




Exploring self-generated thoughts in a resting state with natural language processing

Hui-Xian Li^{1,2,3} · Bin Lu^{1,2,3} · Xiao Chen^{1,2,3} · Xue-Ying Li^{1,3,4,5} · Francisco Xavier Castellanos^{6,7} · Chao-Gan Yan^{1,2,3,6,8} 

Accepted: 13 September 2021 / Published online: 13 October 2021
© The Psychonomic Society, Inc. 2021

Abstract

The present study seeks to examine individuals' stream of thought in real time. Specifically, we asked participants to speak their thoughts freely out loud during a typical resting-state condition. We first examined the feasibility and reliability of the method and found that the oral reporting method did not significantly change the frequency or content characteristics of self-generated thoughts; moreover, its test–retest reliability was high. Based on methodological feasibility, we combined natural language processing (NLP) with the Bidirectional Encoder Representation from Transformers (BERT) model to directly quantify thought content. We analyzed the divergence of self-generated thought content and expressions of sadness and empirically verified the validity and behavioral significance of the metrics calculated by BERT. Furthermore, we found that reflection and brooding could be differentiated by detecting the divergence of self-generated thought content and expressions of sadness, thus deepening our understanding of rumination and depression and providing a way to distinguish adaptive from maladaptive rumination. Finally, this study provides a new framework to examine self-generated thoughts in a resting state with NLP, extending research on the continuous content of instant self-generated thoughts with applicability to resting-state functional brain imaging.

Keywords Self-generated thoughts · Resting state · Think-aloud method · Natural language processing · Rumination

Attention is increasingly focused on understanding mental activities that are largely independent of perceptual input or external experimental tasks (Callard et al., 2013). The term “self-generated thoughts” captures both the active nature of

these experiences and their independence from perception and ongoing action (Perkins et al., 2015; Smallwood & Schooler, 2015), emphasizing that the content of experience arises from internal changes occurring within an individual rather than from external changes directly prompted by perceptual events occurring in the external environment (Smallwood & Schooler, 2015). “Self-generated thoughts” is an umbrella term that encompasses various types of thought. In the Dynamic Framework of Thought (Christoff et al., 2016; Fox et al., 2018), self-generated thoughts include unconstrained forms of cognition (spontaneous thought: dreaming, mind wandering, and creative thinking) and automatically constrained cognition (rumination and obsessive thought).

Self-generated thought is a ubiquitous (Christoff et al., 2016; Killingsworth & Gilbert, 2010; Smallwood et al., 2004) cognitive activity with complex impacts (Wang et al., 2018). On the one hand, self-generated thought can have many negative effects, such as disrupting task performance (McVay & Kane, 2012) and reducing psychological well-being (Killingsworth & Gilbert, 2010; Smallwood & O'Connor, 2011; Stawarczyk, Majerus, & D'Argembeau, 2013b), and it is associated with various mental illnesses (Marchetti et al., 2016; Perkins et al., 2015). On the other hand, self-generated thought can also be beneficial. It can help

Hui-Xian Li and Bin Lu contributed equally to this work.

✉ Chao-Gan Yan
yangc@psych.ac.cn

- ¹ CAS Key Laboratory of Behavioral Science, Institute of Psychology, 16 Lincui Road, Chaoyang District, Beijing 100101, China
- ² Department of Psychology, University of Chinese Academy of Sciences, Beijing, China
- ³ International Big-Data Center for Depression Research, Institute of Psychology, Chinese Academy of Sciences, Beijing, China
- ⁴ Sino-Danish College, University of Chinese Academy of Sciences, Beijing, China
- ⁵ Sino-Danish Center for Education and Research, Beijing, China
- ⁶ Department of Child and Adolescent Psychiatry, NYU Robert I. Grossman School of Medicine, New York, NY, USA
- ⁷ Nathan Kline Institute for Psychiatric Research, Orangeburg, NY, USA
- ⁸ Magnetic Resonance Imaging Research Center, Institute of Psychology, Chinese Academy of Sciences, Beijing, China

in planning for the future (Szpunar, 2010), reflecting on the past (Stawarczyk et al., 2011), processing personal goals (Smallwood et al., 2013; Smallwood & Schooler, 2006), and consolidating self-related memories (Smallwood et al., 2011). Moreover, it can improve the ability to delay gratification (Smallwood et al., 2013) and facilitate creative problem solving (Baird et al., 2012).

The content regulation hypothesis posits that the beneficial or detrimental effects of self-generated thoughts are influenced by their content (Smallwood & Andrews-Hanna, 2013). Indeed, different kinds of thought content can produce adaptive or maladaptive effects (Marchetti et al., 2016). Self-generated thoughts focusing on the past are more likely to induce stress and sadness (Smallwood & O'Connor, 2011; Stawarczyk, Majerus, & D'Argembeau, 2013b), while those focusing on the future can make individuals more adaptive and resilient (Baird et al., 2011). Ruminative self-generated thoughts increase vulnerability to depression (Nolen-Hoeksema et al., 2008; Watkins, 2008), while overactive, confident, and grandiose self-generated thoughts are associated with mania (Gruber et al., 2008). Thus, detailed and fine-grained research on the content of self-generated thoughts is crucial to understand the complex impact of this phenomenon and its link to mental illness.

Investigators have used a variety of experience sampling (ES) methods to measure self-generated thoughts. Currently, dominant ES approaches include the probe-caught, self-caught, and retrospective methods (Smallwood & Schooler, 2015). The probe-caught method involves randomly interrupting participants while they are completing a task or taking a break and then asking about the contents of their experience with questions such as “Where was your attention focused just before the probe?” The self-caught approach requires participants to actively provide ES reports when they capture their self-generated thoughts. In the retrospective method, subjects report their thoughts after an entire task ends or after a rest period (usually lasting a few minutes); participants are usually asked to fill out questionnaires. These approaches have provided invaluable insights into the frequency, content, and impact of self-generated thoughts. However, they also have substantial limitations. First, ES relies exclusively on individual introspection (Smallwood & Schooler, 2006, 2015). Second, the measurements are all indirect measurements of experience and do not allow us to observe the stream of thought as it flows from one state to another in real time. Although the probe-caught method is closer to real-time measurement than the retrospective method, repeatedly interrupting participants can affect self-generated thoughts, such as by reducing their frequency. Researchers have found that longer intervals between thought probes are associated with a higher frequency of off-task thoughts according to self-reports (Seli et al., 2013; Smallwood et al., 2002). Although the retrospective method does not interrupt

individuals, the fleeting nature of self-generated thoughts (Klinger, 2009) makes it likely that participants will forget short-lived thoughts before they can report these thoughts retrospectively. Researchers have pursued a richer analysis of the content of self-generated thoughts by adding additional probe questions through an approach known as multidimensional experience sampling (MDES) (Konu et al., 2020; Smallwood et al., 2016; Wang et al., 2018). However, adding questions inevitably increases the disadvantages of the probe-caught method. Accordingly, current methods are limited in supporting a richer study of the content of self-generated thoughts.

Language plays an undeniable role in our cognitive lives (Tillas, 2015), and various thinkers have suggested that language is the cornerstone for building thoughts (Levinson, 2003; Prinz, 2004). The language of thought hypothesis (Fodor, 1975; Rescorla, 2019) in linguistics, philosophy of mind, and cognitive science posits that thought occurs in a mental language. In addition, researchers have demonstrated that language is a fundamental element in emotion (Lindquist et al., 2015) and plays an important role in the communication of emotions and emotion-related information (Reisenzein & Junge, 2012). Although mental language, or *mentalese*, is not necessarily a form of natural language, we may be able to measure and analyze thought through language, including the fertile content and emotional characteristics of the stream of thought. The think-aloud verbal protocol in which subjects report their thoughts when performing a primary task has long been a useful method for exploring cognitive processes and the content of consciousness (Ericsson et al., 2006; Ericsson & Simon, 1993; Russo et al., 1989). It has been widely used to study dimensions of thought, including self-generated thoughts (Baird et al., 2011; Holleran & Mehl, 2008; Sripada & Taxali, 2020), which suggests that the think-aloud method is one way to measure the content of self-generated thoughts.

In this paper, we established a new framework for understanding self-generated thoughts in a resting state through the think-aloud protocol. Self-generated thoughts are prevalent in resting-state conditions when no goal-directed tasks are performed. Resting-state fMRI quantifies correlates of spontaneous brain activity, which consumes most brain energy (task-evoked increases in neuronal metabolism are remarkably small (< 5%) (Madsen et al., 1995). We are interested in self-generated thoughts in a resting state, as they may be more representative of the psychological phenomenon of “the stream of consciousness”. However, several issues arise.

First, do verbal reports change the stream of thought? The validity of verbal report methods has been questioned with respect to “non-veridicality”, “completeness”, and “reactiveness” (Russo et al., 1989; Wilson, 1994). The question of “non-veridicality” relates to whether verbal protocols accurately reflect the content of our thoughts, including errors of

omission and commission. The accuracy of verbal reports is related to task requirements, and it is difficult to test the veridicality of a concurrent verbal protocol (Ericsson & Simon, 1980; Wilson, 1994). The question of “completeness” pertains to the data collected using verbal protocols and whether they provide sufficient information to understand an individual’s cognitive processes. Incompleteness of verbal reports is inevitable. On the one hand, due to the limited capacity of short-term memory or working memory, conscious elements may not be fully reported (Ericsson & Simon, 1993). On the other hand, unconscious cognitive processes may be used in task performance (Wilson, 1994). Nevertheless, efforts should be made to obtain a complete verbal report. As Simon suggests, it is always better to retain as much information as possible in any oral report analysis. It is also important to note that participants should be sufficiently trained to become accustomed to the verbalization of their thinking before the verbal report method is used in a study. The central question of the “reactiveness” of verbal protocols relates to whether oral reports interfere with the thinking process. Concerns about reactivity take precedence over concerns about lack of veridicality and completeness because it makes little sense for a test report to be veridical and complete if the words have changed the primary process (Russo et al., 1989). Russo emphasized that testing reactivity is an important methodological goal that usually involves comparing a silent control condition to a concurrent protocol condition. The key to ensuring the validity of verbal reports is to define their applicable task areas and task types.

Therefore, we explored whether verbal reports changed the stream of thought in a resting state, i.e., reactivity. We asked participants to report whatever was currently on their minds in a resting-state condition. This concurrent method did not involve interrupting participants; they simply said aloud what they thought without interpretation. We first explored whether verbal reports were reactive. The primary process in our study was the naturally occurring stream of thought. Therefore, we examined whether verbal reports changed this natural thinking process, that is, whether the frequency and content characteristics of self-generated thoughts changed as a result of verbal reports. We incorporated classic methods for comparison. The first was the uninterrupted self-caught method, which simply required participants to press the “H” button when they experienced a self-generated thought to minimize the impact on subjects. The retrospective method was used to ask participants to evaluate the content characteristics of their self-generated thoughts. The first nonverbal report stage was used as a comparable silent control condition. We compared the difference between the frequency of thoughts and the retrospective content assessment in the first nonverbal report stage and the verbal report stage. Further, all participants completed two identical sessions within less than 1 week to test the test-retest reliability of the verbal report method.

Second, after collecting self-reports, most researchers have conducted manual follow-up analyses. For example, judges have been instructed to rate participants’ personality from stream-of-consciousness essays (Holleran & Mehl, 2008); to rate negativity of tone, problem focus, and self-criticism scores based on collected thoughts (Lyubomirsky et al., 1999); and to classify the levels of abstraction of thought content from transcribed think-out-loud reports (D’Argembeau & Mathy, 2011). Such analyses of the content of collected thoughts have mainly been performed through manual coding, which is labor-intensive. Fortunately, a set of distributional semantic models (DSMs) can be used to decode verbally reported text, a less labor-intensive approach that facilitates the analysis of large numbers of verbal reports. Furthermore, the use of a model may provide more consistent and reproducible ratings than human raters. The reproducibility of the output indexes is also higher and the whole experiment can be easily replicated by other researchers. DSMs, such as Latent Semantic Analysis (Landauer & Dumais, 1997), calculate the context-based semantic similarity between words or terms to infer the relationships between terms and the content of texts. DSM studies have usually used word-based semantic models to analyze verbal reports. For example, Faber and D’Mello used latent semantic analysis (LSA) to analyze the content of self-reported task-unrelated thoughts during a primary task (Faber & D’Mello, 2018), suggesting that natural language processing (NLP) has potential value for quantitative analysis of thought content characteristics. However, sentence-based models, such as deep-learning algorithms trained on large-scale data (e.g., Wikipedia), showed superior performance than word-based models (Bhatia et al., 2019) and have become prominent in linguistical analysis (see Bhatia & Richie, 2020; Singh et al., 2020; Zhao et al., 2019 for examples). Of particular note in this regard is the natural language pre-trained Bidirectional Encoder Representation from Transformers (BERT) model (Devlin et al., 2019) released by the Google team in 2018. The vectors in BERT embody richer and more accurate semantic information than the vectors in LSA. The reasons for this are as follows. (1) The vectors in LSA represent words, while the vectors in BERT represent sentences. (2) The vectors in LSA are created by converting a corpus into a term-document matrix by counting the number of times a term_i is present in a document_j and decomposing the matrix into a low-dimensional approximation. However, the vectors in BERT are pre-trained by masking a word in a sentence, predicting the masked words, and predicting the next sentence following the present sentence. By achieving these two semantic inference tasks, the pre-trained model can learn both word-level and sentence-level semantic representations. (3) The LSA model was trained on the Touchstone Applied Science Associates (TASA) corpus, which comprises 60,527 samples from 6333 textbooks. However, the BERT (Chinese version) was trained on the entire Chinese content of

Wikipedia, which contains more than 1,152,136 articles. The enormous amount of native language training data can provide better representations of the verbal reports of participants. BERT is the pre-eminent NLP model in many benchmark comparisons (Devlin et al., 2019), especially in the sentiment classification task (SST-2, Stanford Sentiment Treebank) and semantic similarity discrimination (STS-B, the Semantic Textual Similarity Benchmark), which are closely related to probing self-generated thoughts. In general, BERT is the next-generation technology relative to LSA. The BERT model can achieve both word-level and sentence-level sophisticated semantic representations.

We combine verbal reports with NLP to quantify the content of thought using the BERT model and examine whether the indicators calculated by BERT are behaviorally relevant. Rumination is a stable trait-like response mode triggered by negative events and characterized by repetitive and negative content (Watkins, 2008). Researchers consider rumination to represent a sticky and negatively valenced type of self-generated thought (Christoff et al., 2016; DuPre & Spreng, 2018; Smallwood & Schooler, 2006). Therefore, we developed quantitative metrics of the characteristics of self-generated thought content. Specifically, we used the BERT model to detect the divergence of thoughts (repetitive feature) and expressions of sadness (negatively valenced feature) in thought content. We hypothesized that BERT-generated metrics would be associated with individual differences in the rumination trait and that analysis of divergence and expressions of sadness in self-generated thought content would bear on current views of the relationship between self-generated thoughts and rumination. Of note, researchers have stressed that rumination contains different psychological constructs: reflection and brooding (Treyner et al., 2003). Reflection can be adaptive; it can increase self-knowledge and facilitate problem solving (Watkins, 2008). It is positively related to concurrent depression but negatively related to longitudinal depression (Treyner et al., 2003). In contrast, brooding is viewed as moody pondering and as a maladaptive form of rumination. It involves passively focusing on issues and symptoms and is associated with both concurrent and longitudinal depression (Siegle et al., 2004; Treyner et al., 2003). It would be valuable to differentiate the reflection and brooding components of rumination (Treyner et al., 2003) to help address the question of whether rumination can be adaptive and maladaptive (Barbic et al., 2014; Nolen-Hoeksema et al., 2008; Watkins, 2008). Therefore, these two content characteristic metrics have the potential to distinguish between adaptive and maladaptive rumination.

In summary, we propose a new framework for understanding self-generated thoughts in a resting state by considering the think-aloud protocol, and we attempt to address the following questions. First, is it feasible to have participants report their thoughts generated in a resting state? Specifically, we

examine whether verbal reports change the frequency and content characteristics of self-generated thoughts in a resting state. Second, can the representations generated by DSMs reflect the content of the resting-state self-generated thoughts of people with high levels of negative and repetitive rumination?

Methods

Participants

Forty-six participants recruited by Internet advertisements completed two identical experiments and all self-report questionnaires. Individuals who reported psychiatric or neurological disorders, use of psychotropic medications, or any history of substance or alcohol abuse were excluded. We excluded two participants with missing button press data in the first session and four participants whose verbal reports contained fewer than five thoughts. This yielded a final sample size of 40 (19 males and 21 females; age: 18–33 years, 23.0 ± 3.4 years). All participants gave informed written consent. The protocol was approved by the Institutional Review Board of the Institute of Psychology, Chinese Academy of Sciences.

Measures

The Daydream Frequency Questionnaire (Singer & Antrobus, 1972) was used to assess the frequency of daily daydreaming. It is a subscale of the Imaginal Processes Inventory Questionnaire and consists of 12 items. Participants rated the frequency with which they daydreamed in daily life on a five-point Likert scale with items such as “What is the probability that my mind will wander in my free time?” The five response options for this item include never, rarely, sometimes, often, and always. For “How much of my total conscious waking state time is spent wandering or daydreaming?”, response options are none, less than 10%, at least 10%, at least 25%, and at least 50%. Trait rumination was evaluated with the Ruminative Responses Scale (Treyner et al., 2003), which includes 22 items rated from 1 (*almost never*) to 4 (*almost always*). The items describe responses regarding feelings of sadness and depression. The scale consists of three factors: Reflection, Brooding and Depression-related. An example Reflection item is “Write down what you are thinking about and analyze it.” An example Brooding item is “Why do I have problems other people don’t have?” An example “Depression-related” item is “Think about all your shortcomings, failings, faults, mistakes.” All participants completed all questionnaires. The mean Daydream Frequency Questionnaire score was 42.55 (SD = 8.88), and the mean Ruminative Responses Scale score was 47.88 (SD = 12.42). The distribution of the scores is shown in Fig. S8.

Experimental procedure

All participants completed the task in the same quiet, isolated room. After entering the laboratory, participants were asked to sit comfortably in front of a computer for five minutes to relax and become familiar with their surroundings. Then, we briefly introduced the concept of self-generated thoughts and asked a few questions to deepen their understanding of this phenomenon, such as “Do you pay attention to thoughts or images that come to your mind unrelated to what is going on when you are working or taking a class? How much attention do you pay?” “Have you ever experienced a situation in which you were reminded by others that you seem to have been mind wandering? Does this happen a lot?” All introduction and interview information is presented in Supplementary Material Table S6.

We then informed participants of the overall experimental requirements and started the formal experiment. The experimental task included three 10-min stages. In the first stage, participants were required to press the “H” button whenever a new thought (idea) that was different from their former thought came to mind. In the second stage, participants were asked to say whatever was currently on their minds. The third stage was a combination of the first two. In addition to saying their current thought, participants needed to press the “H” button whenever they realized they were switching to a new thought. It should be noted that the participant pressed the button when a thought was first generated, and each subsequent press of the button needed to be different from the previous thought. Differences and switches in thought need to be judged by the participants based on topic similarity. Button presses reflect the number of self-generated thought episodes (Seli et al., 2017; Stawarczyk, Cassol, & D’Argembeau, 2013a). As our experiment asked participants to self-catch their thoughts, it reflects self-generated thoughts with meta-awareness. At the end of each stage, participants rated the characteristics of their thought content on a scale from 1 to 9 points. They were asked a total of eight questions that included four dimensions, i.e., the temporal, social, emotional valence, and mental experience form of their self-generated thoughts (Gorgolewski et al., 2014)(Table S1). Each participant completed two identical sessions within less than 1 week (Time 1 and Time 2) (1~7 days, 2.6 ± 1.8 days) to assess the reliability of the resting-state think-aloud method.

Our first purpose was to explore whether verbal reporting was feasible, and our first concern was whether verbal reporting might change the content or frequency of self-generated thoughts. In addition, to prevent the verbal reporting from affecting participants, all participants were asked to first complete a resting state without providing any verbal report,

with the only requirement that they press the “H” button (Stage 1). This stage was compared to the last two stages as an accuracy index.

After the first stage, the experimental content and requirements of the last two stages were explained in detail. We informed participants that the last two reporting periods would be audio-recorded throughout but we would not concurrently listen to their reports; that the audio-recording file names would not contain any personally identifiable information; and that we would use the text only for scientific research and would not disclose any of their personal information. Then, participants performed a 5-min verbal report exercise to adapt to the experiment, and they proceeded to the next two formal stages. Nearly all participants expressed their willingness to report their thoughts and said that the experience of talking to themselves was familiar. One participant expressed difficulty reporting his ideas during the practice session, so he did not participate in the experiment. The last stage was a combination of the first two, and it was designed to improve the efficiency of quantifying thought switches by using participants’ button presses so that we could attempt to automate the detection of thought switches.

To ensure a natural and quiet resting state and to avoid embarrassment or discomfort when participants reported their thoughts, the experimenter left the testing room and entered an adjacent experimental room after each experimental stage requirement was explained. Participants controlled the start of every stage by pressing a button. Once they finished a stage, they informed the experimenter sitting next door (by opening the door of the testing room). The experimenter then proceeded to introduce the next stage.

Preprocessing and analysis of verbal reports

We converted the verbal reports into text with the speech-to-text platform iFLYTEK (<https://www.iflyrec.com>) with manual supervision. Only a few participants used filler words such as “um” and “ah” to indicate a pause. To avoid the influence of such words, various stop words were discarded using iFLYTEK. Based on the text, we manually labeled thought switches according to the similarity of the topics and calculated the number of self-generated thought episodes. For example, “When will my hair grow long?” followed by “I’m going to talk to my sister later about how to use the water boiler” was marked as a thought switch, which consisted of two different self-generated thought episodes. It is worth noting that manually coded thought switches have the limitation that we cannot discern connections in participants’ thought stream based on their prior experience. The total number of self-generated thought episodes based on manual text labeling was counted and compared with the number of button presses per participant in the first stage.

Divergence in thought content in the first and second 5-min segments

We then used NLP to process the text using the BERT model (Devlin et al., 2019) and bert-as-service (<https://github.com/hanxiao/bert-as-service>) to map sentences contained in each self-generated thought episode into 768-dimensional fixed-length vector representations. Here, the pre-trained BERT model was Chinese_L-12_H-768_A-12 provided by Google. We focused on the divergence in thought content over time. We divided each 10-min stage into two 5-min segments. For each segment, we calculated the sum vector of all self-generated thought episode vectors (768-dimensional token vectors from the final layer of BERT) to represent the content of the segment. We measured the divergence of thought content across the two segments by computing the inverse cosine value between the two sum vectors.

$$\text{acos}(\text{dot}(A, B) / (\text{norm}(A) * \text{norm}(B)))$$

The above MATLAB formula calculated the angle expressed in radians between two vectors. The smaller this indicator is, the smaller the angle between the two vectors, and the more similar the text content of the two segments. We call this indicator the radian difference. The detailed calculation process is shown in Fig. 1 (section 1) below.

Detection of the level of sadness in self-generated thought content

To detect expressions of sadness in verbal reports, we trained a BERT-based deep-learning emotion classifier on 24,400 Weibo (Chinese version of Twitter) sentences with emotion annotations. This dataset was downloaded from the “The 3rd CCF Conference on Natural Language Processing & Chinese Computing” website (http://tcci.ccf.org.cn/conference/2014/pages/page04_ans.html). These raw Weibo sentences were first preprocessed to be more similar to our texts by removing the Weibo emoji, the “@” symbol and the following username, the “#” symbol, and the “Reply” word, which are part of the Weibo format. Then, these preprocessed sentences were imported into the BERT model and converted to 768-dimensional fixed-length vectors. Three full-connect layers (containing 1024, 256 and 64 neurons, respectively, with ‘relu’ activation), one dropout layer (dropout rate = 0.5), and one softmax layer were concatenated to the BERT model. The last 3 layers of the BERT model (Chinese_L-12_H-768_A-12) used in this section were fine-tuned to achieve better performance. We used categorical cross-entropy as loss function and used default adam optimizer in Keras for training. The batch size was set to 32. The emotion classifier predicts the most predominant emotion (from among anger, disgust, fear, happiness, liking, none, sadness, or surprise) of each sentence

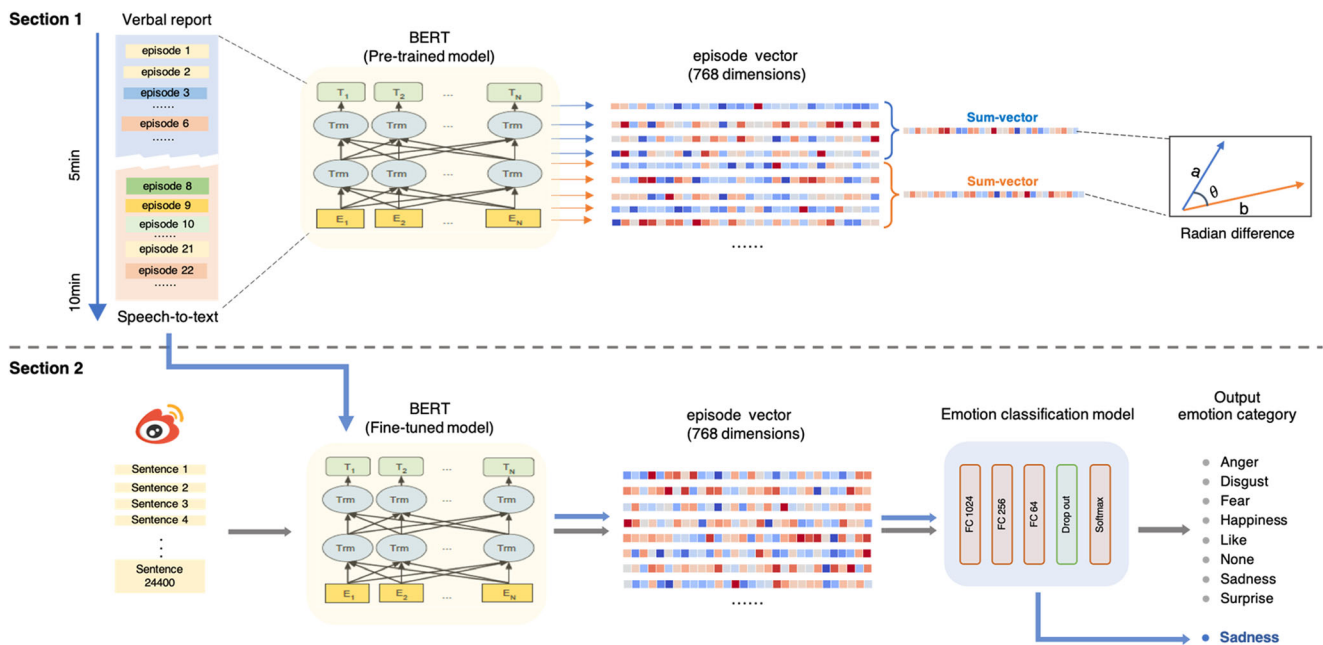


Fig. 1 Flow chart of verbal report analysis: an example of one stage. Section 1: The sentences contained in each self-generated thought episode were represented as a 768-dimensional vector using the BERT model. Then, the radian difference between two sum vectors was calculated to detect the overall divergence of thought content between the two 5-min

segments. Section 2: We trained an emotion recognition model based on 24,400 sentences, and the model achieved 78% accuracy in the eight-class emotion recognition task. Then, the verbal reports were entered into this classifier to quantify the tendency to report sad thoughts

and provides the probability of each emotion. A total of 20,000 Weibo sentences were randomly allocated into the training sample, and the other 4400 were allocated into the testing sample. After 50 epochs of training, the model achieved 78% accuracy in the eight-class emotion recognition task. All performance metrics were calculated based on the testing sample. More information about precision, recall, and F1-scores as performance indices is shown in Table 1. The confusion matrix of the emotion classifier is shown in Table 2. The distribution of true labels in the training and testing samples is provided in supplemental materials Fig. S6. The sentences contained in each self-generated thought episode verbal report was entered into this classifier to obtain the sadness emotion probability of a self-generated thought episode. The average sadness emotion probability of all self-generated thought episodes of a verbal report was used as the degree of NLP expression of sadness for each participant (Fig. 1 section 2).

Correlation analysis

We calculated the correlation between the total score on the Daydream Frequency Questionnaire and the frequency of self-generated thought episodes in the experiment. We also calculated the relationship between two metrics calculated by NLP (content divergence and expressions of sadness in the verbal reports) and rumination trait indices (total score on the Ruminative Responses Scale and three sub-factors of rumination). Rumination is characterized by repetitive negative thoughts; therefore, we hypothesized that a higher rumination trait score would be associated with smaller divergence and greater sadness in the content of self-generated thoughts.

Results

Feasibility of the resting-state think-aloud method

Our first goal was to examine the feasibility of using the resting-state think-aloud method. Would the verbal report change the frequency or content of self-generated thoughts? The first stage was compared to the verbal report stage to assess accuracy. To quantify the frequency of self-generated thoughts in the first stage, we used the number of times participants pressed the “H” button. In the second and third stages, we used the frequency of manually labeled self-generated thoughts. Repeated-measures ANOVA showed no significant differences in the frequency of self-generated thoughts among the three stages (Time 1: $F(1.46, 56.77) = 0.455$, $MSE = 20.347$, $p = .575$ ($\epsilon = 0.728$, Greenhouse–Geisser correction), partial $\eta^2 = 0.012$; Time 2: $F(1.17, 45.64) = 0.782$, $MSE = 66.411$, $p = .40$ ($\epsilon = 0.585$, Greenhouse–Geisser correction), partial $\eta^2 = 0.02$) (Fig. 2a). The frequency of self-generated thoughts was

Table 1 The performance of the emotion classifier on the testing sample

Emotion	Precision	Recall	F1-score	No. of samples
Anger	0.748	0.761	0.755	410
Disgust	0.754	0.772	0.763	723
Fear	0.708	0.708	0.708	72
Happiness	0.846	0.734	0.786	704
Like	0.823	0.854	0.838	1163
None	0.752	0.786	0.768	589
Sadness	0.790	0.799	0.795	538
Surprise	0.711	0.687	0.699	201

significantly correlated across the three stages; that is, the consistency was good (stage 1 _ stage 2: $r(40) = 0.542$, $p < .001$, 95% CI = [0.28, 0.73]; stage 1 _ stage 3: $r(40) = 0.502$, $p = .001$, 95% CI = [0.23, 0.70]; stage 2 _ stage 3: $r(40) = 0.828$, $p < .001$, 95% CI = [0.70, 0.90]) (Fig. S2). These same analyses conducted at Time 2 revealed similar patterns (Fig. S1). Furthermore, frequency in the first and second stages was related to daydream frequency per questionnaire (Fig. 2b, column Time 1: stage 1, $r(40) = 0.492$, $p = .001$, 95% CI = [0.21, 0.70]; stage 2: $r(40) = 0.354$, $p = .025$, 95% CI = [0.05, 0.60]; Fig. 2b, column Time 2: stage 1, $r(40) = 0.316$, $p = .047$, 95% CI = [0.005, 0.57]; stage 2: $r(40) = 0.431$, $p = .006$, 95% CI = [0.14, 0.65]).

In the third stage, participants were asked to speak their thoughts out loud and press the “H” button when they realized their thoughts had changed. Button pressing frequency in the third stage of Time 1 was significantly lower than the labeled frequency from oral reports ($t(39) = -4.12$, $p < .001$, $d = -0.65$, 95% CI = [-7.23, -2.47]). Moreover, it was significantly lower than the frequency of button pressing in the first stage ($t(39) = -2.86$, $p = .007$, $d = -0.45$, 95% CI = [-6.23, -1.07]) and the labeled frequency in the second stage ($t(39) = -3.36$, $p = .002$, $d = -0.53$, 95% CI = [-7.09, -1.76]) (Fig. 2a1). Most participants reported that they often forgot to press the button in the third stage despite having expressed their thoughts out loud. Therefore, in the second identical experimental session (Time 2), we emphasized that the task also included pressing the button in the third stage. At Time 2, the difference between button pressing and labeled frequencies in the third stage no longer differed significantly ($t(39) = -1.87$, $p = .068$, $d = -0.30$, 95% CI = [-7.12, 0.27]), but the button pressing frequency in the third stage was still significantly lower than that in the first stage ($t(39) = -4.07$, $p < .001$, $d = -0.64$, 95% CI = [-8.05, -2.70]) and the labeled frequency in the second stage ($t(39) = -2.16$, $p = .037$, $d = -0.34$, 95% CI = [-8.04, -0.27]) (Fig. 2a2). Additionally, the labeled frequency in the third stage of the two sessions (Time

Table 2 The confusion matrix of the emotion classifier on the testing sample

		Predicted							
		Anger	Disgust	Fear	Happiness	Like	None	Sadness	Surprise
Actual	Anger	312	59	1	5	7	10	8	8
	Disgust	39	558	3	8	32	36	28	19
	Fear	2	5	51	1	0	4	8	1
	Happiness	9	20	0	517	103	30	17	8
	Like	12	20	0	57	993	44	25	12
	None	10	33	8	10	39	463	21	5
	Sadness	17	30	5	11	22	20	430	3
	Surprise	16	15	4	2	10	9	7	138

1 and Time 2) was not significantly correlated with the Daydream Frequency Questionnaire score (Fig. 2b, row Stage 3, Time 1: $r(40) = 0.295$, $p = .065$, 95% CI = $[-0.019, 0.56]$; Time 2: $r(40) = 0.278$, $p = .082$, 95% CI = $[-0.04, 0.54]$). Although the button pressing frequency of the third stage at Time 1 was significantly positively related to the Daydream Frequency Questionnaire score ($r(40) = 0.405$, $p = .010$, 95% CI = $[0.11, 0.64]$), it was not significantly correlated at Time 2 ($r(40) = 0.229$, $p = .156$, 95% CI = $[-0.09, 0.50]$)

(Fig. 2b). Accordingly, we inferred that asking participants to both speak out loud and press the button to indicate thought switching (i.e., Stage 3) may have been too burdensome. Therefore, the subsequent NLP analysis did not include the data from this stage.

In each stage, participants were asked eight questions comprising four dimensions of self-generated thoughts, i.e., temporal (past vs. future), social (self vs. other), emotional valence (positive vs. negative) and mental experience form

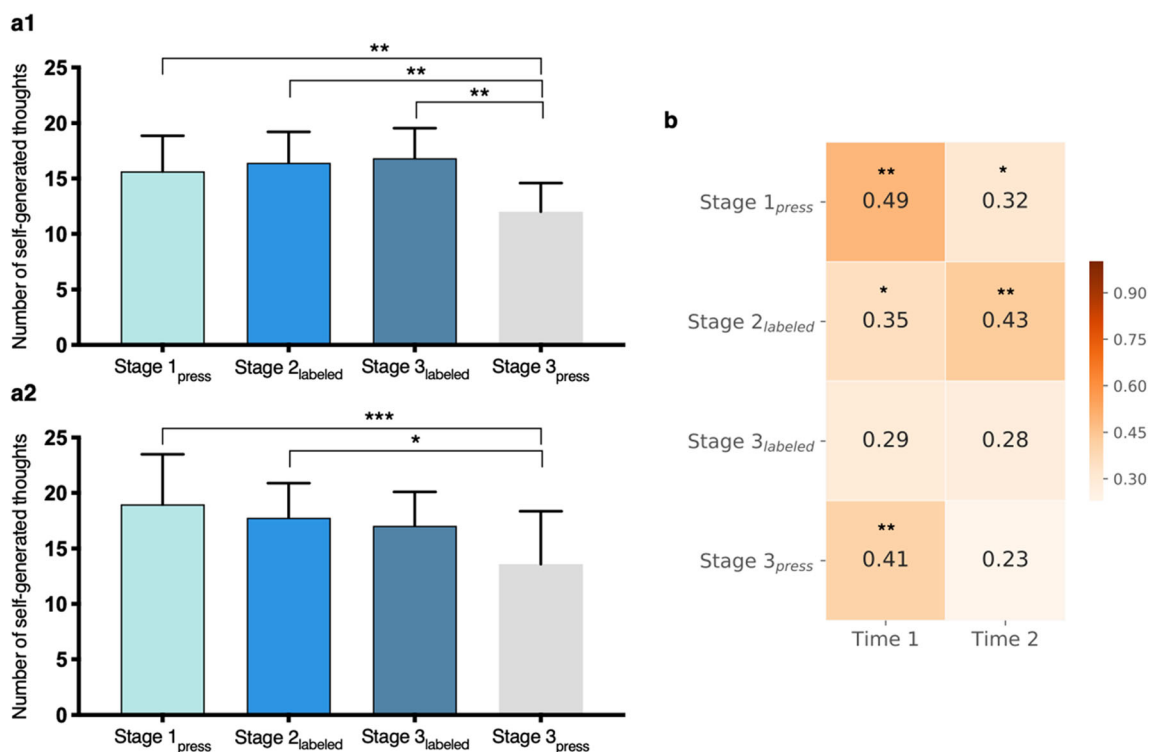


Fig. 2 The frequency of self-generated thoughts during simultaneous verbal reporting and button pressing. **a** Differences between the frequency of button pressing in the third stage and in other stages. Two-tailed paired t tests; data are means with 95% CI. **b** The Pearson correlation coefficients between the frequency of self-generated thoughts and

daydream frequency (two-tailed). “Daydream Frequency” is the total score on the Daydream Frequency Questionnaire. The subscript “press” indicates the frequency of button pressing. The subscript “labeled” indicates the number of manually labeled occurrences. * $p < .05$; ** $p < .01$; *** $p < .001$

(picture vs. language). We did not find any significant differences in the eight content characteristics across the three stages (Table S2, Fig. 3a). Furthermore, we compared the differences in the four dimensions of thought content at each stage (Fig. 3b1, b2, b3, Table S3), and the results were stable across the three stages. Participants thought significantly more about themselves than others and reported more positive than negative content. The content results measured at Time 2 were similar to those measured at Time 1 (Fig. S3).

Reliability of the resting-state think-aloud method

The intraclass correlation coefficient (ICC), which varies between 0 and 1, is a widely used coefficient for measuring test–retest reliability (Bartko, 1966). Liandis and Koch suggested that an ICC greater than 0.80 indicates that reliability is good, 0.61 to 0.80 moderate, 0.41 to 0.60 fair, 0.11 to 0.40 slight, and less than 0.1 negligible (Shrout, 1998). In our experiment, all participants completed two identical experimental measurements within 1 week (Time 1 and Time 2), and we used

the ICC to examine the test–retest reliability of the frequency and content of self-generated thoughts (Table 3). The test–retest reliabilities (ICC) of frequency in stages one to three were 0.765, 0.853, and 0.870, respectively. In the third stage, the frequency measurement was the number of manually labeled thoughts. The test–retest reliability of content characteristics in the three stages was also quantified as the ICC of the answers to the eight questions after each stage (Table 3).

Feasibility and reliability of quantitative metrics based on NLP analysis

Based on the above results, we concluded that oral reporting did not significantly change the frequency or content characteristics of individual self-generated thoughts. Additionally, the verbal report method had moderate-to-good test–retest reliability. Therefore, we proceeded to quantify natural stream of thought content characteristics through NLP (details provided in the Methods section). We focused on thought divergence and expressions of sadness and used the second stage of

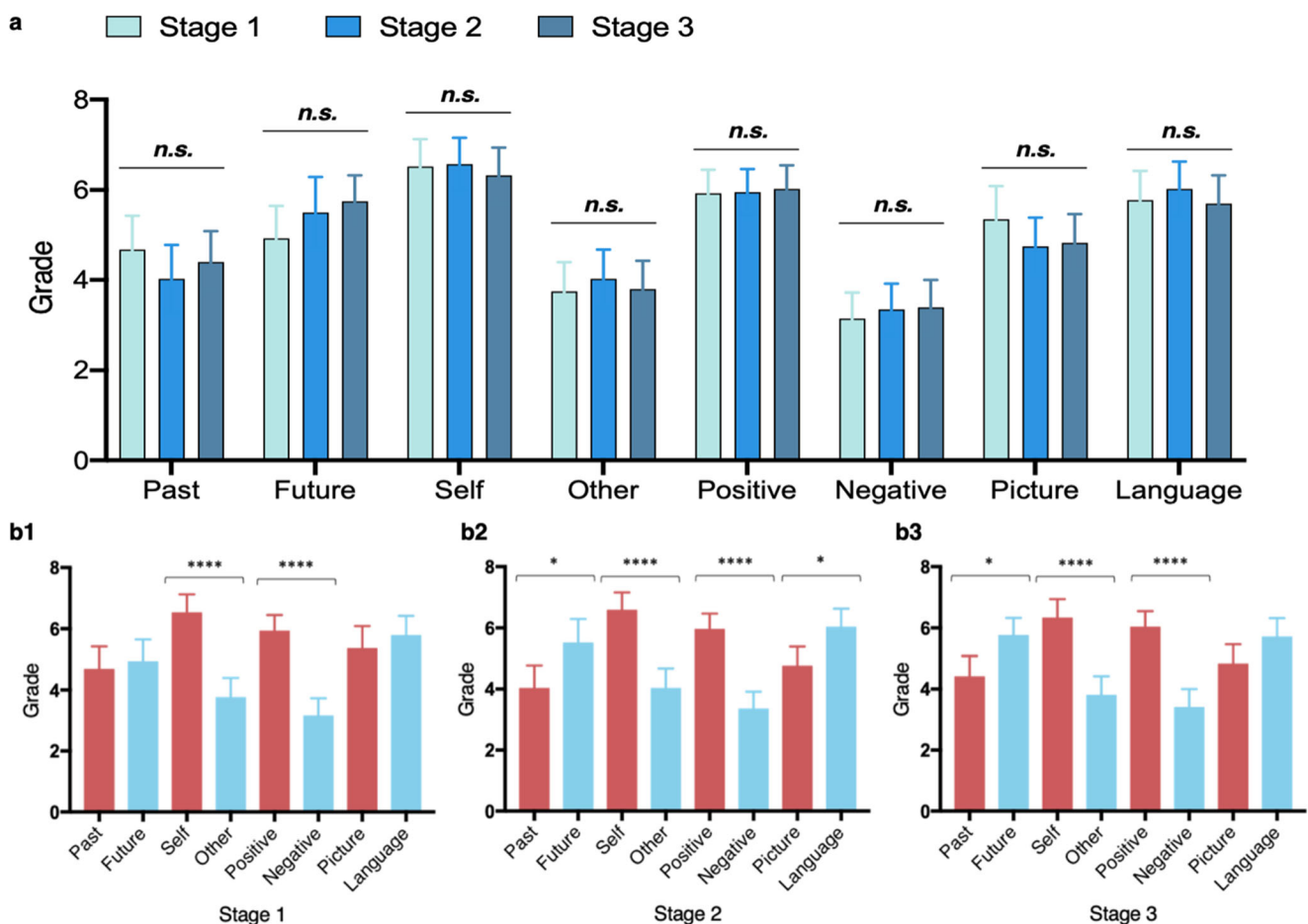


Fig. 3 Content characteristics of self-generated thoughts. **a** Differences among the three stages of each content assessment question. Comparisons of differences (two-tailed paired *t* tests) across four dimensions (temporal: past vs. future; social: self vs. other; emotional valence: positive vs.

negative; mental experience: picture vs. language) of self-generated thought content in the first stage (**b1**), the second stage (**b2**), and the third stage (**b3**). Data are means with 95% CI. * $p < .05$; **** $p < .0001$

Table 3 Test–retest reliability

	Stage1			Stage2			Stage3		
	ICC	<i>P</i>	95% CI	ICC	<i>P</i>	95% CI	ICC	<i>P</i>	95% CI
Frequency	0.765	< .001	0.55, 0.88	0.853	< .001	0.72, 0.92	0.870	< .001	0.75, 0.93
Past	0.499	.017	0.05, 0.74	0.412	.051	-0.11, 0.69	0.396	.060	-0.14, 0.68
Future	0.260	.176	-0.40, 0.61	0.578	.004	0.20, 0.78	0.611	.002	0.27, 0.79
Self	0.697	< .001	0.43, 0.84	0.506	.015	0.07, 0.74	0.413	.050	-0.11, 0.69
Other	0.347	.094	-0.24, 0.66	0.645	.001	0.33, 0.81	0.556	.006	0.16, 0.77
Positive	0.447	.034	-0.05, 0.71	0.747	< .001	0.52, 0.87	0.761	< .001	0.55, 0.87
Negative	0.499	.017	0.05, 0.74	0.752	< .001	0.53, 0.87	0.553	.007	0.16, 0.76
Picture	0.837	< .001	0.69, 0.91	0.734	< .001	0.50, 0.86	0.863	< .001	0.74, 0.93
Language	0.763	< .001	0.55, 0.88	0.788	< .001	0.60, 0.89	0.841	< .001	0.70, 0.92

Note: The test–retest reliability of three experimental methods (three stages) includes the frequency and content of self-generated thoughts in a resting state. SPSS version 25 was used. Mode: two-way mixed; type: consistency

the oral reporting results because that stage was less burdensome than the third stage and had moderate-to-good test–retest reliability. The ICC of the radian difference between the first and second 5-min segments was 0.709 ($p < .001$, 95% CI = [0.45, 0.85]). The ICC of expressions of sadness in the verbal reports detected by the pre-trained emotion recognition model was 0.667 ($p < .001$, 95% CI = [0.37, 0.82]) and was significantly positively correlated with the negative events tendency retrospectively reported by individuals (Fig. 4a, $r(40) = 0.354$, $p = .025$, 95% CI = [0.05, 0.60]; Fig. 4b, $r(40) = 0.341$, $p = .031$, 95% CI = [0.03, 0.59]).

Based on two thought content indicators calculated by NLP, we explored the relationship between the characteristics of self-generated thought content and the rumination trait. First, we found that the index of within-session divergence

in self-generated thought content was significantly negatively correlated with rumination (Fig. 5a1, $r(40) = -0.436$, $p = .005$, 95% CI = [-0.66, -0.14]), indicating that people with higher rumination trait values differ less in self-generated thought content within a session. In addition, to control for possible effects of report length and total number of words on content divergence, we calculated the corresponding partial correlations between content divergence and rumination trait indices (“Partial correlations” in the Supplemental Material). The results remained almost the same (Table S4). In addition, the tendency to report sad thoughts in verbal reports was significantly positively correlated with the rumination trait (Fig. 5b1, $r(40) = 0.348$, $p = .028$, 95% CI = [0.04, 0.60]), suggesting that people with higher rumination trait levels have higher levels of sadness in the content of their self-generated

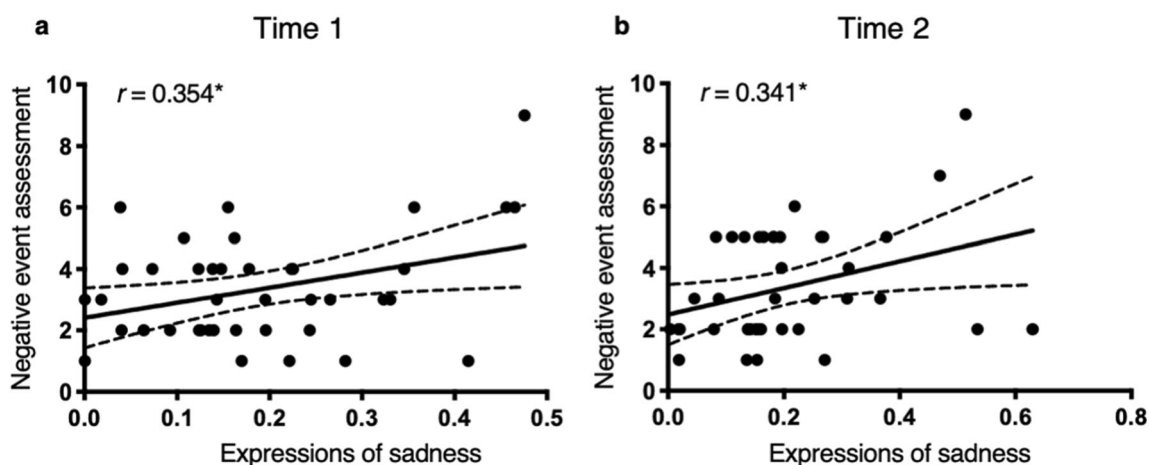


Fig. 4 Pearson correlations between expressions of sadness in verbal reports calculated by the BERT model and negative events assessed retrospectively. **a** The second stage of Time 1. **b** The second stage of Time 2. The *X*-axis is the expressions of sadness detected by the pre-

trained emotion recognition model. The *Y*-axis is the negative events assessment, a retrospective question that participants rated from 1 to 9 points. Two-tailed. * $p < .05$

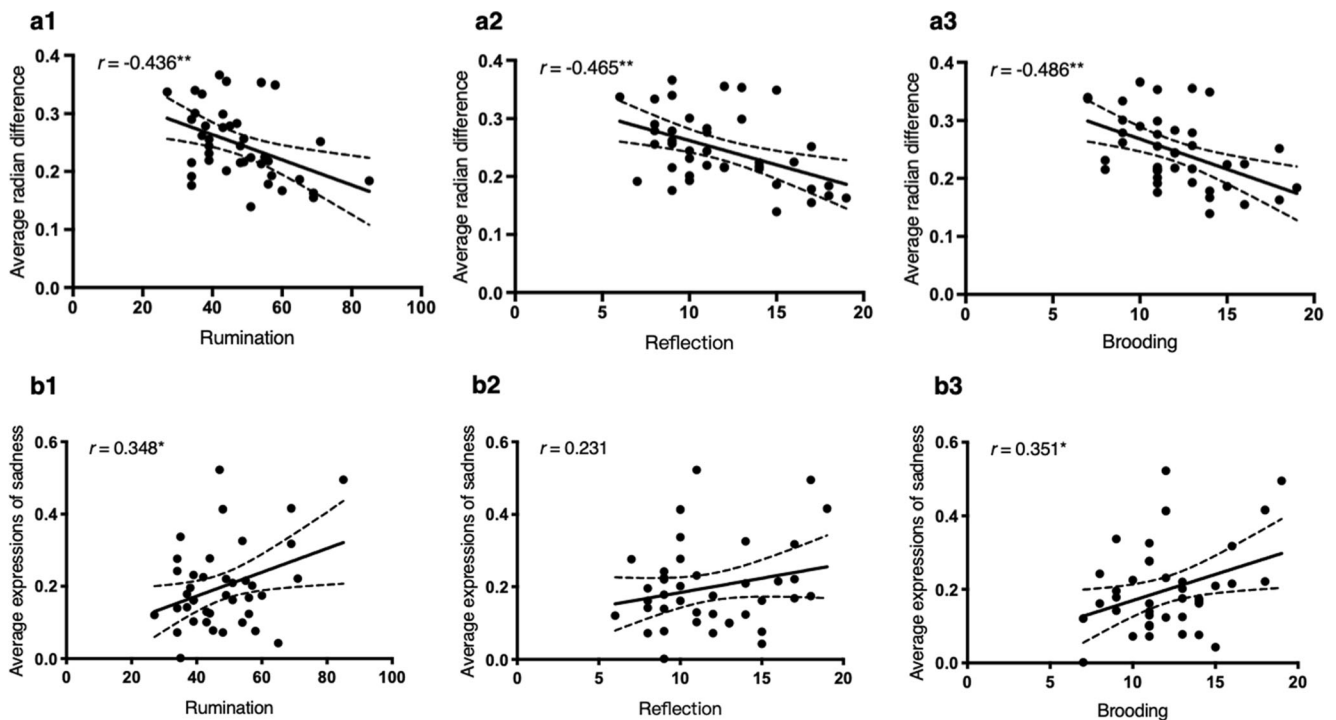


Fig. 5 Pearson correlations between the characteristics of self-generated thoughts calculated by NLP and rumination traits. **a** Correlations between the index of divergence in self-generated thought content and total rumination traits (**a1**), reflection sub-factor (**a2**) and brooding sub-factor (**a3**). The Y-axis is the average radian difference between two consecutive 5-min measurements in the second stage. **b** Correlations between the index of expressions of sadness in verbal reports and total rumination traits (**b1**),

reflection sub-factor (**b2**), and brooding sub-factor (**b3**). The Y-axis shows the average expressions of sadness in verbal reports detected by the pre-trained emotion recognition model in the second stage of two sessions (T1 and T2). “Rumination” is the total score on the Ruminative Responses Scale. “Reflection” and “Brooding” are two factors of the Ruminative Responses Scale. Two-tailed. * $p < 0.05$; ** $p < 0.01$

thoughts. We further examined the performance of two indicators of self-generated thought content characteristics on two different forms of rumination: reflection and brooding. Reflection and brooding were consistent in relation to the divergence of thought content within a session (Fig. 5a2, $r(40) = -0.465$, $p = .003$, 95% CI = $[-0.68, -0.18]$; Fig. 5a3, $r(40) = -0.486$, $p = .002$, 95% CI = $[-0.70, -0.21]$). The partial correlation results were consistent with the full correlation results (Table S4). We also analyzed the divergence in thought content over time, which was the radian difference between the two test–retest sessions. The divergence between Time 1 and Time 2 was also significantly negatively related to rumination, reflection, and brooding (Fig. S4). After we controlled for report length and total number of words, only brooding remained significantly negatively correlated with divergence in thought content over time (Table S5, brooding: $r(36) = -0.376$, $p = .020$). In addition, reflection and brooding differed in their relationships with expressions of sadness. Specifically, the tendency to report sad thoughts was significantly related only to brooding (Fig. 5b3, $r(40) = 0.351$, $p = .027$, 95% CI = $[0.04, 0.60]$), while the correlation with reflection was low and not significant (Fig. 5b2, $r(40) = 0.231$, $p = .151$, 95% CI = $[-0.09, 0.51]$).

Automated NLP analysis attempt

Although many participants performed poorly in pressing the button in the third stage, we explored the possibility of automatically detecting thought switches in the third stage. Participants whose button pressing frequencies were fewer than five were excluded. In addition, two participants with extremely high button press values, 49 and 87 presses, were excluded. The final sample size was 31 (15 males and 16 females). We found that the button pressing frequencies in the third stage of both sessions were significantly positively related to Daydream Frequency Questionnaire scores (Time 1, $r(31) = 0.416$, $p = .020$, 95% CI = $[0.67, 0.07]$; Time 2, $r(31) = 0.377$, $p = .036$, 95% CI = $[0.65, 0.03]$). Participants’ button pressing frequency and the frequency of manually labeled self-generated thought episodes in the third stage differed. Specifically, participants’ button pressing frequencies were less than the frequency of manually labeled self-generated thought episodes (Fig. 2a). Most participants reported that they often forgot to press the button in the third stage despite having expressed their thoughts out loud. So we believed that button presses were missed in the third stage. We examined manually labeled thought switches in the third stage according

to when participants pressed the button. Specifically, if a participant did not press the button at an obvious topic switch we would treat it as a thought switch and labeled it as a missed pressing. If we did not see an obvious topic switch but the participant pressed the button, we respected the participant's judgment of topic switch and labeled it in accordance with the participant's button press. The results of a reanalysis with manual correction of missed button presses showed that the labeled frequency in the third stage of the two sessions was also significantly correlated with Daydream Frequency Questionnaire scores (Time 1, $r(31) = 0.364$, $p = .044$, 95% CI = [0.64, 0.01]; Time 2, $r(31) = 0.420$, $p = .019$, 95% CI = [0.67, 0.08]). Furthermore, we calculated the within-session divergence index of thought content based on participants' button pressing and manually labeled thought switches. The correlation between the two ways of calculating thought switches reached 0.923 ($p < .001$, 95% CI = [0.96, 0.85]). The button pressing frequency was still significantly lower than the labeled frequency (Time 1: $t(30) = -4.324$, $p < .001$, $d = -0.78$, 95% CI = [-3.09, -1.11]; Time 2: $t(30) = -3.308$, $p = 0.002$, $d = -0.59$, 95% CI = [-3.03, -0.72]) because some participants missed presses, but we found that the index of within-session divergence in self-generated thought content based on participants' button pressing was significantly negatively correlated with rumination (Fig. 6, a1: $r(31) = -0.392$, $p = 0.029$, 95% CI = [-0.04, -0.66]),

reflection (Fig. 6, a2: $r(31) = -0.374$, $p = 0.038$, 95% CI = [-0.02, -0.64]) and brooding ((Fig. 6, a3: $r(31) = -0.420$, $p = 0.019$, 95% CI = [-0.08, -0.67])). In addition, thought divergence calculated by combining participants' button pressing and manual checks was also significantly correlated with rumination and its sub-factors (Fig. 6).

Discussion

Our first purpose was to explore whether verbal reporting is feasible for accessing task-independent (resting state) self-generated thoughts and, if so, to explore a new simple behavioral paradigm to expand research on self-generated thoughts. In a 10-min resting-state condition, participants were asked to immediately speak out loud any thoughts that came to their minds. To test the feasibility of this approach, we began our experiments with traditional self-caught and retrospective methods for comparison. In the first stage, participants needed only to press the "H" button whenever new thoughts came to mind. At the end of each stage, they retrospectively evaluated their thoughts by responding to eight questions. The number of "H" button presses represented the frequency of self-generated thoughts that participants perceived. For analysis of the oral reporting stages, we first converted recorded verbal reports into text and manually labeled the frequency of

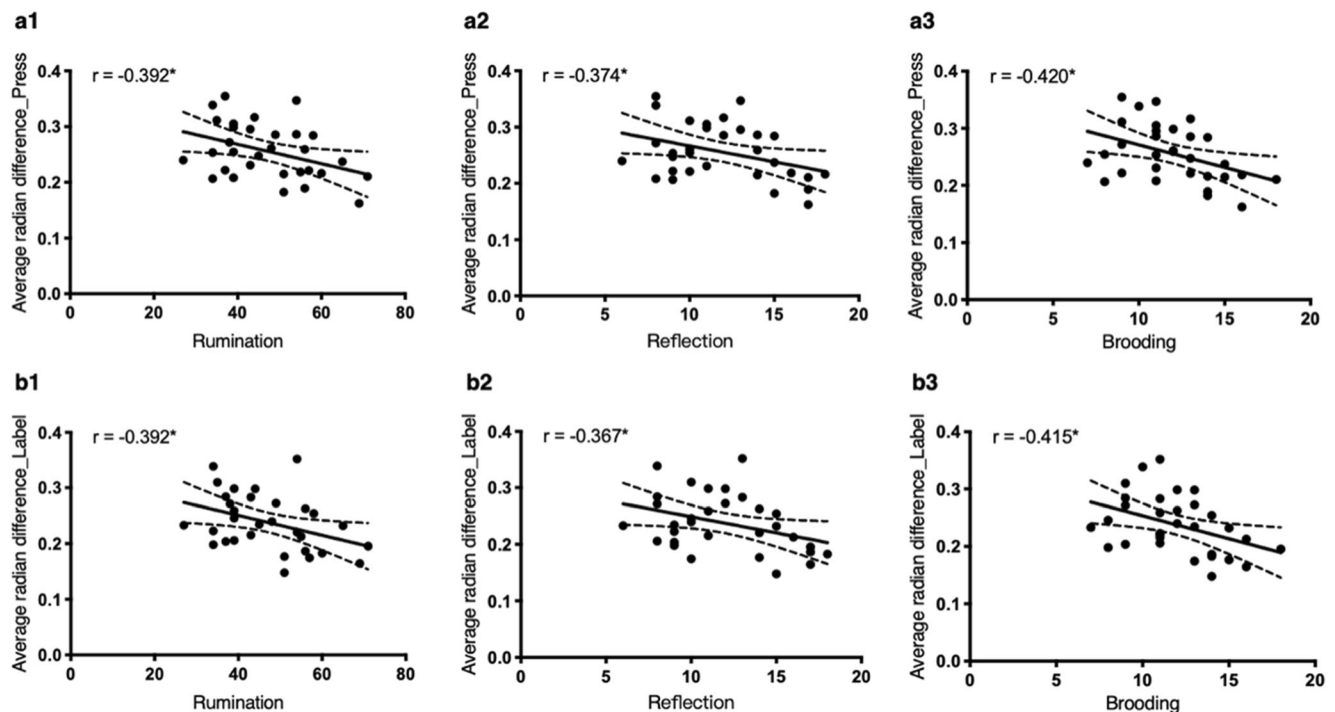


Fig. 6 Correlations between rumination and the radian difference between participants' button pressing and manually labeled thought switches. The Y-axis shows the radian difference in the thought content of the third stage between the two sessions (T1 and T2). "Rumination" is the total score on the Ruminative Responses Scale. "Reflection" and

"Brooding" are sub-factors of the Ruminative Responses Scale. The subscript "press" indicates a topic shift using the participant's button pressing, while the subscript "labeled" indicates manually labeled topic shifts. Two-tailed. * $p < .05$; ** $p < .01$

thought content changes. Our results showed that the resting-state think-aloud method was feasible and reliable. It is worth noting that while the content of individuals' daily thoughts is always changing, what we expect to be reliable are the characteristics of the self-generated thoughts, e.g., the frequency of individual self-generated thoughts, the tendencies of different content dimensions, within-session content divergence, and the degree of sadness expressed. Based on methodological feasibility, we explored whether the thought content representations generated using NLP are behaviorally meaningful and examine the relationship between rumination and self-generated thoughts. We used the pre-trained BERT model to map sentences contained in each self-generated thought episode into 768-dimensional fixed-length vector representations, and quantitatively calculated the divergence of thought content between temporal segments. We also trained a BERT-based NLP text emotion classifier to detect expressions of sadness in verbal reports. Our results showed that the thought content indicators calculated quantitatively by NLP are valid. Meanwhile, we validated the relationship between self-generated thoughts and rumination and found that reflection and brooding could be identified by detecting the divergence of self-generated thought content and expressions of sadness.

To distinguish each time self-generated thoughts arose and calculate the number of self-generated thoughts, we defined self-generated thought episodes on the basis of topic shifts. In the first and third stages, the first time a thought was generated the participant pressed the button, and each subsequent button press indicated a different thought. This difference was based on the difference in topics. The button pressing frequency represents the number of self-generated thought episodes. This definition of the number of self-generated thoughts differs slightly from that based on the task state, however, we believe it is a more operational way to reflect self-generated thoughts in the task-free state. Furthermore, it allows us to dynamically separate self-generated thought episodes for more detailed analysis. Based on this calculation of the number of self-generated thought episodes, we can compare differences in the number of self-generated thought episodes among the three stages. We did not detect significant differences between the button pressing frequency of the first stage and the labeled frequency of the second stage, which supports the comparability between self-caught button pressing and the manually labeled frequency. In addition, we administered the Daydream Frequency Questionnaire, a widely used tool for studying self-generated thoughts (Mason et al., 2007). Previous studies showed that individuals with high daydreaming traits also showed a higher frequency of mind wandering in the laboratory (Smallwood et al., 2004). Consistent with this, we found that both the button pressing frequency of the first stage and the labeled frequency of the second stage were significantly positively related to the total score on the Daydream Frequency Questionnaire. These

results indicate that the labeled frequency of oral reports could be an effective way to study self-generated thoughts.

To integrate the retrospective method, we asked participants to respond to eight questions after each stage corresponding to four dimensions of self-generated thoughts. A two-way repeated-measures ANOVA for each question showed that the main effects of session and stage were not significant, nor was their interaction (Table S2). Of note, non-significant results mean not that the content characteristics measured by the retrospective method in the three stages are statistically equal but rather that they are relatively comparable. Additionally, we measured individual traits of current awareness and attention in daily life with the Mindful Attention Awareness Scale (MAAS) (Brown & Ryan, 2003), which has a total score of 90. The average MAAS score of our participants was 62.2 (range 40 ~ 73), suggesting that most of our subjects could be characterized as self-aware. We found that participants' spontaneous self-generated thoughts were more about themselves than about others and more positive than negative, in accordance with previous research (Hoffmann et al., 2016; Ruby et al., 2013). For example, using the traditional probe-caught method, the frequency of self-generated thoughts was found to be more future-related than past-related, more self-related than other-related, and more positive than negative (Ruby et al., 2013). In a task-free situation (restingstate), Wilson et al. found that most people seem to prefer to do something rather than having nothing to do but think, even if that something is negative (Wilson et al., 2014). We cannot directly compare our research with that of Wilson et al. because we did not measure enjoyment. We assume that people do not choose to put themselves in the default mode by disengaging from the external world, but when they are in this mode, healthy participants, such as those we studied, think more about positive events than negative events overall.

Meta-awareness (or meta-consciousness) is a "monitoring system" that intermittently assesses the contents of consciousness and can be as diverse as the content of experience (Schooler, 2002). Self-generated thoughts with and without meta-awareness is an important topic (Seli et al., 2017). Research has demonstrated that the two states of mind wandering have different implications for task performance (Smallwood et al., 2007, 2008) and depression (Deng et al., 2014; Nayda & Takarangi, 2021). They also differ in patterns of associated brain activity (Christoff et al., 2009). Any time participants self-catch their thoughts, the process involves meta-awareness. Thus, this method is believed to only allow access to self-generated thoughts that are accompanied by meta-awareness, while the probe-caught method is thought to allow access to self-generated thoughts with and without meta-awareness (Schooler et al., 2004; Seli et al., 2017; Smallwood et al., 2007). Our experiment asked participants to self-catch their thoughts, thus involving meta-awareness. The first and third stages need participants to judge if the

emerging thought is on a different topic from the previous thought and press the button, thus requiring meta-awareness of self-generated thought episode switches. The free verbal report stage (stage 2) asks participants to continuously report whatever comes to their mind, regardless of grammar, which does not require meta-awareness regarding their self-generated thought episode switches. There was no significant difference between pressing the button alone and verbal report alone (manually labeled frequency), while the number of button presses was significantly lower for participants in the stage where both button pressing and verbal report were required than in the other two stages (Fig. 2a). We believe that these differences mainly reflect differences in metacognitive load. Intriguing, the third stage differed significantly from the first two stages mainly in participants' button pressing frequencies, while the number of manually labeled self-generated thought episodes did not differ significantly from the first two stages. Moreover, there were no significant differences from the first two stages in retrospective thought content assessment. Most participants reported that they often forgot to press the button in the third stage despite having expressed their thoughts out loud. Therefore, we believed that the higher level of meta-awareness required for Stage 3 is a major limitation of requiring both pressing the button and verbal report.

We designed the third stage (including both button pressing and verbal reporting) to test the concordance of thought switches identified by participants' button presses and thought switches identified manually from the content of verbal reports, which reflected differences in defining topic similarity. Some of the participants' button pressing moments matched verbal report moments within a margin of error of 2 s; however, some button presses did not match, and some were missing. This result may be due to limitations in our experiment. First, the instructions were insufficiently clear. We asked participants to press the button when they noticed a new thought, but we did not specify whether to speak or press first. We found that participants sometimes pressed the button before they spoke and sometimes afterwards. Second, the requirements for monitoring and executive function might have been excessive, which may have led to fewer thought switches. Third, participants may have become fatigued. This stage was always executed last, and continuously speaking aloud might be tiring, but we did not measure the degree of fatigue. Finally, we did not counterbalance task order. To avoid the impact of verbal reporting on self-generated thoughts, stage 1 was always carried out first. In addition, we ordered the three stages by increasing cognitive load, and we did not counterbalance the order to prevent tasks with a higher cognitive load from affecting tasks with a lower cognitive load. Thus, we cannot differentiate between higher cognitive load and

fatigue as causal factors for our finding of fewer button presses in the third stage.

Automation will largely improve the efficiency of NLP analysis and reduce the large amount of manual work and facilitate the reproducibility of research. Although some problems with the design of the third stage may have led to poor execution by participants, we attempted to explore the possibility of automatic detection of thought switches based on the stage 3 method. Fortunately, we found that some of the participants were able to perform the task well in stage 3, and some of them had few missed presses. Accordingly, we conducted a detailed analysis based on 31 participants with sufficient data from stage 3. The 31 participants' results showed that both the button pressing frequency and the labeled frequency in the third stage of the two sessions were significantly correlated with Daydream Frequency Questionnaire scores. Importantly, we found that the index of within-session divergence in self-generated thought content based on participants' button pressing was significantly negatively correlated with rumination, reflection, and brooding (Fig. 6). Therefore, it is feasible to use participants' button pressing to automatically detect thought switches, which requires enhanced training of participants. Furthermore, based on the results of our experiments, we believe that the free verbal report phase is better. Unfortunately, our study does not yet answer the question of automatic machine recognition of self-generated thought episodes. This would greatly improve the efficiency of the NLP method and enable large-scale studies, and it would require the exploration of improved computational approaches.

Having established the effectiveness of the resting-state think-aloud method, we used the second-stage data to quantitatively explore the content characteristics of self-generated thoughts by NLP. We found that the indicator measuring quantitative divergence in thought content was significantly negatively correlated with rumination. In other words, people with high rumination traits generally stick to similar thoughts. The dynamic framework proposed by Christoff et al. (Christoff et al., 2016)(Fig. S7a) defines spontaneous thought as a mental state “that arises relatively freely due to an absence of strong constraints on the contents of each state and on the transitions from one mental state to another.” Mind wandering and rumination seem antithetical based on the dynamics of thought: mind wandering is characterized by free transitions from one idea to another, and ruminative thoughts are often fixed on a single theme or topic. The authors also proposed that “there is a range of low to medium level of automatic constraints that can occur during dreaming, mind-wandering and creative thinking, but thought ceases to be spontaneous at the strongest levels of automatic constraint, such as during rumination or obsessive thought” (Christoff et al., 2016). According to the dynamic framework, thoughts in the resting state arose and proceeded in a relatively free, unconstrained fashion; therefore, they belong to mind wandering under the

concept of spontaneous thought. Our findings suggest that people with high rumination trait have lower thought divergence, supporting the dynamic framework of rumination characterized by strong automatic constraints. As a result, people with high rumination may tend to automatically limit their thoughts to fewer topics or events. At the same time, our results suggest that rumination should be considered a more protracted form of mind wandering and spontaneous thought.

We also found that the tendency to share sad thoughts in the verbal reports detected by the pre-trained emotion recognition model was significantly positively correlated with rumination, suggesting that the self-generated thought content of individuals with high rumination trait is more negative. This aspect is also consistent with the negative tendency of rumination and its relation to negatively valenced self-generated thoughts (Christoff et al., 2016; Nolen-Hoeksema et al., 2008; Watkins, 2008). We further examined the relationship between our two quantitative indicators of self-generated thought content characteristics and different forms of rumination. Reflection and brooding were both correlated with within-session divergence of thought content. Partial correlation results showed the same (Table S4). Furthermore, only brooding retained a significant partial correlation with differences in thought content over longer periods (two test–retest experiments). This is consistent with the notion that brooding represents longer-term repetition of thinking about a few things, while reflection is a short-term reflection on something. This may explain previous research results finding that reflection was related to a decrease in depression over time and could lead to effective problem solving, while brooding was related to more concurrent and longitudinal depression (Treynor et al., 2003). Therefore, longer spans of thought content differences have value in distinguishing between reflection and brooding. In contrast, expressions of sadness in self-generated thoughts were significantly positively related to brooding but not correlated with reflection (Fig. 5b), suggesting that sad thoughts may differentiate reflection and brooding. Indeed, brooding is viewed as moody pondering and a maladaptive form of rumination (Marchetti et al., 2013), while reflection is considered an adaptive form of rumination that involves an attempt to solve problems to improve mood (Treynor et al., 2003).

Our proposed framework using NLP based on verbally reported thoughts is essentially self-reported content because verbal reports are self-reports; however, compared to the traditional ES method in the field of research on self-generated thoughts, our method is innovative in several ways. First, we collected self-generated thoughts in real time and could observe the stream of thoughts from one thought to another while not relying on memory and thus reducing memory bias. Second, in contrast to ES, we directly quantified thoughts without relying on individual introspection. Third, compared to a human rater coding of task-unrelated thoughts, direct

quantitative analysis reduces labor and improves reproducibility of rating and output indexes. Finally, we extended the ability to study the content characteristics of self-generated thoughts. Beyond the divergence and sadness emotion content of sentences, more metrics can be developed based on the rich contents of BERT vectors. Importantly, all such potential metrics would be person-power-free and could facilitate the reproducibility of research on self-generated thoughts.

Our proposed framework for the study of self-generated thoughts in a resting state has several potential applications. We demonstrated that the resting-state think-aloud method was effective and feasible in the absence of an external primary task; i.e., it can be used to study “stream of consciousness” in a resting state. Seeking to understand inner experience during a resting state in MRI scanning, Delamillieure et al. designed a resting state introspective questionnaire and asked participants to complete this questionnaire after an 8-min resting-state MRI scan. They found that most participants showed dominance of a type of mental activity (visual mental imagery, IMAG) at rest (Delamillieure et al., 2010). Gorgolewski and colleagues (Gorgolewski et al., 2014) focused on the relationship between self-generated thoughts and intrinsic neural activity of the brain in resting-state fMRI by administering retrospective questionnaires. They found patterns of self-generated thoughts related to interindividual differences distributed across a wide range of cortical areas and showed that self-generated thoughts are a heterogeneous category of experience. They also proposed that studying the content of self-generated thoughts could help in understanding brain dynamics. These pioneering studies have deepened our understanding of inner experience in a resting state and of the relationship between self-generated thoughts and intrinsic neural fluctuations, despite the difficulty of evaluating spontaneous experiences that occurred in the past. Our results show that the resting-state think-aloud method is an effective way of studying self-generated thoughts with potential applicability to understanding spontaneous brain activity. The method of instant oral reporting, in addition to revealing frequency-related characteristics of self-generated thoughts, affords rich and detailed content characteristics. Specifically, we can quantify the content of self-generated thoughts through NLP, which allows us to more closely examine the content of thoughts. Additionally, we can combine this new method with traditional methods, such as the retrospective method, for more comprehensive studies. Therefore, the resting-state think-aloud method may lead to a comprehensive exploration of self-generated thoughts and spontaneous brain activity through resting-state fMRI. The application of this method in MRI scanning should be explored in the future. In addition, verbal reporting in a resting state can be integrated into

BOLD-fMRI, EEG or fNIRS experiments, providing opportunities to synchronously monitor brain activity related to self-generated thoughts.

Second, we can track the stream of thought fragments over time in individuals, exploring changes in self-generated thought content over time. Specifically, why do some people think adaptively when considering their problems and troubles, while others fall into brooding? How can we help individuals not fall into maladaptive rumination? Our study demonstrated that a brief sample of the stream of thought can reflect the personal trait of rumination. Furthermore, divergence and expressions of sadness of thought content can distinguish reflection and brooding. These results answer the first question while providing guidance and suggestions for the second question. This method has good implementability. We can collect a small sample of individual self-generated thoughts. By analyzing the thought content itself, we can distinguish different forms of rumination and provide targeted help. More future research is required to establish criteria for comparing indicators such as divergence and expressions of sadness to distinguish between reflection and brooding.

Third, we can directly distinguish between self-generated thought models of different mental illnesses to aid in diagnosis and treatment. A brief segment of the individual stream of thought can indeed reflect individual traits. Advances in NLP allow us to perform quantitative analyses of the content of thoughts themselves and to improve the efficiency of our coding. A large sample of thought content can be subjected to machine learning to classify different thought characteristics and help us detect different individual thought forms. Furthermore, the content characteristics of self-generated thoughts differ between various mental illnesses (Gruber et al., 2008). Collecting the stream of thought of participants with mental illnesses, such as depression, may support diagnosis and treatment by detecting their thought content using NLP. This may be a promising area for future research.

Fourth, an interesting possibility for future research is to use BERT models with longer maximum sequence lengths and construct vectors for sequences of thought longer than a single sentence. This would improve the application of the approach to self-generated thought detection.

In summary, our study established a new framework for understanding self-generated thoughts in a resting state. The study offers three main contributions. First, the think-aloud method has been widely used to capture thought processes during the performance of a primary task, such as problem solving and decision making. However, few studies have utilized the think-aloud method to study the stream of thought in a resting state. Here, we applied this method to study resting-state self-generated thoughts and demonstrated its validity. Second, we demonstrated the behavioral significance of quantitative NLP metrics and provided empirical evidence that

rumination is a sticky and negatively valenced type of self-generated thought. Furthermore, we found that thought divergence and expressions of sadness metrics can distinguish between adaptive and maladaptive rumination. Finally, our study demonstrated that simple oral reporting in a resting state is reliable and effective. Our findings will help extend research on the content characteristics of self-generated thoughts to address the complex impact of this phenomenon and its link to mental illness. In particular, NLP can be used to directly quantify the self-generated thought content of individuals to reduce memory and introspection bias and improve reproducibility.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.3758/s13428-021-01710-6>.

Funding This work was supported by the National Key R&D Program of China (2017YFC1309902), the National Natural Science Foundation of China (81671774, 81630031), the Hundred Talents Program and the 13th Five-year Informatization Plan (XXH13505) of Chinese Academy of Sciences, Beijing Municipal Science & Technology Commission (Z161100000216152).

Declarations

Conflict of interest All authors have no conflicts of interest.

Ethics approval All questionnaires and experimental protocols for this study were approved by the Institutional Review Board of the Institute of Psychology, Chinese Academy of Sciences (H19033).

Consent to participate Informed consent was obtained from all individual participants included in the study.

Consent for publication All participants signed informed consent regarding the publication of their data.

References

- Baird, B., Smallwood, J., & Schooler, J. W. (2011). Back to the future: Autobiographical planning and the functionality of mind-wandering. *Consciousness and Cognition*, 20(4), 1604–1611. <https://doi.org/10.1016/j.concog.2011.08.007>
- Baird, B., Smallwood, J., Mrazek, M. D., Kam, J. W., Franklin, M. S., & Schooler, J. W. (2012). Inspired by distraction: Mind wandering facilitates creative incubation. *Psychological Science*, 23(10), 1117–1122. <https://doi.org/10.1177/0956797612446024>
- Barbic, S. P., Durisko, Z., & Andrews, P. W. (2014). Measuring the bright side of being blue: A new tool for assessing analytical rumination in depression. *PLoS One*, 9(11), e112077. <https://doi.org/10.1371/journal.pone.0112077>
- Bartko, J. J. (1966). The Intraclass Correlation Coefficient as a Measure of Reliability. *Psychological Reports*, 19(1), 3–11. <https://doi.org/10.2466/pr0.1966.19.1.3>
- Bhatia, S., & Richie, R. (2020). Transformer networks of human conceptual knowledge. *PsyArXiv*. <https://doi.org/10.31234/osf.io/hs4ra>

- Bhatia, S., Richie, R., & Zou, W. (2019). Distributed semantic representations for modeling human judgment. *Current Opinion in Behavioral Sciences*, 29, 31–36. <https://doi.org/10.1016/j.cobeha.2019.01.020>
- Brown, K. W., & Ryan, R. M. (2003). The benefits of being present: Mindfulness and its role in psychological well-being. *Journal of Personality and Social Psychology*, 84(4), 822–848. <https://doi.org/10.1037/0022-3514.84.4.822>
- Callard, F., Smallwood, J., Golchert, J., & Margulies, D. S. (2013). The era of the wandering mind? Twenty-first century research on self-generated mental activity. *Frontiers in Psychology*, 4, 891. <https://doi.org/10.3389/fpsyg.2013.00891>
- Christoff, K., Gordon, A. M., Smallwood, J., Smith, R., & Schooler, J. W. (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., 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>
- D'Argembeau, A., & Mathy, A. (2011). Tracking the construction of episodic future thoughts. *Journal of Experimental Psychology: General*, 140(2), 258–271. <https://doi.org/10.1037/a0022581>
- Delamillieure, P., Doucet, G., Mazoyer, B., Turbelin, M. R., Delcroix, N., Mellet, E., Zago, L., Crivello, F., Petit, L., Tzourio-Mazoyer, N., & Joliot, M. (2010). The resting state questionnaire: An introspective questionnaire for evaluation of inner experience during the conscious resting state. *Brain Research Bulletin*, 81(6), 565–573. <https://doi.org/10.1016/j.brainresbull.2009.11.014>
- Deng, Y.-Q., Li, S., & Tang, Y.-Y. (2014). The relationship between wandering mind, depression and mindfulness. *Mindfulness*, 5(2), 124–128. <https://doi.org/10.1007/s12671-012-0157-7>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Association for Computational Linguistics*, 4171–4186. <https://doi.org/10.18653/v1/N19-1423>
- DuPre, E., & Spreng, R. N. (2018). Rumination is a sticky form of spontaneous thought. In K.C.R. Fox, & K. Christoff (Eds.), *The Oxford Handbook of Spontaneous Thought* (pp. 509–520). Oxford University Press.
- Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, 87(3), 215–251. <https://doi.org/10.1037/0033-295x.87.3.215>
- Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. The MIT Press. <https://doi.org/10.7551/mitpress/5657.001.0001>
- Ericsson, K. A., Ericsson, K. A., Charness, N., Feltovich, P. J., & Hoffman, R. R. (2006). Protocol analysis and expert thought: Concurrent verbalizations of thinking during experts' performance on representative tasks. In K. A. Ericsson, N. Charness, P. J. Feltovich, & R. R. Hoffman (Eds.), *The Cambridge Handbook of Expertise and Expert Performance* (pp. 223–242). Cambridge University Press. <https://doi.org/10.1017/cbo9780511816796.013>
- Faber, M., & D'Mello, S. K. (2018). How the stimulus influences mind wandering in semantically rich task contexts. *Cognitive Research: Principles and Implications*, 3(1), 35. <https://doi.org/10.1186/s41235-018-0129-0>
- Fodor, J. A. (1975). *The Language of Thought*. Harvard University Press.
- Fox, K. C. R., Andrews-Hanna, J. R., Mills, C., Dixon, M. L., Markovic, J., Thompson, E., & Christoff, K. (2018). Affective neuroscience of self-generated thought. *Annals of the New York Academy of Sciences*. <https://doi.org/10.1111/nyas.13740>
- Gorgolewski, K. J., Lurie, D., Urchs, S., Kipping, J. A., Craddock, R. C., Milham, M. P., Margulies, D. S., & Smallwood, J. (2014). A correspondence between individual differences in the brain's intrinsic functional architecture and the content and form of self-generated thoughts. *PLoS One*, 9(5), e97176. <https://doi.org/10.1371/journal.pone.0097176>
- Gruber, J., Johnson, S. L., Oveis, C., & Keltner, D. (2008). Risk for mania and positive emotional responding: Too much of a good thing? *Emotion*, 8(1), 23–33. <https://doi.org/10.1037/1528-3542.8.1.23>
- Hoffmann, F., Banzhaf, C., Kanske, P., Bermpohl, F., & Singer, T. (2016). Where the depressed mind wanders: Self-generated thought patterns as assessed through experience sampling as a state marker of depression. *Journal of Affective Disorders*, 198, 127–134. <https://doi.org/10.1016/j.jad.2016.03.005>
- Holleran, S. E., & Mehl, M. R. (2008). Let me read your mind: Personality judgments based on a person's natural stream of thought. *Journal of Research in Personality*, 42(3), 747–754. <https://doi.org/10.1016/j.jrp.2007.07.011>
- Killingsworth, M. A., & Gilbert, D. T. (2010). A wandering mind is an unhappy mind. *Science*, 330(6006), 932. <https://doi.org/10.1126/science.1192439>
- Klinger, E. (2009). Daydreaming and fantasizing: Thought flow and motivation. In K. D. Markman, W. M. P. Klein, & J. A. Suhr (Eds.), *Handbook of imagination and mental simulation* (pp. 225–239). Psychology Press.
- Konu, D., Turnbull, A., Karapanagiotidis, T., Wang, H.-T., Brown, L. R., Jefferies, E., & Smallwood, J. (2020). A role for the ventromedial prefrontal cortex in self-generated episodic social cognition. *Neuroimage*, 218, 116977. <https://doi.org/10.1016/j.neuroimage.2020.116977>
- Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2), 211–240. <https://doi.org/10.1037/0033-295X.104.2.211>
- Levinson, S. C. (2003). Language and mind: Let's get the issues straight! In D. Gentner, & S. Goldin-Meadow (Eds.), *Language in mind: Advances in the study of language and thought* (pp. 25–46). Cambridge, MA: MIT Press.
- Lindquist, K. A., MacCormack, J. K., & Shablack, H. (2015). The role of language in emotion: Predictions from psychological constructionism. *Frontiers in Psychology*, 6, 444–444. <https://doi.org/10.3389/fpsyg.2015.00444>
- Lyubomirsky, S., Tucker, K. L., Caldwell, N. D., & Berg, K. (1999). Why ruminators are poor problem solvers: Clues from the phenomenology of dysphoric rumination. *Journal of Personality and Social Psychology*, 77(5), 1041.
- Madsen, P. L., Hasselbalch, S. G., Hagemann, L. P., Olsen, K. S., Bulow, J., Holm, S., Wildschiodtz, G., Paulson, O. B., & Lassen, N. A. (1995). Persistent resetting of the cerebral oxygen/glucose uptake ratio by brain activation: Evidence obtained with the Kety–Schmidt technique. *Journal of Cerebral Blood Flow and Metabolism*, 15(3), 485–491. <https://doi.org/10.1038/jcbfm.1995.60>
- Marchetti, I., Koster, E. H. W., & De Raedt, R. (2013). Rest-related dynamics of risk and protective factors for depression: A behavioral study. *Clinical Psychological Science*, 1(4), 443–451. <https://doi.org/10.1177/2167702613489668>
- Marchetti, I., Koster, E. H. W., Klinger, E., & Alloy, L. B. (2016). Spontaneous thought and vulnerability to mood disorders: The dark side of the wandering mind. *Clinical Psychological Science*, 4(5), 835–857. <https://doi.org/10.1177/2167702615622383>
- Mason, M. F., Norton, M. I., Van Horn, J. D., Wegner, D. M., Grafton, S. T., & Macrae, C. N. (2007). Wandering minds: The default network and stimulus-independent thought. *Science*, 315(5810), 393–395. <https://doi.org/10.1126/science.1131295>
- McVay, J. C., & Kane, M. J. (2012). Why does working memory capacity predict variation in reading comprehension? On the influence of mind wandering and executive attention. *Journal of Experimental*

- Psychology: General*, 141(2), 302–320. <https://doi.org/10.1037/a0025250>
- Nayda, D. M., & Takarangi, M. K. T. (2021). The cost of being absent: Is meta-awareness of mind-wandering related to depression symptom severity, rumination tendencies and trauma intrusions? *Journal of Affective Disorders*, 292, 131–138. <https://doi.org/10.1016/j.jad.2021.05.053>
- Nolen-Hoeksema, S., Wisco, B. E., & Lyubomirsky, S. (2008). Rethinking rumination. *Perspectives on Psychological Science*, 3(5), 400–424. <https://doi.org/10.1111/j.1745-6924.2008.00088.x>
- Perkins, A. M., Arnone, D., Smallwood, J., & Mobbs, D. (2015). Thinking too much: Self-generated thought as the engine of neuroticism. *Trends in Cognitive Sciences*, 19(9), 492–498. <https://doi.org/10.1016/j.tics.2015.07.003>
- Prinz, J. J. (2004). *Furnishing the mind: Concepts and their perceptual basis*. MIT Press.
- Reisenzein, R., & Junge, M. (2012). Language and emotion from the perspective of the computational belief-desire theory of emotion. In E. Lorini, & A. Herzig (Eds.), *Dynamicity in Emotion Concepts*. Springer Verlag.
- Rescorla, M. (2019). The Language of Thought Hypothesis (Summer 2019 Edition). In N. Z. Edward (Ed.). *The Stanford Encyclopedia of Philosophy* URL = <<https://plato.stanford.edu/archives/sum2019/entries/language-thought/>>
- Ruby, F. J., Smallwood, J., Engen, H., & Singer, T. (2013). How self-generated thought shapes mood—the relation between mind-wandering and mood depends on the socio-temporal content of thoughts. *PloS One*, 8(10), e77554. <https://doi.org/10.1371/journal.pone.0077554>
- Russo, J. E., Johnson, E. J., & Stephens, D. L. (1989). The validity of verbal protocols. *Memory and Cognition*, 17(6), 759–769. <https://doi.org/10.3758/bf03202637>
- Schooler, J. W. (2002). Re-representing consciousness: Dissociations between experience and meta-consciousness. *Trends in Cognitive Sciences*, 6(8), 339–344. [https://doi.org/10.1016/s1364-6613\(02\)01949-6](https://doi.org/10.1016/s1364-6613(02)01949-6)
- Schooler, J. W., Reichle, E. D., & Halpern, D. V. (2004). Zoning out while reading: Evidence for dissociations between experience and metaconsciousness. In D. T. Levin (Ed.), *Thinking and seeing: Visual metacognition in adults and children* (pp. 203–226). MIT Press.
- Seli, P., Carriere, J. S., Levene, M., & Smilek, D. (2013). How few and far between? Examining the effects of probe rate on self-reported mind wandering. *Frontiers in Psychology*, 4, 430. <https://doi.org/10.3389/fpsyg.2013.00430>
- Seli, P., Ralph, B. C. W., Risko, E. F., Schooler, J. W., Schacter, D. L., & Smilek, D. (2017). Intentionality and meta-awareness of mind wandering: Are they one and the same, or distinct dimensions? *Psychonomic Bulletin & Review*, 24(6), 1808–1818. <https://doi.org/10.3758/s13423-017-1249-0>
- Shrout, P. E. (1998). Measurement reliability and agreement in psychiatry. *Statistical Methods in Medical Research*, 7(3), 301–317. <https://doi.org/10.1177/096228029800700306>
- Siegle, G. J., Moore, P. M., & Thase, M. E. (2004). Rumination: One construct, many features in healthy individuals, depressed individuals, and individuals with lupus. *Cognitive Therapy and Research*, 28(5), 645–668. <https://doi.org/10.1023/B:COTR.0000045570.62733.9f>
- Singer, J. L., & Antrobus, J. S. (1972). Daydreaming, imaginal processes, and personality: A normative study. In Sheehan, P. (Ed.), *The function and nature of imagery* (pp. 175–202). Academic Press, New York.
- Singh, M., Richie, R., & Bhatia, S. (2020). Representing and predicting everyday behavior. *PsyArXiv*. <https://doi.org/10.31234/osf.io/zb53h>
- Smallwood, J., & Andrews-Hanna, J. (2013). Not all minds that wander are lost: The importance of a balanced perspective on the mind-wandering state. *Frontiers in Psychology*, 4, 441. <https://doi.org/10.3389/fpsyg.2013.00441>
- Smallwood, J., & O'Connor, R. C. (2011). Imprisoned by the past: Unhappy moods lead to a retrospective bias to mind wandering. *Cognition & Emotion*, 25(8), 1481–1490. <https://doi.org/10.1080/02699931.2010.545263>
- 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, 487–518. <https://doi.org/10.1146/annurev-psych-010814-015331>
- Smallwood, J., Obonsawin, M., & Reid, H. (2002). The effects of block duration and task demands on the experience of task unrelated thought. *Imagination, Cognition and Personality*, 22(1), 13–31. <https://doi.org/10.2190/tbml-n8jn-w5yb-419r>
- Smallwood, J., Davies, J. B., Heim, D., Finnigan, F., Sudberry, M., O'Connor, R., & Obonsawin, M. (2004). Subjective experience and the attentional lapse: Task engagement and disengagement during sustained attention. *Consciousness and Cognition*, 13(4), 657–690. <https://doi.org/10.1016/j.concog.2004.06.003>
- Smallwood, J., McSpadden, M., & Schooler, J. W. (2007). The lights are on but no one's home: Meta-awareness and the decoupling of attention when the mind wanders. *Psychon Bull Rev*, 14(3), 527–533. <https://doi.org/10.3758/bf03194102>
- Smallwood, J., McSpadden, M., & Schooler, J. W. (2008). When attention matters: The curious incident of the wandering mind. *Memory and Cognition*, 36(6), 1144–1150. <https://doi.org/10.3758/mc.36.6.1144>
- Smallwood, J., Schooler, J. W., Turk, D. J., Cunningham, S. J., Burns, P., & Macrae, C. N. (2011). Self-reflection and the temporal focus of the wandering mind. *Consciousness and Cognition*, 20(4), 1120–1126. <https://doi.org/10.1016/j.concog.2010.12.017>
- Smallwood, J., Ruby, F. J., & Singer, T. (2013). Letting go of the present: Mind-wandering is associated with reduced delay discounting. *Consciousness and Cognition*, 22(1), 1–7. <https://doi.org/10.1016/j.concog.2012.10.007>
- Smallwood, J., Karapanagiotidis, T., Ruby, F., Medea, B., de Caso, I., Konishi, M., Wang, H. T., Hallam, G., Margulies, D. S., & Jefferies, E. (2016). Representing representation: Integration between the temporal lobe and the posterior cingulate influences the content and form of spontaneous thought. *PloS One*, 11(4), e0152272. <https://doi.org/10.1371/journal.pone.0152272>
- 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, 103007. <https://doi.org/10.1016/j.concog.2020.103007>
- 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., Cassol, H., & D'Argembeau, A. (2013a). Phenomenology of future-oriented mind-wandering episodes. *Frontiers in Psychology*, 4, 425. <https://doi.org/10.3389/fpsyg.2013.00425>
- Stawarczyk, D., Majerus, S., & D'Argembeau, A. (2013b). Concern-induced negative affect is associated with the occurrence and content of mind-wandering. *Consciousness and Cognition*, 22(2), 442–448. <https://doi.org/10.1016/j.concog.2013.01.012>
- Szpunar, K. K. (2010). Episodic future thought: An emerging concept. *Perspectives on Psychological Science*, 5(2), 142–162. <https://doi.org/10.1177/1745691610362350>

- Tillas, A. (2015). Language as grist to the mill of cognition. *Cognitive Processing*, 16(3), 219–243. <https://doi.org/10.1007/s10339-015-0656-2>
- Treynor, W., Gonzalez, R., & Nolen-Hoeksema, S. (2003). Rumination reconsidered: A psychometric analysis. *Cognitive Therapy and Research*, 27(3), 247–259. <https://doi.org/10.1023/A:1023910315561>
- Wang, H. T., Poerio, G., Murphy, C., Bzdok, D., Jefferies, E., & Smallwood, J. (2018). Dimensions of experience: Exploring the heterogeneity of the wandering mind. *Psychological Science*, 29(1), 56–71. <https://doi.org/10.1177/0956797617728727>
- 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>
- Wilson, T. D. (1994). The proper protocol: Validity and completeness of verbal reports. *Psychological Science*, 5(5), 249–252. <https://doi.org/10.1111/j.1467-9280.1994.tb00621.x>
- Wilson, T. D., Reinhard, D. A., Westgate, E. C., Gilbert, D. T., Ellerbeck, N., Hahn, C., Brown, C. L., & Shaked, A. (2014). Just think: The challenges of the disengaged mind. *Science*, 345(6192), 75–77. <https://doi.org/10.1126/science.1250830>
- Zhao, W. J., Richie, R., & Bhatia, S. (2019). Process and content in decisions from memory. *PsyArXiv*. <https://doi.org/10.31234/osf.io/nwxs2>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.