Wolsink, L. N., Jacob, G., & Merz, C. J. (2026). Memory Suppression Does Not Slow Down, But Retrieval Practice Speeds up Later Recall in the Think/No-Think Task. *Collabra: Psychology*, *12*(1). https://doi.org/10.1525/collabra.163092

Cognitive Psychology

Memory Suppression Does Not Slow Down, but Retrieval Practice Speeds Up Later Recall in the Think/No-Think Task

Lianne N. Wolsink[1][a], George Jacob[2], Christian J. Merz[1]

[1] Department of Cognitive Psychology, Institute of Cognitive Neuroscience, Faculty of Psychology, Ruhr University Bochum, Bochum, Germany,
[2] Department of Neuropsychology, Institute of Cognitive Neuroscience, Faculty of Psychology, Ruhr University Bochum, Bochum, Germany

Keywords: Memory suppression, Recall delay, Response time, Suppression-Induced Forgetting, Think/No-Think

https://doi.org/10.1525/collabra.163092

Collabra: Psychology

Vol. 12, Issue 1, 2026

In the Think/No-Think (TNT) task, the effect of memory suppression on later recall is typically assessed by the difference in total recall score between No-Think and baseline words, termed suppression-induced forgetting (SIF). Considering the usually small, and occasionally null, effects, we reason that this recall score, that is based on a binary measure (remembered or forgotten), might be too insensitive to detect the effects of suppression on later recall. In this study, we therefore tested recall delay as a potentially more sensitive measure, as it is measured on a continuum. Recent studies using modified TNT tasks suggest that No-Think words are indeed recalled slower than baseline or Think words. We tested this effect using a typical TNT task with word pairs, in a larger sample (N = 80), and with a preregistered analysis plan complemented by Bayesian analyses. Moreover, we tested recall delay for emotionally negative words. Using a same-probe recall test, the SIF effect was significant but small, and it was not supported by a Bayesian analysis. Moreover, the SIF effect was not significant for neutral and negative words separately. Retrieval suppression did not slow down recall time, but retrieval practice (i.e., Think condition) sped up later recall compared to baseline words. These findings support growing concerns regarding the robustness of suppression effects on later recall.

1. Introduction

Intrusive memories of traumatic events, such as in patients with post-traumatic stress disorder (PTSD), can be distressing and hinder recovery (Ehlers & Clark, 2000). Research in the last decades has shown that it is possible to suppress memory intrusions (e.g., Gagnepain et al., 2017; Levy & Anderson, 2012), and that this suppression can cause forgetting (Clark et al., 2025; Stramaccia et al., 2021). The Think/No-Think task (TNT; Anderson & Green, 2001) is an experimental paradigm designed to test this retrieval suppression and its effects on later memory recall. The TNT task consists of three main phases: learning, TNT, and recall. First, participants learn word pairs containing one cue and one target word (e.g., waffle – maple). During the TNT phase, the cue (e.g., waffle) is presented, and participants must either think of the target (e.g., maple; Think condition) or suppress any thoughts of the target (No-Think condition). During the recall phase, participants are presented with the cues again and must recall the target (e.g., maple). No-Think targets are typically poorly recalled in this later cued recall phase compared to words that were learned but not manipulated (baseline words), an effect referred to as suppression-induced forgetting (SIF; Anderson & Green, 2001). In this study, we investigated if retrieval suppression can also influence the time it takes to recall suppressed memories.

In healthy individuals, the SIF effect size is typically small to moderate (d ≈ 0.20 – 0.40 in Clark et al., 2025; d = 0.28 in Stramaccia et al., 2021), but not all studies find a SIF effect (Bergström et al., 2009; Bulevich et al., 2006; Mecklinger et al., 2009; Wiechert et al., 2023). Recent studies have raised concerns that the SIF effect may be less robust and replicable than previously assumed (Wessel et al., 2020, 2025; Wiechert et al., 2023), as well as dependent on optimal test conditions (Zhao et al., 2025), thereby limiting the potential for generalization outside a controlled lab setting. Besides the potential lack of robustness of SIF itself, null findings could also be attributed to, for example, the use of an online context (Wiechert et al., 2023) or suboptimal test conditions (Zhao et al., 2025). The effect size is further negatively related to unspecified suppression instructions (Stramaccia et al., 2021) and a lack of instruction compliance (Liu et al., 2021).

Another possible explanation for null findings could be the insensitive measurement of the SIF effect. Many TNT

a Corresponding author: lianne.wolsink@ruhr-uni-bochum.de

studies use word pairs as stimuli (e.g., 17 of 25 studies in the meta-analysis by Stramaccia et al., 2021). Besides, the SIF effect is smaller for words than for pictures (Stramaccia et al., 2021). This may not reflect words being harder to forget, but rather the greater sensitivity of recall measures for pictures compared to the typically binary measures used for words. While picture recall is scored on a continuum, words are scored as either recalled or forgotten, and considering the small effect size of SIF, many suppressed words are later successfully recalled. The effects of memory suppression on later recall may be more subtle, and typical recall measures might be too insensitive to detect these effects.

In the current study, we investigate the use of recall response time as a potentially more sensitive measure of suppression effects. Slower recall time of suppressed target words could reflect a weakening of the cue-target association, such that perceiving the cue no longer leads to the automatic retrieval (or intrusion) of the target as quickly as before suppression. In daily life, this delay in memory retrieval allows for interference from thoughts or external stimuli, which can halt retrieval and prevent the intrusion from entering awareness. Slower memory retrieval could therefore reduce the likelihood of intrusions.

Recall response time has been assessed in previous studies, but often as an additional measure without a hypothesis or specified definition (e.g., if final recall time was corrected for baseline recall time, i.e., recall delay; Cano & Knight, 2016; Noreen & MacLeod, 2013; Schechtman et al., 2021; Waldhauser et al., 2012). Furthermore, heterogeneity in experimental designs makes it difficult to compare the findings across studies. Only two studies have specifically tested recall response time (Haghighi et al., 2020; Nishiyama & Saito, 2021: Study 2). In these studies, participants needed more time to recall No-Think words compared to Think or baseline words, respectively. These findings indicate that memory suppression influenced later recall, even if the word was successfully recalled.

Both studies, however, used modified versions of the TNT task. One included a dominant number of No-Think trials and assessed baseline performance on a separate day (Haghighi et al., 2020), and the other study used word triplets, including substitutes, to be used by half of the participants during suppression, and a dot-probe task with the same words in between the TNT and recall phase (Nishiyama & Saito, 2021). Notably, these modifications could have affected memory performance. In the current study, we tested the effects of memory suppression on recall delay using a typical TNT task with word pairs, in a larger sample (planned N = 80, compared to N = 20 (Haghighi et al., 2020) and N = 45 (Nishiyama & Saito, 2021)), and with a preregistered analysis plan complemented by Bayesian analyses. Our first hypothesis concerns SIF: We predicted that the recall score of No-Think words will be lower than the baseline recall score. The second hypothesis concerns suppression-induced recall delay: We predicted that recall delay of No-Think words will be larger than baseline recall delay, i.e., slower responses to No-Think words.

Memory intrusions are particularly problematic when they are distressing and emotionally negative, such as in PTSD. It is therefore even more important to be able to suppress negative memories. Previous studies on emotionally negative stimuli in the TNT task have reported mixed results (Murray et al., 2011; Nørby et al., 2010; van Schie et al., 2013). Therefore, we included both emotionally neutral and negative word pairs and explored the SIF effect and recall delay for these words, separately. Moreover, in one previous study (Nishiyama & Saito, 2021), the effect on recall delay was larger when participants used thought substitution (i.e., thinking of a substitute word; Hertel & Calcaterra, 2005) as a strategy during memory suppression, compared to thought avoidance (i.e., creating a “blank mind”). The memory of the substitute could interfere with the original memory (Racsmány et al., 2012), thereby potentially slowing down the recall of the original memory. In the current study, we included the memory suppression strategy (thought avoidance and thought substitution) in a randomized, between-subjects design. Due to limited resources, and therefore a potential lack of statistical power, we analyzed the effects of word emotionality (neutral vs. negative) and strategy in an exploratory manner.

2. Methods

We preregistered this study on OSF before observing the data (https://doi.org/10.17605/OSF.IO/A34CJ).

2.1. Participants

An a-priori power calculation using G*Power (Faul et al., 2009; for ANOVA: Repeated measures, within factors with α = 0.05, 1-β = 0.95, effect size f = 0.20, number of groups = 1, number of conditions = 3, correlation among repeated measures = 0.50, nonsphericity correction ε = 0.80) indicated a required sample size of 78 participants. We based the effect size on the lower boundary of the effect size for the SIF effect after a specific suppression/substitution instruction, as reported in a recent meta-analysis (Stramaccia et al., 2021). We preregistered a target sample size of 80 participants. In total, we tested 92 participants, 12 of whom had their sessions terminated prematurely because they did not learn the preregistered minimum number of word pairs within a reasonable time window (see section 2.2). Consequently, the data of 80 participants (24 men and 56 women, 0 non-binary, 73 undergraduate students, age: M = 24.71, SD = 5.75) were eligible for analysis. The thought avoidance group comprised 11 men and 29 women (age: M = 23.30, SD = 3.16), whereas the thought substitution group comprised 13 men and 27 women (age: M = 26.13, SD = 7.27).

Participants were mainly recruited on the campus of Ruhr University Bochum in Bochum, Germany. Participation was voluntary and participants were reimbursed with course credit points or money (€20). Exclusion criteria for participation were red-green color blindness, insufficient German language skills, and age below 18 years. We additionally checked whether participants had already participated in another TNT study that was conducted at the faculty (N=0). The study was framed as a study investigating

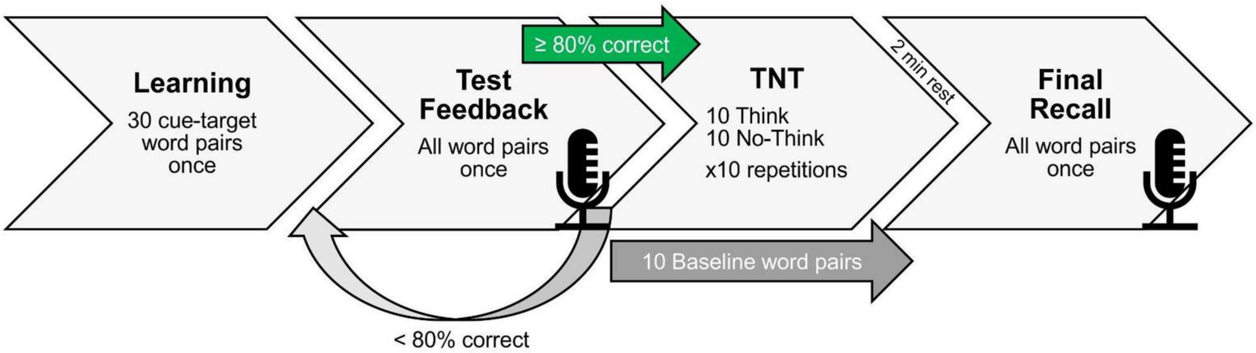

Figure 1. Overview of the phases of the Think/No-Think task

Note: The Test Feedback phase was repeated until at least 80% (24 out of 30) of the word pairs were learned (see section 2.2 for details). Baseline word pairs are not manipulated and therefore not part of the TNT phase. The microphone indicates the audio recording of verbal responses.

the cognitive mechanisms behind attention. Participants were not informed about the actual purpose of the study, and mentioning the word "memory" was avoided until the end of the test session.

2.2. Think/No-Think Task

The TNT task was administered using Presentation® (Version 23.1, Neurobehavioral Systems, Inc., Berkeley, CA, www.neurobs.com). The TNT task comprised four phases: Learning, Test Feedback, TNT, and Final Recall. All phases were performed consecutively (Figure 1). During the task, words were presented in white, unless specified otherwise, on a black background. During the first phase, participants learned 30 German word pairs containing one cue and one target. The cue was a neutral word for all word pairs. However, 15 word pairs contained a neutral target (e.g., procedure – cardboard) and 15 word pairs contained a negative target (e.g., minute – death). We used an adapted version of a TNT task from a previous study (Schmidt et al., 2022). Words were selected from the Berlin Affective Word List Reloaded (BAWL-R; Võ et al., 2009) based on valence and randomly allocated to word pairs. Negative and neutral words differed in valence and arousal ratings, but not in imageability. Words within one word pair were not semantically associated (for a detailed description, see Schmidt et al., 2022). The word pairs and instructions are available on OSF (https://doi.org/10.17605/OSF.IO/WB4CQ). During this learning phase, each word pair was presented once for five seconds, and participants were instructed to memorize the word pairs. The presentation order of the word pairs was randomized per participant.

Subsequently, these word pairs were tested in a Test Feedback phase. Each cue was presented once for four seconds and participants had to verbally report the target. Again, the presentation order of the word pairs was randomized per participant. Trials lasted until the experimenter pressed a key on the keyboard indicating a correct or incorrect answer. If the answer was incorrect, the correct target was displayed on the screen for one second. Participants had to learn at least 80% of the word pairs to pass the Test Feedback phase, to ensure the encoding of a sufficient number of word pairs for manipulation. If they failed to reach this threshold, the Test Feedback phase was repeated until the criterion was reached. After each Test Feedback phase, participants were allowed to rest for up to one minute. If participants needed many Test Feedback phase repetitions and appeared to have a flat learning curve, the session was terminated as preregistered, and participants were debriefed and reimbursed accordingly (12 out of 92 participants: 13%). To allow for individual differences in learning rate, we did not specify a strict criterion for this termination and relied on the assessment of the experimenter regarding participants' ability to learn enough word pairs within a reasonable time window. Eligible participants learned on average 83.83% of the 30 word pairs (M = 25.15, SD = 1.33, range: 24-29), of which 81.08% of the 15 neutral word pairs (M =12.16, SD=1.35, range: 9-15) and 86.59% of the 15 negative word pairs (M = 13.00, SD = 1.12, range: 10-15) were learned. The number of learned word pairs did not significantly differ between baseline, Think, and No-Think word pairs (over all word pairs: $F(2,158)$ = 1.81, p = .167; neutral word pairs: $F(2,158)$ = 1.81, p = .166; negative word pairs: $F(2,158)$ = 0.49, p = .612). Participants completed on average 6.91 Test Feedback phase repetitions (SD = 5.27), with a median value of 6 repetitions (Range: 1 – 31; with 4 participants with >15 repetitions).

The TNT phase included 20 word pairs (excluding 10 word pairs used as baseline) of the initially learned 30 word pairs. Baseline words were learned but not manipulated during the TNT phase. Each trial started with a central fixation cross in white for two seconds, then a central fixation cross in green or red for one second, and the presentation of the cue for four seconds, again either in green or red, depending on the word condition. Participants were instructed to think of the target and say the word out loud in case of a green word (Think words, 10 word pairs), and not to think of the target in case of a red word (No-Think words, 10 word pairs). The instruction for the No-Think trials depended on the experimental group. Participants in the Thought Avoidance group were instructed to clear their minds and to not distract themselves in any way. Participants in the Thought Substitution group were instructed to think of a substitute word, and to not think of anything but the substitute. They were instructed to refrain from saying the substitute out loud, to not induce an additional difference between the thought avoidance and thought substitution instructions. They did not learn substitutes beforehand and had to decide upon a substitute themselves. Further, they were not allowed to use any of the words from the task and were instructed that each red cue needed to have its own substitute. The substitutes were not documented by the experimenter.

Participants were randomly assigned to one of the two experimental groups (Thought Avoidance or Thought Substitution) before they entered the laboratory, via a simple randomization technique (alternating one-by-one). Before beginning the TNT phase, participants received written instructions on the screen, additional verbal instructions from the experimenter, and one Think and one No-Think practice trial on paper. The experimenter ensured that participants understood the instructions well before starting the TNT phase. We did not use additional diagnostic questionnaires or practice trials.

In the TNT phase, each word pair was repeated 10 times, leading to a total of 200 trials. For each repetition, word pairs were presented in random order. Word pairs were either in the Think or No-Think condition, i.e., they did not switch between conditions during the task. The experimenter scored each trial as correct or incorrect using the keyboard. Participants reported to the experimenter after each No-Think trial whether suppression was successful (i.e., correct or incorrect). When No-Think trials were scored as incorrect, a short beeping sound was played. The TNT phase lasted approximately 25 minutes and did not contain breaks. Afterwards, participants rested for two minutes.

During the last phase of the experiment, the Final Recall phase, participants underwent a same-probe recall test. They were presented with the cue of each of the 30 initially learned word pairs (including baseline word pairs) in white color for four seconds and were instructed to recall the target. Recall time could exceed this four-second window, though, as trials lasted until the experimenter pressed a key on the keyboard indicating a correct or incorrect answer. Participants had, in theory, unlimited time to respond. However, trials were terminated by the experimenter when the participant indicated that they did not know the target, and at least after the four-second trial window. The word pairs were presented in random order. During all trials of the TNT task, participants' verbal responses were recorded using a table microphone.

2.3. Procedure

Experimenters (for 63 participants: undergraduate students; for 29 participants: first author) were extensively trained before data collection. They familiarized themselves with the literature, general lab procedures, and subtleties of the TNT instructions, before administering the experimental procedure multiple times on fellow students who were unfamiliar with the TNT task, for practice. In addition, the initial test sessions with real participants were supervised by the first author.

Upon entering the laboratory, participants read the study information and signed informed consent. The experiment started with a short demographic questionnaire, asking for age, gender, occupation, the existence of mental disorders and/or neurological illnesses, medication use, and alcohol and/or drug use one day before and on the day of testing (see section 2.5. for data exclusion criteria). We did not ask participants about their sleep duration. Then, participants performed the TNT task. After completing the task, participants rated the emotionality of the target of each word pair on a 5-point Likert scale (1 = very negative to 5 = very positive). These ratings confirmed that negative targets (M = 1.56, SD = 0.60) were rated as more negative than neutral targets (M = 3.09, SD = 0.30; $t(79)$ = -23.16, p < .001, d = -2.59). Then, participants completed a post-experimental questionnaire that assessed self-reported difficulty, ability, instruction compliance, and distraction by accident and on purpose, on a 4-point Likert scale (not at all, sometimes, often, always). Participants also indicated whether they had expected the Final Recall phase (yes, no, not sure). Results from the post-experimental questionnaire are reported in Supplement I. Lastly, participants were debriefed and reimbursed. The entire experiment lasted approximately 90 minutes, depending on the number of Test Feedback phases required and time taken to complete the questionnaires.

2.4. Outcome Measures

Raw data was processed using MATLAB (Version 2024b; The MathWorks Inc., 2024). The main outcome measures were recall score and recall delay. Recall score is the proportion of correctly recalled target words per word condition (Think, No-Think, baseline). Only words that were correctly recalled during the final Test Feedback phase were included in data analysis (i.e., conditionalized data; Nardo & Anderson, 2024).

The audio recordings were further processed to extract the onset of the verbal response relative to trial onset in seconds. For this, we manually determined the onset of the correct verbal response in each trial by listening to the audio recording and viewing the audio waveform using Audacity (Version 3.6.4). For each word, recall delay was calculated by subtracting the response time during the final Test Feedback phase from the response time during the Final Recall phase, to correct for individual response speed. Only words that were correctly recalled during the final round of the Test Feedback phase and Final Recall phase were included in data analysis. This means that trials without a response or an incorrect response were excluded from data analysis. However, if participants first gave an incorrect response and then corrected themselves, the onset of the second, correct, response was taken as response time and the trial was included in the analyses. For each participant, recall delay was averaged over word conditions (Think, No-Think, baseline). Due to technical problems with the microphone, the final dataset for the audio responses contained data of 62 participants.

In addition to our preregistered analysis, we also included the timings of the keyboard presses by the experimenter during the respective recall phases as an additional outcome measure for calculating recall delay. Although keyboard presses are temporally less precise than audio recordings, we were able to include data from all 80 participants. Results from these analyses are reported in Supplement II.

2.5. Confirmatory Data Analysis

Data analysis was done using JASP (Version 0.19.1; JASP Team, 2024), using a statistical significance level of α = .05. Since our dependent variables were not normally distributed as confirmed by Shapiro-Wilk tests of normality, we had to deviate from our preregistered analysis plan. Instead of parametric analyses of variance (ANOVA), we performed non-parametric Friedman tests with Holm-adjusted Conover's post-hoc comparisons. To test the first and second hypotheses, we performed Friedman tests with word condition (baseline, Think, No-Think) as the within-subjects factor and as dependent variable recall score, for hypothesis one, and recall delay, for hypothesis two. All Friedman tests were performed with conditionalized word pairs to test the respective hypothesis, and for negative and neutral word pairs (conditionalized) separately, for exploratory purposes. In addition, we compared the recall score and recall delay of neutral vs. negative words for each word condition (baseline, Think, No-Think) separately, using Wilcoxon signed-rank tests. We complemented these analyses with Bayesian one-tailed paired samples t-tests for all three outcome measures, to determine the amount of evidence in favor of the alternative or null hypothesis. We used the default Cauchy prior with scaling factor r = .707. Bayes factors (BF_{10}) were interpreted using Jeffreys (1961) terminology.

Primary analyses were first performed without accounting for the answers to the post-experimental questionnaire. The analyses were then repeated after excluding the data from participants who indicated a 1 (not at all) on instruction compliance and/or a 4 (always) on distraction on purpose (N = 3). Similarly, in case of self-reported mental disorders and/or neurological illnesses (N = 9), we repeated the analyses without the data from these participants to examine whether the presence of mental disorders and/or neurological illnesses affected the results.

2.6. Exploratory Data Analysis

As additional exploratory analyses, we examined the effect of strategy (Thought Avoidance vs. Thought Substitution) and word emotionality (neutral vs. negative) on recall score and recall delay. We calculated two indices: SIF-score and a retrieval index. The SIF-score for recall score and recall delay was calculated as follows: ((baseline – No-Think) / baseline; cf., Ashton et al., 2020). A score of zero indicates no difference between baseline and No-Think recall. Positive scores indicate forgetting (suppression effect), while negative scores indicate better recall of No-Think words. The SIF-score was calculated across all words, and separately for negative and neutral words.

We also assessed a potential retrieval facilitation effect of Think words compared to baseline words. Therefore, a retrieval index was calculated, again, separately for recall score and recall delay. This index is defined as follows: ((baseline – Think) / baseline). A score of zero indicates no difference between the recall of baseline and Think words. Positive scores indicate better recall of baseline words, while negative scores indicate better recall of Think words (facilitation effect). Next, we conducted four 2 x 2 mixed ANOVAs with word emotionality (neutral vs. negative) as within-subjects factor, strategy (Thought Avoidance vs. Thought Substitution) as between-subjects factor, and SIF score, recall delay SIF score, retrieval index, and recall delay retrieval index, as dependent variables. A sensitivity analyses revealed that the analyses for SIF score and retrieval index (with N=80 each) have 95% power to detect effect sizes of Cohen's $f \geqslant .20$. The recall delay analyses (with N=62 each) have 95% power to detect effect sizes of Cohen's $f \geqslant .23$ (when assuming a correlation of r=.50 between repeated measures for both sensitivity analyses).

3. Results

The data and JASP files are available on OSF (https://doi.org/10.17605/OSF.IO/WB4CQ/). Unless stated otherwise, the results did not differ in terms of significance when excluding either participants with mental or neurological conditions (N = 9) or participants who failed compliance (N = 3; Supplement III).

3.1. Confirmatory Analyses on Suppression-Induced Forgetting

In line with our first hypothesis on the SIF effect, recall score differed by word condition ($\chi^2(2)$ = 30.59, p < .001, W = .19; Figure 2). A smaller proportion of No-Think words were correctly recalled compared to baseline (ΔM = 0.03, 95% CI [-0.01, 0.07], $t(158)$ = 2.07, p = .040, r_{rb} = .23) and Think words (ΔM = 0.10, 95% CI [0.06, 0.17], $t(158)$ = 6.02, p < .001, r_{rb} = .80). The Bayesian one-tailed paired samples t-test for No-Think vs. baseline revealed anecdotal evidence for no effect (BF_{10} = 0.89), but extreme evidence in favor of the alternative hypothesis for the difference between No-Think and Think recall (BF_{10} = 1077*10^6). Moreover, Think words were better recalled than baseline words (ΔM = -0.07, 95% CI [-0.11, -0.03], $t(158)$ = 3.94, p < .001, r_{rb} = -.64, BF_{10} = 646.02). When excluding participants with mental or neurological conditions, the difference between No-Think and baseline recall was no longer significant (ΔM = 0.03, 95% CI [-0.02, 0.07], $t(140)$ = 1.86, p = .066, r_{rb} = .19, BF_{10} = 0.62). However, since the effect size remained almost unchanged, we speculate that this non-significant result could be due to reduced power.

We also analyzed the data from negative and neutral word pairs separately. For negative word pairs, No-Think recall did not significantly differ from baseline words (main effect word condition: $\chi^2(2)$ = 15.50, p < .001, W = .10; post

Table 1. Recall scores and recall delays per word condition

	Baseline	No-Think	Think
Recall score (proportion)	.89 ± .12	.87 ± .13	.96 ± .08
Negative	.92 ± .13	.90 ± .14	.96 ± .12
Neutral	.87 ± .19	.83 ± .19	.97 ± .08
Recall delay (s)	0.08 ± 0.51	0.35 ± 1.28	-0.35 ± 0.67
Negative	0.06 ± 0.56	0.32 ± 1.58	-0.24 ± 0.86
Neutral	0.12 ± 0.86	0.38 ± 1.28	-0.45 ± 0.82

Note. Values are reported as $M \pm SD$. Negative: Negative word pairs only; Neutral: Neutral word pairs only.

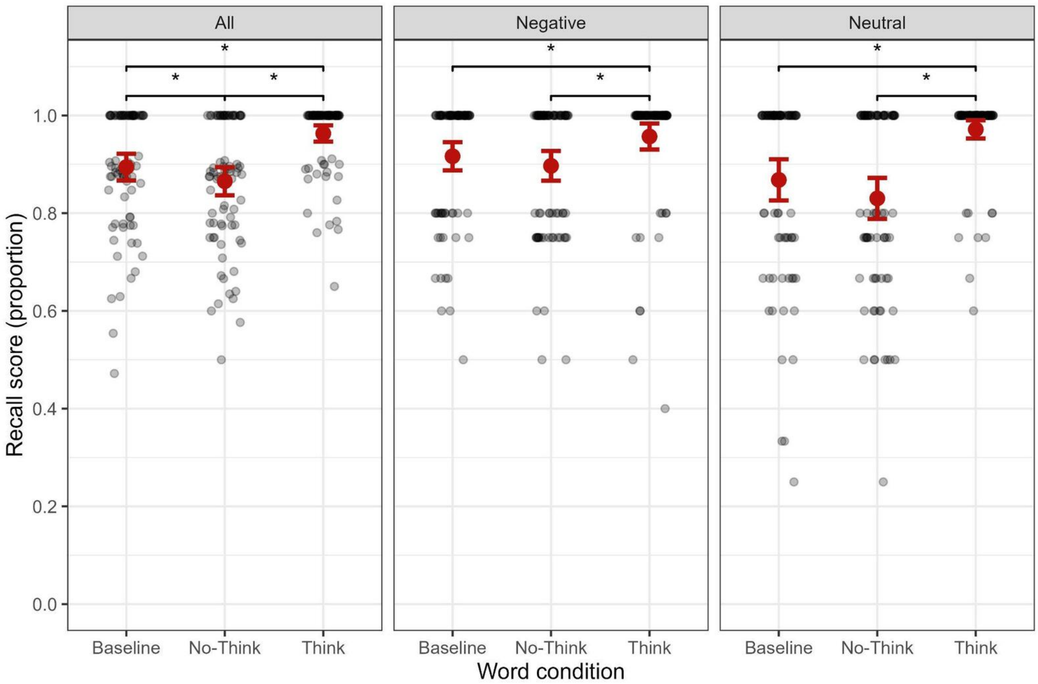

Figure 2. Individual mean recall scores per word condition for all (conditionalized) word pairs, and negative and neutral word pairs, separately.

Note: The red line depicts the mean and 95% confidence interval.

hoc test: ΔM = 0.02, 95% CI [-0.03, 0.07], $t(158)$ = 1.60, p = .112, r_{rb} = .16), but was significantly lower than Think words (ΔM = 0.06, 95% CI [0.01, 0.11], $t(158)$ = 4.09, p <.001, r_{rb} = .54). The Bayesian analysis indicated anecdotal evidence for no difference between No-Think and baseline words (BF_{10} = 0.34), but strong evidence in favor of the alternative hypothesis for the difference between No-Think and Think recall (BF_{10} = 15.66). The difference between Think and baseline recall was significant (ΔM = -0.04, 95% CI [-0.09, 0.01], $t(158)$ = 2.49, p = .028, r_{rb} = -.36, BF_{10} = 1.82). When excluding participants who failed compliance, the difference between Think and baseline recall was no longer significant (ΔM = -0.04, 95% CI [-0.09, 0.01], $t(152)$ = 2.23, p = .055, r_{rb} = -.34, BF_{10} = 1.19).

For neutral word pairs, recall of No-Think words did not significantly differ from baseline (main effect word condition: $\chi^2(2)$ = 31.31, p < .001, W = .20; post hoc test: ΔM = 0.04, 95% CI [-0.03, 0.11], $t(158)$ = 1.61, p = .109, r_{rb} = .21), but it was significantly lower than recall of Think words (ΔM = 0.14, 95% CI [0.09, 0.20], $t(158)$ = 5.99, p < .001, r_{rb} = .90). The Bayesian analysis indicated anecdotal evidence for no difference between No-Think and baseline words (BF_{10} = 0.53), but extreme evidence in favor of the alternative hypothesis for the difference between No-Think and Think recall (BF_{10} = 2488*10^6). Further, Think words were better recalled than baseline words (ΔM = -0.10, 95% CI [-0.16, -0.05], $t(158)$ = 4.38, p < .001, r_{rb} = -.79, BF_{10} = 2586.04).

Comparing the recall of neutral vs. negative words using Wilcoxon signed-rank tests revealed that negative baseline words were better recalled than neutral baseline words (z = -1.97, p = .049, r_{rb} = -.35), negative No-Think words were better recalled than neutral No-Think words (z = -2.89, p = .004, r_{rb} = -.48), and Think recall did not differ between neutral and negative words (z = 0.81, p = .429, r_{rb} = .22).

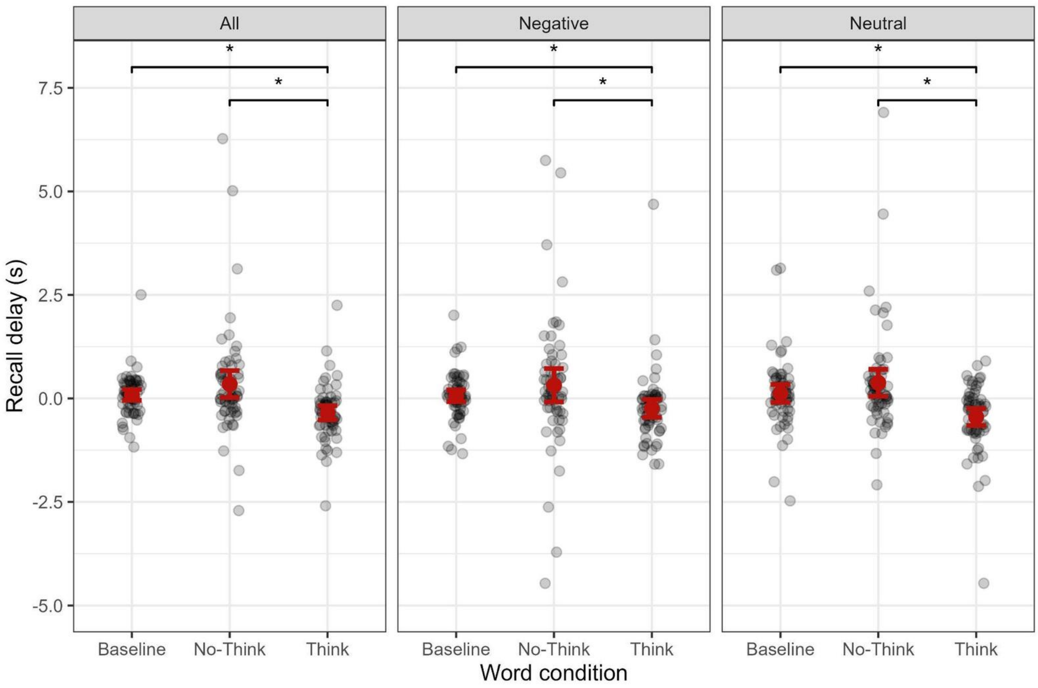

Figure 3. Individual mean recall delays per word condition for all (conditionalized) word pairs, and negative and neutral words, separately.

Note: Recall delay = (Response time during Final Recall phase) – (Response time during final Test Feedback phase). A value of 0 indicates no difference in recall response time between both phases. The more negative the value, the faster the response during the Final Recall phase compared to that during the final Test Feedback phase. The red line depicts the mean and 95% confidence interval.

3.2. Confirmatory Analyses on Suppression-Induced Recall Delay

Recall delay differed by word condition ($\chi^2(2) = 33.52$, $p < .001$, $W = .27$; Figure 3), but the difference between No-Think and baseline words was not significant ($\Delta M = -0.26$, 95% CI [-0.65, 0.13], $t(122) = 0.73$, $p = .467$, $r_{rb} = -.20$). Recall delay was smaller for Think words compared to baseline ($\Delta M = 0.43$, 95% CI [0.21, 0.65], $t(122) = 5.42$, $p < .001$, $r_{rb} = .68$) and No-Think words ($\Delta M = -0.69$, 95% CI [-1.12, -0.26], $t(122) = 6.15$, $p < .001$, $r_{rb} = -.67$). The Bayesian one-tailed paired samples t-test for No-Think vs. baseline revealed anecdotal evidence for no effect ($BF_{10} = 0.94$), but extreme evidence in favor of the alternative hypothesis for the difference between Think and baseline ($BF_{10} = 3953.89$) and Think and No-Think ($BF_{10} = 232.42$). Thus, the data do not confirm a suppression-induced response delay, but only a retrieval practice effect on later recall time.

Once again, we analyzed the data from negative and neutral word pairs separately. For negative word pairs, No-Think recall delay did not significantly differ from baseline words (main effect word condition: $\chi^2(2) = 21.32$, $p < .001$, $W = .17$; post hoc test: $\Delta M = -0.26$, 95% CI [-0.76, 0.25], $t(122) = 1.08$, $p = .284$, $r_{rb} = -.26$, $BF_{10} = 0.51$), but it was significantly different from Think words ($\Delta M = -0.56$, 95% CI [-1.09, -0.02], $t(122) = 4.80$, $p <.001$, $r_{rb} = -.50$, $BF_{10} = 5.64$). The difference between Think and baseline recall delay was significant ($\Delta M = 0.30$, 95% CI [0.02, 0.59], $t(122) = 3.72$, $p < .001$, $r_{rb} = 0.57$, $BF_{10} = 6.33$).

For neutral word pairs, recall delay for No-Think words did not significantly differ from baseline (main effect word condition: $\chi^2(2) = 24.81$, $p < .001$, $W = .20$; post hoc test: $\Delta M = -0.26$, 95% CI [-0.71, 0.20], $t(122) = 0.20$, $p = .842$, $r_{rb} = -.09$), but it was significantly different from Think recall ($\Delta M = -0.83$, 95% CI [-1.27, -0.38], $t(122) = 4.68$, $p < .001$, $r_{rb} = -.70$). The Bayesian analysis indicated anecdotal evidence for no difference between No-Think and baseline words ($BF_{10} = 0.63$), but extreme evidence in favor of the alternative hypothesis for the difference between No-Think and Think recall ($BF_{10} = 1628.74$). Once again, Think words were faster recalled than baseline words ($\Delta M = 0.57$, 95% CI [0.21, 0.93], $t(122) = 4.88$, $p < .001$, $r_{rb} = .57$, $BF_{10} = 183.55$).

Comparing recall delay of neutral vs. negative words using Wilcoxon signed-rank tests revealed no significant differences for baseline ($z = 0.78$, $p = .441$, $r_{rb} = .11$), Think ($z = -1.62$, $p = .107$, $r_{rb} = -.24$), or No-Think words ($z = -.33$, $p = .742$, $r_{rb} = -.05$).

3.3. Exploratory Analyses on Word Emotionality and Strategy

Suppression-induced forgetting did not differ between neutral and negative words ($F(1,78) = 0.26$, $p = .611$) or suppression strategy (Thought Avoidance vs. Thought Substitution; $F(1,78) = 0.29$, $p = .593$). The facilitation effect (better recall of Think vs. baseline words) was larger for neutral compared to negative words ($F(1,78) = 6.29$, $p = .014$, $\eta 2 = .03$; $\Delta M = -0.14$, 95% CI [-0.25, -0.03], $t(79) = -2.51$, $p = .014$, $d = -0.37$), probably due to worse baseline recall for neutral words (section 3.1.). The facilitation effect was not affected by suppression strategy (F(1,78) = 0.38, p = .542).

For recall delay, SIF score and retrieval index were not affected by either word emotionality or suppression strategy (p's > .05 for all).

4. Discussion

The current study was inspired by occasional null findings in the TNT literature and recent work investigating the robustness and replicability of the SIF effect. We reasoned that the commonly used recall score, that is based on a binary measure (remembered or forgotten), might be too insensitive to detect the effects of suppression on later recall, thereby possibly explaining any null findings. Instead, recall delay may be a more sensitive measure, and we tested this in a typical TNT task in a large sample. However, we only observed a small SIF effect, and no evidence for suppression-induced delay of recall time. Contrastingly, retrieval practice sped up later recall.

Fewer No-Think words were correctly recalled compared to baseline during the same-probe recall test. The effect size was small (r_{rb} = .23), and the Bayes factor indicated anecdotal evidence for the null hypothesis (BF_{10} = 0.89). This result corroborates previous reports regarding a small SIF effect (e.g., Stramaccia et al., 2021), and casts further doubt over the robustness of SIF (Wessel et al., 2020, 2025; Wiechert et al., 2023). Optimal test conditions appear to be of critical importance for observing a SIF effect (Zhao et al., 2025). Although we took substantial care to ensure optimal conditions for observing SIF, the test conditions could be further improved; for example, by adding breaks during the TNT phase, including a diagnostic questionnaire, or screening participants for sleep duration (Zhao et al., 2025). Since we did not include a diagnostic questionnaire during the TNT phase, we cannot confirm that participants, especially in the thought substitution condition, followed the instructions precisely. A lack of instruction compliance can weaken the SIF effect (Liu et al., 2021). Participants answered a question regarding compliance in the post-experimental questionnaire that revealed generally high levels of compliance (see Supplement I). Yet, using existing compliance questionnaires that are even more specific can provide better insights into the degree of instruction compliance (see e.g., Liu et al., 2021). Relatedly, a large number of participants (45 out of 80 participants; 56.25%) reported to have expected a final recall test (see Supplement I). Expecting a recall test may have affected the suppression of memory retrieval during the TNT phase and thereby also weakened the SIF effect. However, one study found that whereas test expectation is related to memory checking behaviors during the TNT phase, it was not related to SIF (Liu et al., 2021).

In the current study, retrieval suppression did not slow down later recall time. This finding is in contrast with our hypothesis and previous studies that found suppression-induced recall delay (Cano & Knight, 2016; Nishiyama & Saito, 2021; Noreen & MacLeod, 2013; Schechtman et al., 2021; Waldhauser et al., 2012) or slower reaction time on a recognition test (Benoit et al., 2015; Gagnepain et al., 2017). This discrepancy could be due to differences in experimental design. For example, some studies used pictures (Benoit et al., 2015; Gagnepain et al., 2017) or auditorily presented words (Cano & Knight, 2016) as stimuli. Perhaps more importantly, previous studies generally tested small samples (~20-30 participants) and thus had low statistical power. Statistically significant findings in these studies are therefore less reliable. In contrast, due to our larger sample size, our analyses had more statistical power and consequently, retrieval suppression may, in fact, not slow down later recall. However, we draw this conclusion cautiously, as we had insufficient power to detect small effects.

Participants were instructed to employ either thought avoidance or thought substitution during the No-Think trials. In our main analyses involving SIF and recall delay, we combined the data from both conditions. However, these strategies are presumed to have different underlying processes. While thought avoidance draws on an inhibitory process, thought substitution causes interference (Racsmány et al., 2012). Therefore, combining both strategies in a single analysis may have obscured any theoretically meaningful differences. Our sample was relatively small to be split up, due to limited resources, so the effects of these specific strategies should be investigated in more detail in future studies. Exploratory analyses revealed that SIF and recall delay did not differ between participants who used thought avoidance and those who used thought substitution. Notably, one previous study reported that thought substitution was more effective for forgetting than no specific strategy instruction (Hertel & Calcaterra, 2005). However, in studies that assessed thought substitution (e.g., Hertel & Calcaterra, 2005; Nishiyama & Saito, 2021), participants had to learn specific substitutes instead of deciding upon the substitutes themselves. Learning word triplets (cue, target, and substitute) may be more demanding and therefore lead to stronger interference than direct suppression. The strategies may, however, not necessarily lead to differences in SIF, especially when using a same-probe cued recall test (Racsmány et al., 2012). Future research should use well-powered designs to test the differences between these strategies.

Individual differences in memory suppression ability can predict SIF (Levy & Anderson, 2012), as well as slower recall time (Noreen & MacLeod, 2013). Participants who are better at controlling their awareness during No-Think trials may experience larger forgetting effects later. Future studies could investigate recall delay in relation to memory suppression ability, for example, by assessing the intrusion slope during the TNT phase (Levy & Anderson, 2012). To do so, the trial-by-trial measurement of intrusions needs to be improved to a 3-point scale ("never", "briefly", "often"; see Levy & Anderson, 2012) that participants respond to themselves, in order to reduce any social desirability effects and obtain more accurate responses.

Think words were better and faster recalled than baseline and No-Think words, across all conditionalized word pairs and for neutral and negative word pairs separately. This finding aligns with previous findings regarding TNT (Nishiyama & Saito, 2021; Noreen & MacLeod, 2013) and the retrieval practice effect (Roediger & Butler, 2011). By repeatedly retrieving the target word during Think trials in the TNT phase, participants likely improved the cue-target associations and automatized the retrieval process, leading to better and faster recall (Racsmány et al., 2018). Furthermore, the facilitation effect (better recall of Think vs. baseline words) was larger for neutral compared to negative words. Instead of emotionally neutral stimuli being more strongly facilitated by retrieval practice, this effect could be due to baseline differences. Baseline recall for negative words was better than that of neutral words, in line with research on beneficial effects of emotions on memory (e.g., Kensinger et al., 2007). Yet, this baseline difference allowed for larger effects of retrieval practice on neutral recall, whereas the improvement of negative word recall was restricted by a ceiling effect.

A limitation of the current study is the number of word pairs. To allow participants to learn a sufficient number of word pairs within a reasonable time window, we included 30 word pairs, of which at least 24 (=80%) had to be learned. Increasing the number of word pairs could increase the statistical power for finding an effect. For example, SIF was not present for neutral and negative words separately, potentially due to a maximum of only five word pairs per condition. Further studies could consider using a larger number of stimuli. It should be noted, however, that the use of 30 or less stimuli is not uncommon for studies that find a SIF effect (e.g., Ashton et al., 2020; Noreen & MacLeod, 2013). Moreover, more stimuli will increase the difficulty of, and time needed for learning the word pairs, potentially inducing fatigue, boredom, and/or frustration in the participant and leading to more exclusions due to insufficient learning. These factors could, in turn, deteriorate test conditions and affect SIF. A potential limitation in our measurement of recall delay is that we did not instruct participants to respond to the cue words in the Final Recall Test as quickly as possible. Although we assume that participants would not withhold their responses after retrieving them from memory, we cannot exclude the possibility that omitting this instruction may have influenced the response time results. Further research should test the effects of specific recall instructions.

To conclude, we tested recall delay as a potentially more sensitive measure of memory suppression than the typically used recall score. In a large sample, we did not find evidence for suppression-induced recall delay, but only for the retrieval practice effect. Moreover, the SIF effect for recall scores was small and not supported by Bayesian analyses. These findings should be interpreted with some caution, however, as we had insufficient statistical power to detect small effect sizes. Nevertheless, our findings support growing concerns regarding the robustness of suppression effects on later memory recall.

Author Contributions

Lianne N. Wolsink: Conceptualization, Methodology, Validation, Data curation, Formal analysis, Investigation, Project administration, Visualization, Writing - original draft, Writing - review and editing. **George Jacob:** Conceptualization, Methodology, Validation, Data curation, Formal analysis, Writing - review and editing. **Christian J. Merz:** Conceptualization, Supervision, Writing - review and editing.

Acknowledgements

We would like to thank Sascha Schmitz and Christian Weitz for collecting pilot data, Cara Göhl for literature research, Cara Göhl, Angelina Hoßfeld, Dimitrios Kiriakidis, and Sofia Siarkou for help with data collection, and Carla Evers, Christina Fast, Angelina Hoßfeld, and Coraly Wollstein for help with processing of the audio data.

Funding Information

All authors are funded by the Deutsche Forschungsgemeinschaft (DFG; German Research Foundation) within SFB 1280 Extinction Learning (grant no. 316803389 - SFB1280).

Ethics Statement

All procedures were in accordance with the Declaration of Helsinki and approved by the ethics committee of the Faculty of Psychology at Ruhr University Bochum (reg. nr. 847).

Competing Interests

We declare no competing interests.

Data Accessibility

All stimulus material, instructions, participant data, and JASP files can be found on OSF (https://doi.org/10.17605/OSF.IO/WB4CQ).

Editors: Peter Verkoeijen (Associate Editor)

Submitted: October 02, 2025 PDT. Accepted: May 12, 2026 PDT. Published: June 25, 2026 PDT.

References

Anderson, M. C., & Green, C. (2001). Suppressing unwanted memories by executive control. *Nature*, *410*, 366–369. https://doi.org/10.1038/35066572

Ashton, S. M., Benoit, R. G., & Quaedflieg, C. W. (2020). The impairing effect of acute stress on suppression-induced forgetting of future fears and its moderation by working memory capacity. *Psychoneuroendocrinology*, *120*, 104790. https://doi.org/10.1016/j.psyneuen.2020.104790

Benoit, R. G., Hulbert, J. C., Huddleston, E., & Anderson, M. C. (2015). Adaptive top–down suppression of hippocampal activity and the purging of intrusive memories from consciousness. *Journal of Cognitive Neuroscience*, *27*(1), 96–111. https://doi.org/10.1162/jocn_a_00696

Bergström, Z. M., de Fockert, J. W., & Richardson-Klavehn, A. (2009). ERP and behavioural evidence for direct suppression of unwanted memories. *NeuroImage*, *48*(4), 726–737. https://doi.org/10.1016/j.neuroimage.2009.06.05

Bulevich, J. B., Roediger, H. L., Balota, D. A., & Butler, A. C. (2006). Failures to find suppression of episodic memories in the think/no-think paradigm. *Memory & Cognition*, *34*(8), 1569–1577. https://doi.org/10.3758/BF03195920

Cano, M. E., & Knight, R. T. (2016). Behavioral and EEG evidence for auditory memory suppression. *Frontiers in Human Neuroscience*, *10*, 133. https://doi.org/10.3389/fnhum.2016.00133

Clark, C., Todorovic, D., Levy, B. J., Eschmann, K. C. J., Klein, M., Anderson, M., ... Fawcett, J. M. (2025). Forgetting as a Consequence of Retrieval Suppression: A Meta-Analytic Review. *PsyArXiv*. https://doi.org/10.31234/osf.io/zd6ah_v1

Ehlers, A., & Clark, D. M. (2000). A cognitive model of posttraumatic stress disorder. *Behaviour Research and Therapy*, *38*(4), 319–345. https://doi.org/10.1016/S0005-7967(99)00123-0

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Gagnepain, P., Hulbert, J., & Anderson, M. C. (2017). Parallel regulation of memory and emotion supports the suppression of intrusive memories. *Journal of Neuroscience*, *37*(27), 6423–6441. https://doi.org/10.1523/JNEUROSCI.2732-16.2017

Haghighi, S. S., Ghorbani, M., Dehnavi, F., Safaie, M., & Moghimi, S. (2020). Motivated forgetting increases the recall time of learnt items: Behavioral and event related potential evidence. *Brain Research*, *1729*, 146624. https://doi.org/10.1016/j.brainres.2019.146624

Hertel, P. T., & Calcaterra, G. (2005). Intentional forgetting benefits from thought substitution. *Psychonomic Bulletin & Review*, *12*(3), 484–489. https://doi.org/10.3758/BF03193792

JASP Team. (2024). *JASP* (Version 0.19.1) [Computer software].

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford University Press.

Kensinger, E. A., Garoff-Eaton, R. J., & Schacter, D. L. (2007). Effects of emotion on memory specificity in young and older adults. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *62*(4), P208–P215. https://doi.org/10.1093/geronb/62.4.P208

Levy, B. J., & Anderson, M. C. (2012). Purging of memories from conscious awareness tracked in the human brain. *Journal of Neuroscience*, *32*(47), 16785–16794. https://doi.org/10.1523/JNEUROSCI.2640-12.2012

Liu, P., Hulbert, J. C., Yang, W., Guo, Y., Qiu, J., & Anderson, M. C. (2021). Task compliance predicts suppression-induced forgetting in a large sample. *Scientific Reports*, *11*(1), 20166. https://doi.org/10.1038/s41598-021-99806-8

Mecklinger, A., Parra, M., & Waldhauser, G. T. (2009). ERP correlates of intentional forgetting. *Brain Research*, *1255*, 132–147. https://doi.org/10.1016/j.brainres.2008.11.073

Murray, B. D., Muscatell, K. A., & Kensinger, E. A. (2011). Effects of emotion and age on performance during a think/no-think memory task. *Psychology and Aging*, *26*(4), 940. https://doi.org/10.1037/a0023214

Nardo, D., & Anderson, M. C. (2024). Everything you ever wanted to know about the Think/No- Think task, but forgot to ask. *Behavior Research Methods*, *56*(4), 3831–3860. https://doi.org/10.3758/s13428-024-02349-9

Nishiyama, S., & Saito, S. (2021). Transferable inhibition of direct suppression: Evidence from a dot-probe task. *Psychologia*, *63*, 20–41. https://doi.org/10.2117/psysoc.2020-A112

Nørby, S., Lange, M., & Larsen, A. (2010). Forgetting to forget: On the duration of voluntary suppression of neutral and emotional memories. *Acta Psychologica*, *133*(1), 73–80. https://doi.org/10.1016/j.actpsy.2009.10.002

Noreen, S., & MacLeod, M. D. (2013). It's all in the detail: intentional forgetting of autobiographical memories using the autobiographical think/no-think task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(2), 375. https://doi.org/10.1037/a0028888

Racsmány, M., Conway, M. A., Keresztes, A., & Krajcsi, A. (2012). Inhibition and interference in the think/no-think task. *Memory & Cognition*, *40*(2), 168–176. https://doi.org/10.3758/s13421-011-0144-6

Racsmány, M., Szőllősi, Á., & Bencze, D. (2018). Retrieval practice makes procedure from remembering: An automatization account of the testing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(1), 157. https://doi.org/10.1037/xlm0000423

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20–27. https://doi.org/10.1016/j.tics.2010.09.003

Schechtman, E., Lampe, A., Wilson, B. J., Kwon, E., Anderson, M. C., & Paller, K. A. (2021). Sleep reactivation did not boost suppression-induced forgetting. *Scientific Reports*, *11*(1), 1383. https://doi.org/10.1038/s41598-020-80671-w

Schmidt, A. C., Waldhauser, G. T., Hestermann, F., Kehyayan, A., Kessler, H., & Axmacher, N. (2022). Testing the relationship between empirical paradigms of repression and suppression. *Psychoanalytic Psychology*, *39*(3), 244–252. https://doi.org/10.1037/pap0000407

Stramaccia, D. F., Meyer, A. K., Rischer, K. M., Fawcett, J. M., & Benoit, R. G. (2021). Memory suppression and its deficiency in psychological disorders: A focused meta-analysis. *Journal of Experimental Psychology: General*, *150*(5), 828. https://doi.org/10.1037/xge0000971

The MathWorks Inc. (2024). *MATLAB version: 24.2.0.2773142 (R2024b)*. The MathWorks Inc. https://www.mathworks.com

van Schie, K., Geraerts, E., & Anderson, M. C. (2013). Emotional and non-emotional memories are suppressible under direct suppression instructions. *Cognition & Emotion*, *27*(6), 1122–1131. https://doi.org/10.1080/02699931.2013.765387

Võ, M. L., Conrad, M., Kuchinke, L., Urton, K., Hofmann, M. J., & Jacobs, A. M. (2009). The Berlin affective word list reloaded (BAWL-R). *Behavior Research Methods*, *41*(2), 534–538. https://doi.org/10.3758/BRM.41.2.534

Waldhauser, G. T., Lindgren, M., & Johansson, M. (2012). Intentional suppression can lead to a reduction of memory strength: behavioral and electrophysiological findings. *Frontiers in Psychology*, *3*, 401. https://doi.org/10.3389/fpsyg.2012.00401

Wessel, I., Albers, C. J., Zandstra, A. R. E., & Heininga, V. E. (2020). A multiverse analysis of early attempts to replicate memory suppression with the think/no-think task. *Memory*, *28*(7), 870–887. https://doi.org/10.1080/09658211.2020.1797095

Wessel, I., Lehmann, R., ... Wiechert, S. (2025). Two replications of Wiechert et al.'s (2023) online Think/No-Think study in undergraduate students. *Memory*, *33*(1), 134–144. https://doi.org/10.1080/09658211.2024.2412025

Wiechert, S., Loewy, L., Wessel, I., Fawcett, J. M., Ben-Shakhar, G., Pertzov, Y., & Verschuere, B. (2023). Suppression-induced forgetting: A pre-registered replication of the think/no-think paradigm. *Memory*, *31*(7), 989–1002. https://doi.org/10.1080/09658211.2023.2208791

Zhao, J., Verschuere, B., & Wiechert, S. (2025). Can People Really Suppress Unwanted Thoughts? On optimal test conditions and authorship effects in the Think/No-Think (TNT) Task. *PsyArXiv*. https://doi.org/10.31234/osf.io/gqnyd_v1