christoff et al., 2009; Kam et al., 2025). Although using a different cutoff did not influence our results, future research may benefit from using binary forced-choice classification of mind- wandering episodes, as recent studies suggest that dichotomous judgments may increase the reliability and consistency of reports across participants and paradigms (e.g., Kane et al., 2021).

In conclusion, this study aimed to address a gap in the literature regarding the validity of the TAP for assessing spontaneous thought relative to other methods, and particularly in comparison to the TPP which is the most commonly used method in the field. Our findings revealed that the TAP and TPP yield highly similar thoughts, suggesting that both methods effectively capture spontaneous mental content. These results thus support the usefulness of the TAP for studying spontaneous cognition in the laboratory, offering real-time access to ongoing thought processes with minimal reactivity. However, our results also indicate that caution is warranted when generalizing thought features from laboratory settings to daily life, as thoughts reported in the DLESP differed from those reported in the TAP and TPP on one-third of the assessed dimensions. Finally, our results indicate that certain features may be overrepresented in retrospective thought reports, suggesting that real-time (concurrent) experience sampling procedures may offer a less biased assessment of spontaneous thought features.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT and DeepL Write in order to improve language and readability. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

CRediT authorship contribution statement

Arya Gilles: Writing – review & editing, Writing – original draft, Visualization, Supervision, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Gaëlle Panneels:** Writing – review & editing, Investigation, Data curation. **Arnaud D'Argembeau:** Writing – review & editing, Methodology, Funding acquisition, Conceptualization. **David Stawarczyk:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

A. D'Argembeau and D. Stawarczyk are, respectively, Research Director and Research Associate at the Fonds de la Recherche Scientifique (F.R.S.-FNRS). Arya Gilles is funded by a Doctoral Grant in Human Sciences from the University of Liège.

Open practices

Anonymized data and analysis scripts are available on the Open Science Framework (https://osf.io/43b5x/). The experiment was not preregistered.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.concog.2025.103910.

Data availability

Data will be made available on request.

References

Alea, N., & Bluck, S. (2003). Why are you telling me that? A conceptual model of the social function of autobiographical memory. *Memory, 11*(2), 165–178. https:// doi.org/10.1080/741938207

Andrews-Hanna, J. R., Kaiser, R. H., Turner, A. E. J., Reineberg, A. E., Godinez, D., Dimidjian, S., & Banich, M. T. (2013). A penny for your thoughts: Dimensions of self-generated thought content and relationships with individual differences in emotional wellbeing. *Frontiers in Psychology, 4*. https://doi.org/10.3389/ fpsyg.2013.00900

Andrews-Hanna, J. R., Woo, C.-W., Wilcox, R., Eisenbarth, H., Kim, B., Han, J., Reynolds Losin, E. A., & Wager, T. D. (2022). The conceptual building blocks of everyday thought: Tracking the emergence and dynamics of ruminative and non-ruminative thinking. *Journal of Experimental Psychology: General, 151*(3), 628–642. https://doi.org/10.1037/xge0001096

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

Barzykowski, K., & Niedzwienska, A. (2018). Priming involuntary autobiographical memories in the lab. *Memory, 26*(2), 277–289. https://doi.org/10.1080/09658211.2017.1353102

Barzykowski, K., Ilczuk, E., & Kvavilashvili, L. (2024). A comprehensive guide to research protocols for collecting and coding involuntary past and future thoughts. *MethodsX, 12*, Article 102732. https://doi.org/10.1016/j.mex.2024.102732

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/ 10.18637/jss.v067.i01

Baumeister, R. F., Hofmann, W., Summerville, A., Reiss, P. T., & Vohs, K. D. (2020). Everyday thoughts in time: Experience sampling studies of mental time travel. *Personality and Social Psychology Bulletin, 46*(12), 1631–1648. https://doi.org/10.1177/0146167220908411

Berntsen, D. (2021). Involuntary autobiographical memories and their relation to other forms of spontaneous thoughts. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 376*(1817). https://doi.org/10.1098/rstb.2019.0693. Article 20190693.

Bluck, S., & Alea, N. (2009). Thinking and talking about the past: Why remember? *Applied Cognitive Psychology, 23*(8), 1089–1104. https://doi.org/10.1002/acp.1612

Brewer, W. F. (1988). Memory for randomly sampled autobiographical events. In U. Neisser, & E. Winograd (Eds.), *Remembering reconsidered: Ecological and traditional approaches to the study of memory* (pp. 21–90). Cambridge University Press.

Carriere, J. S. A., 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), 19–31. https://doi.org/10.1037/a0031438

Christensen, R. (2023). *Ordinal-regression models for ordinal data*. R package version 2023.12 4.1. https://CRAN.R-project.org/package=ordinal.

Christoff, K., Gordon, A. M., Smallwood, J., Schooler, J. W., & Smith, R. (2009). Experience sampling during fMRI reveals default network and executive system contributions to mind wandering. *Proceedings of the National Academy of Sciences of the United States of America, 106*(21), 8719–8724. https://doi.org/10.1073/ pnas.0900234106

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

Chu, M. T., Marks, E., Smith, C. L., & Chadwick, P. (2023). Self-caught methodologies for measuring mind wandering with meta-awareness: A systematic review. *Consciousness and Cognition, 108*, Article 103463. https://doi.org/10.1016/j.concog.2022.103463

Cole, S., & Kvavilashvili, L. (2019). Spontaneous future cognition: The past, present and future of an emerging topic. *Psychological Research Psychologische Forschung, 83*(4), 631–650. https://doi.org/10.1007/s00426-019-01193-3

Cole, S., & Kvavilashvili, L. (2021). Spontaneous and deliberate future thinking: A dual process account. *Psychological Research Psychologische Forschung, 85*(2), 464–479. https://doi.org/10.1007/s00426-019-01262-7

Conway, M. A. (2001). Sensory–perceptual episodic memory and its context: Autobiographical memory. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 356*(1413), 1375–1384. https://doi.org/10.1098/rstb.2001.0940

Conway, M. A., & Holmes, A. (2004). Psychosocial stages and the accessibility of autobiographical memories across the life cycle. *Journal of Personality, 72*(3), 461–480. https://doi.org/10.1111/j.0022-3506.2004.00269.x

Ebbinghaus, H. (1913). *Memory: A contribution to experimental psychology*. (H. A. Ruger & C. E. Bussenius, Trans.). Teachers College Press. https://doi.org/10.1037/ 10011-000.

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. https://doi.org/10.3758/BF03193146 Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (3rd edition). Sage.

Fox, M. C., Ericsson, K. A., & Best, R. (2011). Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. *Psychological Bulletin, 137*(2), 316–344. https://doi.org/10.1037/a0021663

Garg, A., Shelat, S., Gross, M. E., Smallwood, J., Seli, P., Taxali, A., Sripada, C. S., & Schooler, J. W. (2025). Opening the black box: Think Aloud as a method to study the spontaneous stream of consciousness. *Consciousness and Cognition, 128*, Article 103815. https://doi.org/10.1016/j.concog.2025.103815

Gonçalves, O. F., Rego, G., Oliveira-Silva, P., Leite, J., Carvalho, S., De Souza-Queiroz, J., Fregni, F., Amaro, E., & Boggio, P. S. (2017). Is the relationship between mind wandering and attention culture-specific? *Psychology & Neuroscience, 10*(2), 132–143. https://doi.org/10.1037/pne0000083

Hervé, M. (2023). *RVAideMemoire: Testing and plotting procedures for Biostatistics*. R package version 0.9-83-7. https://CRAN.R-project.org/package=RVAideMemoire.

James, W. (1890). The stream of thought. In *The principles of psychology, Vol I.* (Book: 2004-16192-009; pp. 224–290). Henry Holt and Co. https://doi.org/10.1037/10538-009.

Jeunehomme, O., & D'Argembeau, A. (2017). Accessibility and characteristics of memories of the future. *Memory, 25*(5), 666–676. https://doi.org/10.1080/ 09658211.2016.1205096

Jeunehomme, O., & D'Argembeau, A. (2021). The role of self-reference and personal goals in the formation of memories of the future. *Memory & Cognition, 49*(6), 1119–1135. https://doi.org/10.3758/s13421-021-01150-9

Jordano, M. L., & Touron, D. R. (2018). How often are thoughts metacognitive? Findings from research on self-regulated learning, think-aloud protocols, and mind- wandering. *Psychonomic Bulletin & Review, 25*(4), 1269–1286. https://doi.org/10.3758/s13423-018-1490-1

Kam, J. W. Y., Mittner, M., & Knight, R. T. (2022). Mind-wandering: Mechanistic insights from lesion, tDCS, and iEEG. *Trends in Cognitive Sciences, 26*(3), 268–282. https://doi.org/10.1016/j.tics.2021.12.005

Kam, J. W. Y., Rahnuma, T., Nanjappan Jothiraj, S., Ouellette-Zuk, A. A., & Knight, R. T. (2025). Electrophysiological signatures of ongoing thoughts during naturalistic behavior. *Imaging Neuroscience, 3*, Article IMAG.a.20. https://doi.org/10.1162/IMAG.a.20

Kane, M. J., Gross, G. M., Chun, C. A., Smeekens, B. A., 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*(9), 1271–1289. https://doi.org/10.1177/0956797617706086

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. https://doi.org/10.3758/s13428-021-01557-x

Kawashima, I., Hinuma, T., & Tanaka, S. C. (2023). Ecological momentary assessment of mind-wandering: Meta-analysis and systematic review. *Scientific Reports, 13*(1). https://doi.org/10.1038/s41598-023-29854-9. Article 2873.

Kensinger, E. A., & Ford, J. H. (2020). Retrieval of emotional events from memory. *Annual Review of Psychology, 71*(1), 251–272. https://doi.org/10.1146/annurev- psych-010419-051123

Kesler, T., Tinio, P. P. L., & Nolan, B. T. (2016). What's our position? a critical media literacy study of popular culture websites with eighth-grade special education students. *Reading & Writing Quarterly, 32*(1), 1–26. https://doi.org/10.1080/10573569.2013.857976

Konu, D., Mckeown, B., Turnbull, A., Siu Ping Ho, N., Karapanagiotidis, T., Vanderwal, T., McCall, C., Tipper, S. P., Jefferies, E., & Smallwood, J. (2021). Exploring patterns of ongoing thought under naturalistic and conventional task-based conditions. *Consciousness and Cognition*, *93*, Article 103139. https://doi.org/10.1016/j.concog.2021.103139.

Kvavilashvili, L., & Rummel, J. (2020). On the nature of everyday prospection: A review and theoretical integration of research on mind-wandering, future thinking, and prospective memory. *Review of General Psychology, 24*(3), 210–237. https://doi.org/10.1177/1089268020918843

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. https://doi.org/10.2307/2529310

Lenth, R. (2023). *Emmeans: Estimated marginal means, aka least-squares means*. R package version 1.8.9. https://CRAN.R-project.org/package=emmeans.

Li, H.-X., Chen, X., Wang, Z.-H., Lu, B., Liao, Y.-F., Li, X.-Y., Wang, Y.-W., Liu, Y.-S., Castellanos, F. X., & Yan, C.-G. (2024). Characterizing human spontaneous thoughts and its application in major depressive disorder. *Journal of Affective Disorders, 365*, 276–284. https://doi.org/10.1016/j.jad.2024.08.060

Li, H.-X., Lu, B., Chen, X., Li, X.-Y., Castellanos, F. X., & Yan, C.-G. (2022). Exploring self-generated thoughts in a resting state with natural language processing. *Behavior Research Methods, 54*(4), 1725–1743. https://doi.org/10.3758/s13428-021-01710-6

Li, H.-X., Lu, B., Wang, Y.-W., Li, X.-Y., Chen, X., & Yan, C.-G. (2023). Neural representations of self-generated thought during think-aloud fMRI. *NeuroImage, 265*, Article 119775. https://doi.org/10.1016/j.neuroimage.2022.119775

Linz, R., Pauly, R., Smallwood, J., & Engert, V. (2019). Mind-wandering content differentially translates from lab to daily life and relates to subjective stress experience. *Psychological Research Psychologische Forschung, 85*. https://doi.org/10.1007/s00426-019-01275-2

Mace, J. H., Petersen, E. P., & Kruchten, E. A. (2021). Elucidating the mental processes underlying the direct retrieval of autobiographical memories. *Consciousness and Cognition, 94*, Article 103190. https://doi.org/10.1016/j.concog.2021.103190

Martinon, L. M., Smallwood, J., Hamilton, C., & Riby, L. M. (2019). Frogs' legs versus roast beef: How culture can influence mind-wandering episodes across the lifespan. *Europe's Journal of Psychology, 15*(2), 211–239. https://doi.org/10.5964/ejop.v15i2.1597

Matthews, G. (2021). Stress states, personality and cognitive functioning: A review of research with the Dundee stress state questionnaire. *Personality and Individual Differences, 169*. https://doi.org/10.1016/j.paid.2020.110083

Mathews, N. K., Bin Faiz, U., & Brosowsky, N. P. (2024). How do you know if you were mind wandering? Dissociating explicit memories of off task thought from subjective feelings of inattention. *Open Mind: Discoveries in Cognitive Science, 8*, 666–687. https://doi.org/10.1162/opmi_a_00142

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language, 94*, 305–315. https://doi.org/10.1016/j.jml.2017.01.001

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), 11–18. https://psycnet.apa.org/doi/10.1037/a0031569.

Murray, C. A., & Shams, L. (2023). Crossmodal interactions in human learning and memory. *Frontiers in Human Neuroscience, 17*. https://doi.org/10.3389/ fnhum.2023.1181760

R Core Team (2023). *R: A language and environment for statistical computing* [Logiciel]. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project. org/.

Raffaelli, Q., Malusa, R., de Stefano, N.-A., Andrews, E., Grilli, M. D., Mills, C., Zabelina, D. L., & Andrews-Hanna, J. R. (2024). Creative minds at rest: Creative individuals are more associative and engaged with their idle thoughts. *Creativity Research Journal, 36*(3), 396–412. https://doi.org/10.1080/ 10400419.2023.2227477

Raffaelli, Q., Mills, C., de Stefano, N.-A., Mehl, M. R., Chambers, K., Fitzgerald, S. A., Wilcox, R., Christoff, K., Andrews, E. S., Grilli, M. D., O'Connor, M.-F., & Andrews-Hanna, J. R. (2021). The think aloud paradigm reveals differences in the content, dynamics and conceptual scope of resting state thought in trait brooding. *Scientific Reports, 11*(1). https://doi.org/10.1038/s41598-021-98138-x. Article 19362.

Raffaelli, Q., Rai, S., Galbraith, A., Krupa, A., Buerkner, J., Andrews-Hanna, J. R., Callahan, B. L., & Kam, J. W. Y. (2025). Hyperactive ADHD symptoms are associated with increased variability in thought content in less constrained contexts. *Scientific Reports, 15*(1), 9792. https://doi.org/10.1038/s41598-025-93053-x

Randall, J. G., Oswald, F. L., & Beier, M. E. (2014). Mind-wandering, cognition, and performance: A theory-driven meta-analysis of attention regulation. *Psychological Bulletin, 140*(6), 1411–1431. https://doi.org/10.1037/a0037428

Rubin, M. (2024). Inconsistent multiple testing corrections: The fallacy of using family-based error rates to make inferences about individual hypotheses. *Methods in Psychology, 10*, Article 100140. https://doi.org/10.1016/j.metip.2024.100140

Samson, A., Simpson, D., Kamphoff, C., & Langlier, A. (2017). Think aloud: An examination of distance runners' thought processes. *International Journal of Sport and Exercise Psychology, 15*(2), 176–189. https://doi.org/10.1080/1612197X.2015.1069877

Smallwood, J., & Schooler, J. W. (2006). The restless mind. *Psychological Bulletin, 132*(6), 946–958. https://doi.org/10.1037/0033-2909.132.6.946

Smallwood, J., & Schooler, J. W. (2015). The science of mind wandering: Empirically navigating the stream of consciousness. *Annual Review of Psychology, 66*(1), 487–518. https://doi.org/10.1146/annurev-psych-010814-015331

Sripada, C., & Taxali, A. (2020). Structure in the stream of consciousness: Evidence from a verbalized thought protocol and automated text analytic methods. *Consciousness and Cognition, 85*, Article 103007. https://doi.org/10.1016/j.concog.2020.103007

Stawarczyk, D., & D'Argembeau, A. (2019). The dynamics of memory retrieval for internal mentation. *Scientific Reports, 9*(1). https://doi.org/10.1038/s41598-019- 50439-y. Article 13927.

Stawarczyk, D. (2018). Phenomenological properties of mind-wandering and daydreaming: A historical overview and functional correlates. In K. Christoff, & K. C. R. Fox (Eds.), *The Oxford Handbook of Spontaneous Thought: Mind-Wandering, Creativity, and Dreaming* (pp. 193–214). Oxford University Press. https://doi. org/10.1093/oxfordhb/9780190464745.013.18.

Stawarczyk, D., Cassol, H., & D'Argembeau, A. (2013). Phenomenology of future-oriented mind-wandering episodes. *Frontiers in Psychology, 4*. https://doi.org/ 10.3389/fpsyg.2013.00425. Article 225.

Stawarczyk, D., Jeunehomme, O., & D'Argembeau, A. (2018). Differential contributions of default and dorsal attention networks to remembering thoughts and external stimuli from real-life events. *Cerebral Cortex, 28*(11), 4023–4035. https://doi.org/10.1093/cercor/bhx270

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. https://doi.org/10.1016/j.actpsy.2011.01.002

Stawarczyk, D., Majerus, S., Van der Linden, M., & D'Argembeau, A. (2012). Using the daydreaming frequency scale to investigate the relationships between mind-wandering, psychological well-being, and present-moment awareness. *Frontiers in Psychology, 3*. https://doi.org/10.3389/fpsyg.2012.00363. Article 363.

Szpunar, K. K., Addis, D. R., & Schacter, D. L. (2012). *Memory for Emotional Simulations: Remembering a Rosy Future: (520602012-355)*. *American Psychological Association (APA)*.

Uzer, T., Lee, P. J., & Brown, N. R. (2012). On the prevalence of directly retrieved autobiographical memories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(5), 1296–1308. https://doi.org/10.1037/a0028142

Wang, Y.-H. (2016). Reading strategy use and comprehension performance of more successful and less successful readers: A think-aloud study. *Educational Sciences: Theory & Practice, 16*(5). https://doi.org/10.12738/estp.2016.5.0116.

Watkins, E. R. (2008). Constructive and unconstructive repetitive thought. *Psychological Bulletin, 134*(2), 163–206. https://doi.org/10.1037/0033-2909.134.2.163

Watson, J. B. (1920). Is thinking merely the action of language mechanisms? *British Journal of Psychology, 11*(2), 87–104.

Weinstein, Y. (2018). Mind-wandering, how do I measure thee with probes? Let me count the ways. *Behavior Research Methods, 50*(2), 642–661. https://doi.org/ 10.3758/s13428-017-0891-9

Yang, C., & Zhang, L. J. (2023). Think-aloud protocols in second language writing: A mixed-methods study of their reactivity and veridicality (Vol. 34). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-39574-1.