Memory & Cognition (2022) 50:1061–1077
https://doi.org/10.3758/s13421-021-01258-y

Soliciting judgments of forgetting reactively enhances memory as well as making judgments of learning: Empirical and meta-analytic tests

Baike Li[1] · Wenbo Zhao[2] · Jun Zheng[2] · Xiao Hu[2] · Ningxin Su[2] · Tian Fan[2] · Yue Yin[2] · Meng Liu[3] · Chunliang Yang[1,4] · Liang Luo[1,2]

Accepted: 8 November 2021 / Published online: 2 December 2021
© The Psychonomic Society, Inc. 2022

Abstract

Recent studies found that making judgments of learning (JOLs) can reactively facilitate memory, a phenomenon termed the *reactivity effect* of JOLs. The current study was designed to explore (1) whether making judgments of forgetting (JOFs) can also enhance memory and (2) whether there is any difference between the reactivity effects of JOFs and JOLs. Experiment 1 found that soliciting JOFs significantly enhanced retention of single words. Experiments 2 and 3 observed minimal difference in reactivity effects between JOFs and JOLs on learning of single words and word pairs. Finally, a meta-analysis was conducted to integrate results across studies to explore whether retention of items studied with JOLs differed from that of items studied with JOFs. The meta-analytic results showed minimal difference. Overall, the documented findings imply that (1) making JOFs reactively enhances memory, and (2) there is little difference in reactivity effects between JOFs and JOLs. These findings support the positive-reactivity theory to account for the reactivity effect.

Keywords Judgments of forgetting · Judgments of learning · Reactivity effect · Memory · Meta-analysis

Introduction

Metamemory consists of two key components: metamemory monitoring and metamemory control (Nelson & Leonesio, 1996; Nelson & Narens, 1994). It is well known that metamemory monitoring can affect learning and memory in an indirect way through its influences on metamemory control. For instance, according to *discrepancy-reduction* models of self-regulated learning (Dunlosky & Hertzog, 1997; Nelson & Narens, 1994; Verkoeijen et al., 2005), before study learners set a desired learning goal, and during study, they continuously monitor their ongoing learning progress. When a gap is detected between the perceived and desired learning status, further efforts (e.g., more study time) will be expended toward narrowing the perceived gap.

Besides this indirect influence, an emerging body of recent studies found that making a metamemory judgment can also affect learning and memory in a direct way. For instance, recent studies found that instructing learners to make item-by-item judgments of learning (JOLs; metacognitive judgments about the likelihood of remembering a studied item in a future memory test) can alter (typically enhance) memory itself, a phenomenon referred to as the *reactivity* effect, because making JOLs reactively changes memory (Double et al., 2018; Double & Birney, 2019; Dougherty et al., 2018; Janes et al., 2018; Myers et al., 2020; Rivers, 2018).

Akin to JOLs, judgments of forgetting (JOFs) are another form of metamemory judgment about the likelihood of forgetting a studied item in a future memory test (Chen et al., 2016; Finn, 2008; Koriat et al., 2004; Serra & England, 2012, 2019). Although a variety of studies have been conducted to explore the reactive influences of JOLs on memory, so far (to our knowledge) no research has been

✉ Chunliang Yang
chunliang.yang@bnu.edu.cn

[1] Institute of Developmental Psychology, Faculty of Psychology, Beijing Normal University, 19 Xinjiekou Wai Street, Beijing 100875, China

[2] Collaborative Innovation Center of Assessment for Basic Education Quality, Beijing Normal University, Beijing, China

[3] School of Psychology, South China Normal University, Guangzhou, China

[4] Beijing Key Laboratory of Applied Experimental Psychology, National Demonstration Center for Experimental Psychology Education, Beijing Normal University, Beijing, China

conducted to investigate whether JOFs can also induce a reactivity effect. Hence, the first aim of the current study was to fill this gap. It is noteworthy that previous findings about the similarities and differences between JOLs and JOFs are largely inconsistent, and it remains unknown whether JOLs and JOFs are distinct types of metamemory judgments (see below for detailed discussion). Hence, the second aim of the current study was to further explore this critical question by exploring whether JOLs and JOFs have different reactivity effects on memory.

Below we briefly summarize empirical findings and potential mechanisms underlying the reactivity effect of JOLs, then discuss potential differences between JOLs and JOFs, and finally provide an overview of the current study.

Reactivity effect of judgments of learning (JOLs)

Recent studies documented that making JOLs can reactively change memory (Janes et al., 2018; Mitchum et al., 2016; Soderstrom et al., 2015). For instance, in Soderstrom et al.'s (2015) Experiment 1, participants were randomly divided into two groups (JOLs vs. no-JOL) and were instructed to study strongly related (e.g., *blunt-sharp*) and weakly related (e.g., *boxer-terrible*) word pairs. The total exposure time for each word pair was 8 s in both the JOL and the no-JOL groups. Specifically, for the JOL group, a word pair was first presented for 4 s for study, and then this pair continued appearing on-screen for another 4 s, during which participants were required to make a JOL. By contrast, for the no-JOL group, each pair appeared on screen for 8 s in total, and participants did not need to make JOLs. In a later cued recall test, the JOL group recalled significantly more strongly related pairs and numerically more weakly related pairs than the no-JOL group, demonstrating a positive reactivity effect (for related findings, see Double et al., 2018; Janes et al., 2018; Mitchum et al., 2016; Rivers, 2018; Schmoeger et al., 2020; C. Yang et al., 2021). In a meta-analysis, Double et al. (2018) found that there is a small-to-medium enhancing effect of making JOLs on memory of related word pairs (Hedges' $g = 0.323$) and word lists (Hedges' $g = 0.384$). Furthermore, another meta-analysis conducted by C. Yang et al. (2021), which included a larger set of data and more recent studies, also found an overall positive reactivity effect of making JOLs on retention of word pairs and word lists.

Several theories have been proposed to explain why making JOLs reactively affects memory itself (for a detailed discussion about relevant theories, see C. Yang et al., 2021). For instance, a *positive reactivity* theory proposes that making item-by-item JOLs may motivate participants to exert greater encoding effort, adopt more effective encoding strategies, and engage in more elaborative processing, leading to enhanced retention and a positive reactivity effect (Mitchum et al., 2016; Rivers, 2018). For instance, to provide a JOL for each item, participants have to sustain their attention across the learning task. In addition, to make an appropriate JOL for each item, participants have to search for "diagnostic cues" to inform JOL formation, and the cue searching process may in turn induce more elaborative processing. Overall, the positive reactivity theory mainly hypothesizes that soliciting item-by-item JOLs benefits memory through enhancing learning engagement, refining study strategies, and inducing more elaborative processing.

Indeed, a set of recent findings provides support to the positive reactivity theory. For instance, Sahakyan et al. (2004) found that asking participants to make a JOL (i.e., predicting the number of words they would remember in a later memory test) following the study of a list of words caused them to shift from poor learning strategies (e.g., rote rehearsal) to more effective ones during study of the subsequent list. Tekin and Roediger (2020) found that the reactivity effect interacted with the *level of processing* effect. Specifically, they found that items receiving shallow processing (e.g., perceptual judgment) exhibited a larger positive reactivity effect than those receiving deep processing (e.g., semantic judgment), suggesting that the reactivity effect may result from the fact that making JOLs induces more elaborative processing.

The relation of the positive reactivity theory to the current study is discussed below.

Judgments of forgetting (JOFs) versus JOLs

Many studies have been conducted to explore whether JOLs and JOFs are distinct forms of metamemory judgments (England et al., 2017; Finn, 2008; Koriat et al., 2004; Rhodes & Castel, 2008; Schmoeger et al., 2020; Serra & England, 2012, 2019; Tauber & Rhodes, 2012). In these studies, many aspects of JOLs and JOFs have been compared, but the research findings are largely inconsistent or even conflicting.

For instance, Koriat et al. (2004) found that, compared with JOLs, JOFs are more sensitive to retention interval (i.e., the interval between study and test). Specifically, Koriat and colleagues asked participants to imagine a hypothetical student who studied 60 word pairs and to predict how many items they thought that participant would either remember or forget in a test administered after 10 min, 1 day, and 1 week. The results showed that participants predicted that that person would remember equal number of word pairs after 10 min, 1 day, and 1 week, reflecting that JOLs are relatively insensitive to retention interval. By contrast, participants predicted that that person would forget more word pairs after a long retention interval (e.g., 1 week) than after a short interval (e.g., 10 min), reflecting that JOFs are sensitive to retention interval. These findings imply that JOLs and JOFs differ in their sensitivity to retention interval. However,

Serra and England (2019) recently failed to replicate Koriat al.'s (2004) findings by showing that both JOLs and JOFs are insensitive to retention interval.

Finn (2008) explored whether there is any difference in remembering confidence and absolute accuracy (i.e., signed difference between judgments and test performance) between JOLs and JOFs. They instructed two groups of participants to study word pairs, with a JOL group making a JOL after studying each pair and a JOF group making a JOF for each pair. To compare remembering confidence between the JOL and JOF groups, Finn reversed JOF scores (i.e., 100 – JOF) to make JOFs and JOLs comparable. The results showed that participants in the JOF group were overall less confident in their memory ability than those in the JOL group. In addition, absolute accuracy was superior in the JOF group than in the JOL group. However, in a subsequent study, Serra and England (2012) failed to replicate Finn's (2008) main findings by showing a minimal difference in remembering confidence and absolute accuracy between JOLs and JOFs.

Previous studies also explored whether JOLs and JOFs differ in relative accuracy (also known as judgment resolution), which is typically calculated as an intra-individual Gamma correlation between judgments and test performance. Along the same lines, the research findings are substantially inconsistent. For instance, Chen et al. (2016) found that relative accuracy of JOLs was better than that of JOFs (also see England et al., 2017; Serra & England, 2012). However, Finn (2008) did not detect any reliable difference in relative accuracy between JOLs and JOFs.

In summary, many studies have been conducted to explore whether JOLs and JOFs are distinct forms of metamemory judgments through investigating whether JOLs and JOFs differ in their sensitivity to study-test interval, remembering confidence, absolute accuracy, and relative accuracy. However, their research findings are largely inconclusive and conflicting, and it remains unknown whether JOLs and JOFs are different in nature. Hence, the current study is aimed at further exploring this question by investigating whether JOLs and JOFs have different reactivity effects on memory. To our knowledge, the question regarding whether JOLs and JOFs exert different reactivity effects on memory has never been explored before. Below we provide discussions about protentional difference between the reactivity effects of JOLs and JOFs.

Potential difference in reactivity between JOLs and JOFs

As discussed above, the positive reactivity theory assumes that the positive reactivity effect derives from the fact that making item-by-item JOLs enhances participants' engagement in the ongoing learning task and induces study strategy changes (and more elaborative processing). Accordingly, it is reasonable to expect that making JOFs can induce positive reactivity as well as making JOLs because making both kinds of metamemory judgments can maintain attention and induce elaborative processing. Overall, according to the positive reactivity theory, we expect to observe (1) a positive reactivity effect of JOFs and (2) a minimal difference between the reactivity effects of JOLs and JOFs.

However, according to the *framing* explanation (Finn, 2008), we expect to observe a larger reactivity effect of JOFs compared to that of JOLs. As discussed above, Finn (2008) observed that participants were less confident in their memory ability when they were prompted to predict the likelihood of forgetting a given item on an upcoming test than when they were instructed to predict the likelihood of remembering a study item. Furthermore, Finn's (2008) Experiment 2 observed that because participants were less confident in their memory ability in the JOF condition, they selected more items to restudy than they did in the JOL condition. Finn (2008) proposed the framing explanation to account for their findings. Specifically, Finn (2008) hypothesized that framing the judgment question in terms of forgetting activates beliefs about forgetting and makes participants more sensitive to the fallibility of their memories. Hence, the forgetting frame debases participants' subjective confidence about how well they have memorized the study items, and the diminished confidence, in turn, drives them to expand greater study effort (e.g., selecting more items to restudy). Accordingly, based on the framing explanation, we expect to observe a larger reactivity effect of JOFs than that of JOLs, because making JOFs, compared with making JOLs, reduces memory confidence and motivates greater study effort.

In summary, the positive reactivity theory expects little difference between the reactivity effects of JOLs and JOFs. By contrast, the framing explanation predicts a larger reactivity effect of JOFs than that of JOLs. Hence, through investigating the difference between the reactivity effects of JOLs and JOFs, the current study also aims to test the positive reactivity theory and the framing explanation.

Overview of the current study

The current study aimed to: (1) determine whether making JOFs can reactively affect memory, (2) investigate whether JOLs and JOFs differ in their reactivity effects on memory, and (3) test the positive reactivity theory and the framing explanation. The first question was explored in Experiment 1, in which half the words were studied with concurrent JOFs and the other half were studied without JOFs. The reactivity effect of JOFs on memory was quantified as the difference in recognition performance between JOF and no-JOF words. To foreshadow, Experiment 1 documented a positive reactivity effect of JOFs on learning of single words.

The second and third questions were investigated in Experiments 2 and 3, which explored whether JOLs and JOFs exert different reactivity effects on learning of single words (Experiment 2) and word pairs (Experiment 3). To foreshadow, these two experiments consistently observed that soliciting both JOLs and JOFs significantly enhanced retention of single words and word pairs. More importantly, there was minimal difference in their reactivity effects.

Finally, a meta-analysis was conducted to increase statistical power to explore whether JOLs and JOFs have different effects on memory. This meta-analysis again showed little difference in test performance between items studied with JOLs and those studied with JOFs.

Experiment 1

Double et al. (2018) and C. Yang et al. (2021) recently conducted meta-analyses to explore the reactivity effect of JOLs, and they consistently observed that making JOLs significantly enhances retention of single words. Different from previous studies, our Experiment 1 was implemented to explore whether making JOFs can also enhance retention of single words.

Method

Participants

According to a pilot study (Cohen's $d = 0.60$), a power analysis, conducted via G*Power (Faul et al., 2007), showed that 24 participants were required to observe a significant (two-tailed, $\alpha = .05$) reactivity effect of JOFs at 0.80 power. To be more conservative, we decided to increase the sample size to 30. Accordingly, 30 students ($M = 20.833$ years old, $SD = 2.730$; 23 female) were recruited from Beijing Normal University (BNU). They provided informed consent, were tested individually in a sound-proofed cubicle, and received financial remuneration. The protocol was approved by the Institutional Review Board of BNU Faculty of Psychology.

Materials

The stimuli were 352 two-character Chinese words extracted from the Chinese word database developed by Cai and Brysbaert (2010). The word frequency of these words ranged from 0.03 to 19.44 per million. Thirty-two words were used for practice and the other 320 words were used in the formal experiment. For each participant, half of these 320 words were randomly selected by computer to present in the study phase, which also served as "old" items in the recognition test, with the other half of the words serving as "new" items.

To prevent any item-selection effects, for each participant, the 160 to-be-studied words were randomly divided into four lists, with 40 words in each list. Two lists were randomly assigned to the JOF condition and the other two lists were divided into the no-JOF condition. In addition, the present sequence of words in each list and list sequence were randomly decided by computer for each participant. All stimuli were presented via the Matlab *Psychtoolbox* package (Kleiner et al., 2007).

Design and procedure

Experiment 1 involved a within-subjects design (JOF vs. no-JOF). Participants were informed that they would study four lists of words in preparation for a later memory test. For two lists, they would be asked to predict the likelihood of forgetting each word in a later memory test, and they would not need to make such predictions for the other two lists of words. Importantly, they were informed that they should remember all words equally well regardless of whether they had to make predictions or not because all of them would be finally tested.

The procedure was adapted from Soderstrom et al. (2015). Before the formal experiment, participants completed a practice task to familiarize themselves with the experimental procedure. The procedure of the practice task was the same as that of the main experiment (see below for details).

In the formal experiment, participants studied four lists of words, with 40 words in each list. Before studying each list, the computer informed participants whether or not they would need to make forgetting predictions for the following list of words. In a no-JOF list, the 40 words were presented one by one in random order. Before the presentation of each word, a cross sign appeared at the center of the screen for 0.5 s to mark the inter-stimulus interval. Immediately following, a word appeared on-screen for 6 s in total. Then, the next trial started. This cycle repeated until the end of the list, with a new word studied in each cycle.

The procedure for the JOF lists was similar to that for the no-JOF lists, but with several differences. Specifically, in the JOF lists, each word was first presented for 3 s, following which the word remained on-screen for another 3 s with a slider presented below it. Participants were instructed to predict how likely it was that they would forget the word in a later memory test. Their predictions were made on a slider ranging from 0 (*Sure I will not forget it*) to 100 (*Sure I will forget it*). The scale was presented for 3 s, and participants made their JOFs by dragging and clicking the scale pointer. If they successfully made a JOF within the 3-s time window, the word remained on screen for the left duration of the 3 s to ensure that the total exposure duration for each word was 6 s. If they did not successfully make a JOF during the required time window, a message box appeared to remind them to

Table 1 *M*s (*SD*s) for hit rates, false alarm (FA) rates, d', and c' in Experiments 1 and 2

Experiment	Hit		FA	d'		c'
	Judgement	No-judgement		Judgement	No-judgement	
Experiment 1	0.837 (0.104)	0.700 (0.216)	0.127 (0.092)	2.302 (0.578)	1.882 (0.842)	1.238 (0.414)
Experiment 2						
JOF group	0.822 (0.118)	0.702 (0.196)	0.115 (0.097)	2.413 (0.679)	2.005 (0.996)	1.387 (0.581)
JOL group	0.857 (0.092)	0.699 (0.173)	0.083 (0.061)	2.662 (0.668)	2.089 (0.747)	1.497 (0.412)

carefully make predictions for the following words during the required time window. Participants clicked the mouse to remove the message box and to trigger the next trial.

After participants studied all four lists of words, they solved math problems (e.g., *17 + 38= ___?*) for 5 min, which served as a distractor task. Then all participants completed an old/new recognition test. The 160 studied (old) and 160 new words were presented one by one in a random order. Participants were instructed to judge whether the on-screen word was "old" (studied) or "new." If the word was "old," keycode "Z" should be pressed. Otherwise, the "M" keycode should be pressed. There was no time pressure and no feedback in the recognition test.

Results and discussion

The primary research interest was the reactivity effect of JOFs, and hence the results of test performance are presented below. Item-by-item JOFs were not the focus of the current study and hence are reported in Appendix 1.

Table 1 lists hit rates, false-alarm (FA) rates, discriminability (d', an index reflecting the ability to discriminate the signal (i.e., old words) from the noise (i.e., new words)),[1] and response criterion (c', an index reflecting an individual' propensity for the "old" response in the recognition test) (for detailed explanations of d' and c', see Banks, 1970). The main measure of recognition performance employed in the current study was d', following precedents (e.g., Winograd & Vom Saal, 1966; H. Yang et al., 2015).

Bayesian analysis was conducted to assess whether the documented findings favor the null (H0; i.e., absence of the reactivity effect of JOFs) over the alternative (H1; i.e., existence of the reactivity of JOFs) hypothesis. BF_{10} represents the strength of evidence favoring the alternative hypothesis over the null hypothesis, and provides equivalence tests between groups, with $BF_{10} > 3$ representing evidence supporting the alternative hypothesis over the null and $BF_{10} < 0.33$ indicating evidence supporting the null hypothesis over the alternative hypothesis (Barchard, 2015; Mulder & Wagenmakers, 2016). All Bayesian analyses presented below were conducted via JASP 0.12.2 (http://jasp-stats.org/), with all parameters set as default.

A pre-planned paired *t*-test showed that d' for JOF words ($M = 2.302$, $SD = 0.578$) was significantly greater than that for no-JOF words ($M = 1.882$, $SD = 0.842$), difference = 0.420, 95% confidence interval [0.221, 0.619], $t(29) = 4.312$, $p < .001$, Cohen's $d = 0.787$, $BF_{10} = 160.632$ (see Fig. 1), indicating a positive reactivity effect of JOFs. Twenty-four participants presented a positive reactivity effect, five showed the converse pattern, and the remaining one was a tie.

Overall, the above findings reflect that making concurrent JOFs can reactively enhance retention of single words.

Experiment 2

As discussed in the *Introduction*, many studies have been conducted to explore if JOLs and JOFs differ in some respects, such as sensitivity to retention interval, remembering confidence, absolute accuracy, and relative accuracy. But the research findings are largely inconsistent. Going beyond previous studies, Experiment 2 aimed to explore if these two kinds of metamemory judgments have different reactivity effects on memory. Because a single experiment (Experiment 1) is insufficient to make a firm conclusion, another goal of Experiment 2 was to replicate the reactivity effect of JOFs documented in Experiment 1.

Method

Participants

According to the effect size observed in Experiment 1 (Cohen's $d = 0.785$), a power analysis, conducted via G*Power (Faul et al., 2007), showed that 15 participants in each group were required to observe a significant

[1] In the current study, d' was calculated as the signed difference between *Z*-transformed hit rate and *Z*-transformed false-alarm rate (Banks, 1970; Stanislaw & Todorov, 1999).

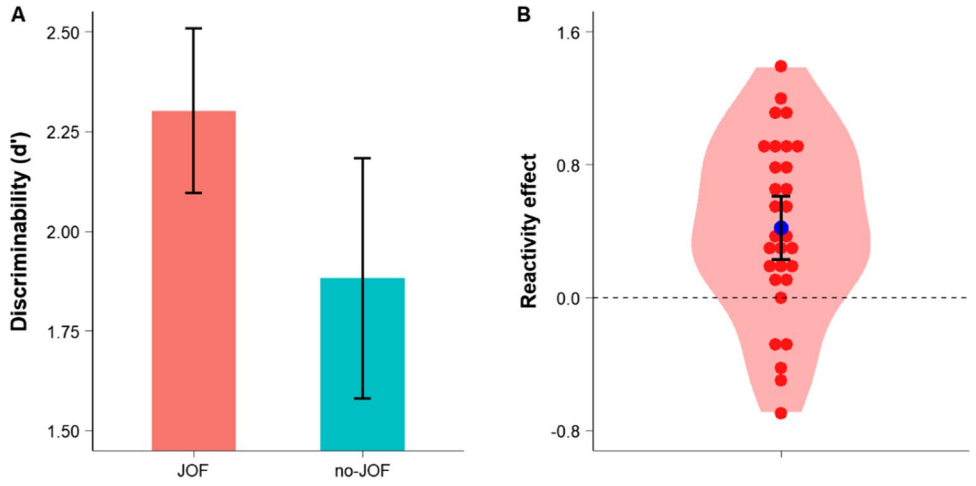

Fig. 1 Panel A: *d′* for JOF and no-JOF words in Experiment 1. Panel B: Violin plot depicting the distribution of the reactivity effect of JOFs (i.e., the difference in *d′* between JOF and no-JOL words). Each red dot represents one participant's reactivity effect score and the blue point represents group average. Error bars represent 95% CI

(two-tailed, $\alpha = .05$) reactivity effect at 0.80 power. To be more conservative and following Experiment 1, we decided to increase the sample size to 30 in each group. It is possible that such a sample size might be underpowered to detect a significant interaction between judgment type and study method.[2] To mitigate the concern about statistical power, we conducted Bayesian analyses to determine whether the obtained results support the existence or absence of the interaction. In addition, as shown below, a meta-analysis was performed to integrate results across studies to increase statistical power to further test potential difference between the reactivity effects of JOLs and JOFs.

Finally, 64 participants with a mean age of 19.688 (*SD* =1.622) years were recruited from the BNU participant pool; 57 were female. Thirty-two participants were randomly assigned to a JOL group, and the other 32 to a JOF group. All participants signed agreements to participate, were tested individually in a sound-proofed cubicle, and received financial remuneration. The protocol was approved by the Institutional Review Board of BNU Faculty of Psychology.

Materials

The materials were identical to those used in Experiment 1.

Procedure and design

Experiment 2 involved a 2 (judgment type: JOLs vs. JOFs) × 2 (study method: judgment vs. no-judgment) mixed design. Judgment type was manipulated as a between-subjects factor, and study method was a within-subjects factor.

The procedure for the JOF group was identical to that in Experiment 1. The procedure for the JOL group was highly similar, but with one exception. Specifically, for the two JOL lists in the JOL group, participants were asked to predict the likelihood that they would remember (rather than forget) each word in a later memory test. JOLs were made on a scale ranging from 0 (*Sure I will not remember it*) to 100 (*Sure I will remember it*).

Results and discussion

Table 1 lists hit rates, FA rates, *d′*, and *c′* for each group. A Bayesian mixed analysis of variance (ANOVA) was conducted via JASP 0.12.2, with all parameters set as default. All models were compared with the best performing model. In this ANOVA, study method was treated as the within-subjects variable, and judgment type was taken as the between-subjects variable, with *d′* as the dependent variable. BF_{incl} represents to what extent the documented findings support inclusion, by comparison with exclusion, of a given effect in a fitting model.

The results showed a main effect of study method, $F(1, 62) = 54.405$, $p < .001$, $\eta_p^2 = .467$, $BF_{\text{incl}} = 1.532\text{e+}7$ (see Appendix 2 for detailed results). As shown in Fig. 2, words studied with concurrent metamemory judgments were memorized better than those without metamemory

[2] As shown below, Experiments 2 and 3 and the meta-analysis consistently showed little difference in reactivity effects between JOLs and JOFs. It is, hence, unlikely for the current study to pre-determine the required sample size to observe a significant interaction at a specific power.

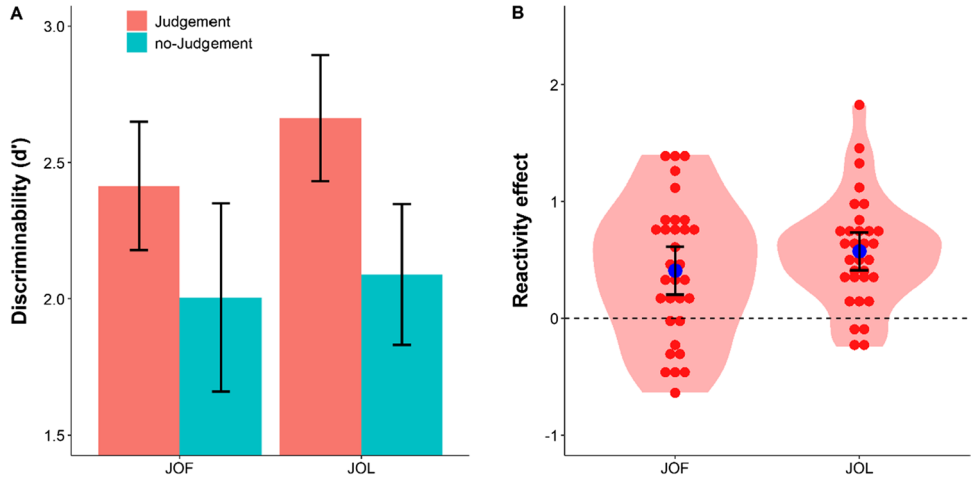

Fig. 2 Panel A: *d′* as a function of judgment type and study method in Experiment 2. Panel B: Violin plot depicting the distribution of the reactivity effects (i.e., the difference in *d′* between words studied with and those without metamemory judgments). Each red dot represents one participant's reactivity effect score and the blue points represent group averages. Error bars represent 95% CI

judgments, indicating an overall positive reactivity effect of making metamemory judgments on memory. The main effect of judgment type was non-significant, $F(1, 62) = 0.818$, $p = .369$, $\eta_p^2 = .013$, $BF_{incl} = 0.654$, indicating little difference in recognition performance between the JOL ($M = 2.376$, $SD = 0.760$) and the JOF ($M = 2.209$, $SD = 0.871$) groups. Of critical interest, there was no significant interaction between study method and judgment type, $F(1, 62) = 1.537$, $p = .220$, $\eta_p^2 = .024$, $BF_{incl} = 0.390$, indicating a minimal difference between the reactivity effects of JOFs and JOLs on learning of single words.

A pre-planned paired *t*-test showed that JOL words ($M = 2.662$, $SD = 0.668$) were memorized better than no-JOL ones ($M = 2.089$, $SD = 0.747$) in the JOL group, difference = 0.573 [0.405, 0.742], $t(31) = 6.947$, $p < .001$, $d = 1.228$, $BF_{10} = 1.731$+e5, replicating the classic positive reactivity effect of JOLs. Twenty-eight participants presented a positive reactivity effect, and the other four showed the converse pattern.

Along the same lines, JOF words ($M = 2.413$, $SD = 0.679$) were memorized better than no-JOF ones ($M = 2.005$, $SD = 0.996$) in the JOF group, difference = 0.408 [0.195, 0.621], $t(31) = 3.911$, $p < .001$, $d = 0.691$, $BF_{10} = 63.932$, replicating the positive reactivity effect of JOFs documented in Experiment 1. Twenty-three participants presented a positive reactivity effect, eight showed the converse pattern, and the other one was a tie.

An independent *t*-test showed no significant difference in *c′* between the JOL ($M = 1.497$, $SD = 0.412$) and JOF ($M = 1.387$, SD= 0.581) groups, difference = 0.110 [-0.143, 0.362], $t(62) = 0.872$, $p = .387$, $d = 0.218$, $BF_{10} = 6.067$e-4, suggesting minimal influence of judgment type on response criterion. Results of item-by-item JOFs and JOLs are reported in Appendix 1.

In summary, the above results imply that soliciting both JOLs and JOFs can reactively enhance recognition performance, and there is minimal difference in their reactivity effects on learning of single words.

Experiment 3

Experiment 3 was conducted to explore whether JOLs and JOFs have different reactivity effects on learning of word pairs, another type of material that has been widely used in previous JOL reactivity studies (Janes et al., 2018; Soderstrom et al., 2015).

It is worth noting that different mechanisms may underlie the reactivity effects on recognition and recall performance. In Experiments 1 and 2, the employed test format was old/new recognition, and the enhanced recognition performance might be attributed to the fact that making metamemory judgments facilitates recollection or familiarity (or a combination of both) of studied words. By contrast, Experiment 3 used a cued recall test to evaluate the reactivity effect on cued recall of word pairs. If a positive reactivity effect emerges in the cued recall test, this positive effect should mainly be attributed to enhanced recollection (rather than enhanced familiarity) induced by the requirement of making metamemory judgments.

As discussed above, different mechanisms may underlie the reactivity effects on recognition and recall performance. Hence, it is critical to test the generalizability of the findings

documented in Experiment 2 to different types of material (i.e., word pairs) and test format (i.e., cued recall).

Method

Participants

Previous studies observed the effect size (i.e., Cohen's *d*) of the reactivity effect of JOLs on learning of word pairs was 0.66 (C. Yang et al., 2021). We conducted a power analysis using G*Power (Faul et al., 2007) and found that about 21 participants were required in each group to observe a significant reactivity effect at 0.80 power. Given that Experiment 2 showed evidence supporting the absence of a difference between the reactivity effects of JOLs and JOFs, we did not estimate the required sample size to detect a significant interaction between study method and judgment type. To supplement, Bayesian analyses were performed to evaluate the strength of the documented findings (see below for details).

Finally, 42 participants (with 21 in each group) were recruited from the BNU participant pool, with a mean age of 20.571 ($SD = 2.307$) years; 38 were female. All participants signed agreements to participate, were tested individually in a sound-proofed cubicle, and received financial remuneration. The protocol was approved by the Institutional Review Board of BNU Faculty of Psychology.

Materials

The stimuli were 100 semantically related Chinese word pairs (e.g., *SCARF-GLOVES*) selected from Hu et al. (2016). Hu and colleagues asked participants to rate semantic relatedness of the word pairs on a scale ranging from 1 ("*completely unrelated*") to 4 ("*strongly related*"). The average relatedness ratings for the selected word pairs was 3.416 ($SD = 0.262$). Eighty pairs were used in the formal experiment, and the other 20 were used for practice. All stimuli were presented via the Matlab *Psychtoolbox* package.

Design and procedure

The experiment involved a 2 (judgment type: JOLs vs. JOFs) × 2 (study method: judgment vs. no-judgment) mixed design. Judgment type was a between-subjects variable, and study method was a within-subjects variable.

The procedure in Experiment 3 was similar to that in Experiment 2. Specifically, participants made item-by-item JOLs for two lists of word pairs in the JOL group, and they made item-by-item JOFs for two lists of word pairs in the JOF group. They did not need to make judgments for the other two lists of word pairs.

For the two lists without concurrent JOLs or JOFs, each word pair was presented on the screen for 8 s in total for participants to study. For the two lists with concurrent JOLs or JOFs, each word pair was firstly presented on the screen for 4 s, following which the word pair remained on the screen for another 4 s with a slider presented below it. Participants were instructed to drag and click the slider to make a JOL (in the JOL group) or JOF (in the JOF group) during the 4-s time window.

After studying all four lists, both groups engaged in a 5-min distractor task, during which they solved as many math problems as they could. After that, they took a cued recall test on all word pairs. Specifically, the cue words were presented one-by-one in a random order, and participants were required to recall the corresponding targets. There was no time pressure and no feedback in the cued recall test.

Results and discussion

Cued recall performance in each condition is depicted in Fig. 3. A Bayesian ANOVA, with study method as the within-subjects variable, judgment type as the between-subjects variable, and recall performance as the dependent variable, revealed a main effect of study method, $F(1, 40) = 17.232$, $p < .001$, $\eta_p^2 = .301$, $BF_{incl} = 169.518$. As shown in Fig. 3a, word pairs studied with concurrent metamemory judgments ($M = 0.857$, $SD = 0.106$) were recalled better than those without concurrent metamemory judgments ($M = 0.770$, $SD = 0.170$), indicating an overall positive reactivity effect of making metamemory judgments on memory of word pairs. The main effect of judgment type was non-significant, $F(1, 40) = 0.807$, $p = .374$, $\eta_p^2 = .020$, $BF_{incl} = 0.517$, indicating little difference in recall performance between the JOL ($M = 0.830$, $SD = 0.143$) and JOF ($M = 0.796$, $SD = 0.151$) groups. Of critical interest, the interaction between study method and judgment type was non-significant, $F(1, 40) = 0.013$, $p = .910$, $\eta_p^2 < .001$, $BF_{incl} = 0.308$, indicating minimal difference between the reactivity effects of JOFs and JOLs on learning of word pairs.

A pre-planned paired *t*-test showed that JOL pairs ($M = 0.873$, $SD = 0.100$) were recalled better than no-JOL ones ($M = 0.788$, $SD = 0.168$) in the JOL group, difference = 0.085 [0.020, 0.149], $t(20) = 2.923$, $p = .013$, $d = 0.593$, $BF_{10} = 3.979$, replicating the classic positive reactivity effect of JOLs on learning of word pairs. Fifteen participants showed a positive reactivity effect, four showed the converse pattern, and the other two were ties.

Along the same lines, JOF pairs ($M = 0.841$, $SD = 0.111$) were recalled better than no-JOF ones ($M = 0.751$, $SD = 0.174$) in the JOF group, difference = 0.089 [0.031, 0.147], $t(20) = 3.188$, $p < .01$, $d = 0.696$, $BF_{10} = 9.673$, reflecting a positive reactivity effect of JOFs on learning of word pairs. Fifteen participants presented a positive reactivity effect, and the other six showed the converse pattern. Results of item-by-item JOFs and JOLs are reported in Appendix 1.

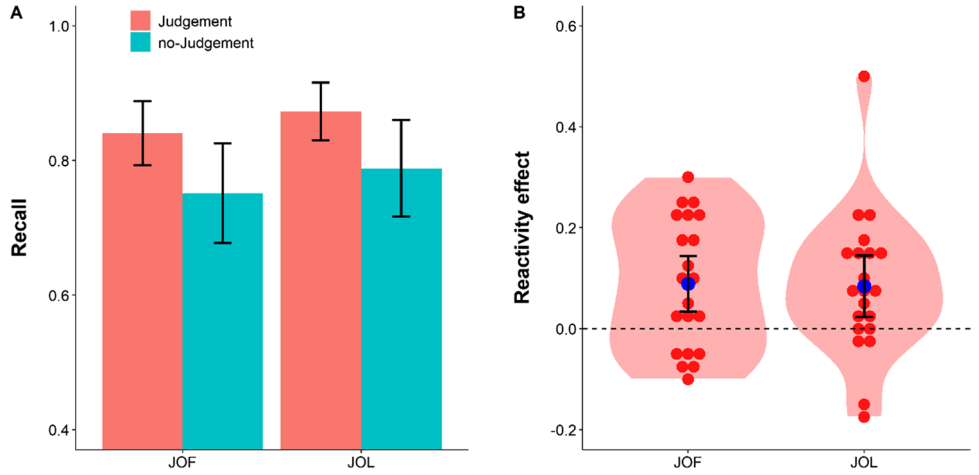

Fig. 3 Panel A: Cued recall performance as a function of judgment type and study method in Experiment 3. Panel B: Violin plot depicting the distribution of the reactivity effects (i.e., the difference in *d'* between items studied with and those without metamemory judgments). Each red dot represents one participant's reactivity effect score and the blue points represent group averages. Error bars represent 95% CI

In summary, the above results suggest that soliciting both JOLs and JOFs can reactively enhance retention of word pairs, and there is minimal difference between their reactivity effects on learning of word pairs.

Meta-analysis

As shown in Experiments 2 and 3, there was no significant difference between the reactivity effects of JOFs and JOLs on learning of single words and word pairs. There are two possible explanations for these non-significant results. The first is that there is truly no difference between their reactivity effects. The second is that these non-significant results are false negative, deriving from low statistical power in each experiment. To test the second explanation, we conducted a meta-analysis to integrate results across previous studies to increase statistical power and to further explore if there is any difference in memory performance between items studied with concurrent JOLs and those studied with concurrent JOFs. Recall that the framing explanation predicts a larger reactivity effect of JOFs compared to that of JOLs, whereas the positive reactivity theory expects minimal difference in their reactivity effects.

Before moving forward, it should be noted that, as discussed above, no previous research has been conducted to compare the reactivity effects of JOFs and JOLs. Instead, a set of previous studies employed two groups of participants, with one group making item-by-item JOLs and the other group making item-by-item JOFs for all study items (e.g., England et al., 2017; Finn, 2008; Serra & England, 2012; Tauber & Rhodes, 2012). In these studies, there was no control condition in which participants did not make metamemory judgments, making it impossible to quantify the magnitudes of the reactivity effects of JOLs and JOFs (i.e., difference in test performance between items studied with and those studied without concurrent judgments).[3] It is worth noting that our Experiments 2 ($t(62) = 0.383$, $p = .703$, $BF_{10} = 0.272$) and 3 ($t(40) = 0.701$, $p = .487$, $BF_{10} = 0.369$) found no significant difference in test performance between no-JOL and no-JOF items. Therefore, it is reasonable to take the difference in test performance between JOL and JOF items as a measure of the difference in reactivity effects between JOLs and JOFs.

Put differently, the difference between the reactivity effects of JOLs and JOFs ought to be computed as [(JOL items' test performance – no-JOL items' test performance) – (JOF items' test performance – no-JOF items' test performance)]. This calculation formula can also be expressed as [(JOL items' test performance – JOF items' test performance) – (no-JOL items' test performance – no-JOF items' test performance)]. Because all previous studies did not include no-JOL and no-JOF items in their experiments, it is impossible for us to directly compute the difference between the reactivity effects of JOLs and JOFs in previous studies. However, our Experiments 2 ($p = .703$, $BF_{10} = 0.272$) and 3 ($p = .487$, $BF_{10} = 0.369$) detected minimal difference in test performance between no-JOL and no-JOF items (that is, no-JOL items' test performance – no-JOF items' test performance ≈ 0), which means that the difference between the reactivity effects of JOLs and JOFs can be approximately

[3] For the sake of brevity, below we term items studied with concurrent JOLs as "JOL items," and items studied with concurrent JOFs as "JOF items."

computed as "(JOL items' test performance – JOF items' test performance) – 0". Accordingly, the current meta-analysis took the difference in test performance between JOL and JOF items as an approximate measure of the difference between the reactivity effects of JOLs and JOFs.

We re-highlight that the difference in test performance between JOL and JOF items is just an approximate measure of the difference between the reactivity effects of JOLs and JOFs, and readers are warned to be cautious when interpreting the meta-analytic results. To foreshadow, the current meta-analysis observed no statistically detectable difference in test performance between JOL and JOF items. Experiments 2 and 3 consistently provided Bayesian evidence supporting no difference between the reactivity effects of JOLs and JOFs. These consistent findings should help to allay the concern regarding the approximate measure of the difference between the reactivity effects of JOLs and JOFs.

Method

Literature search

To obtain a comprehensive set of eligible studies, we conducted a systematic search using the following search terms: (judgment* of forgetting OR judgment* of forgetting OR JOF*).[4] The search was performed in the following electronic databases: PubMed, Google Scholar, PsychINFO, Web of Sciences, ScienceDirect, ProQuest, and China National Knowledge Infrastructure (CNKI).

Inclusion and exclusion criteria

1. Duplicates were excluded. In addition, if the same results were reported in both a thesis and a journal article published by the same authors (e.g., Chen et al., 2016; Zhao, 2015), the thesis was excluded.
2. Empirical studies that employed objective measures of memory performance were included. Qualitative interviews, questionnaire surveys, and review articles were excluded because they did not measure objective memory performance (e.g., Koriat et al., 2004; Serra & England, 2019).
3. Only studies reporting sufficient information for effect size calculation were included.
4. Only English and Chinese studies were considered because of the authors' language proficiency.

The screening procedure and results are reported in a flowchart (see Fig. 4).

Effect size identification and calculation

Two authors independently performed data extraction and moderator coding. All differences were settled through group discussion. The database search procedure identified 11 studies as eligible, and we also included two experiments from the current study (Experiments 2 and 3). In total, the current meta-analysis included 33 effects from 12 studies (see the Reference list for included studies, which are marked by an asterisk).

Figure 5 depicts the characteristics (material types and test formats) of the 33 effects. If Cohen's *d*s were reported in the original reports, we directly extracted the reported *d* values. Otherwise, Cohen's *d*s were calculated using the formulae provided by Borenstein et al. (2009). To mitigate potential bias in effects with small sample sizes, all Cohen's *d*s were transformed into Hedges' *g*s using the bias correction function provided by Hedges (1982).

For three within-subject effects (Chen et al., 2016), correlations between test performance in the JOL and JOF conditions are required to adjust within-group standard deviations (*SD*s) and Hedges' *g*s. Following previous meta-analyses (Chan et al., 2018; Cumming, 2013; Pan & Rickard, 2018), we used $r = 0.5$ to adjust Cohen's *d*s and Hedges' *g*s.

Results and discussion

All meta-analyses were performed using random-effects models via the R *metafor* package (Viechtbauer, 2010). A positive value of Hedges' *g* indicates better memory for JOL items than that for JOF items, and a negative value represents the reverse pattern.

Difference in test performance between JOL and JOF items

A random-effects meta-analysis showed no significant difference in test performance between JOL and JOF items, Hedges' $g = 0.052$ [-0.070, 0.173], $Z = 0.832$, $p = .405$. Heterogeneity amongst the effects was significant, $Q(32) = 54.567$, $p = .008$. It should be noted, out of these 33 effects, three observations involved within-subjects comparison and the true correlations between JOL and JOF items' test performance were unknown (Chen et al., 2016). Therefore, a new meta-analysis was conducted, in which these three effects were excluded. After removing these three effects, the results again showed no significant difference between JOL and JOF items, Hedges' $g = -0.013$ [-0.118, 0.091], $Z =$

[4] To identify as many eligible studies as possible, we did not implement the combined terms [(judgement* of learning OR judgment* of learning OR JOL*) AND (judgement* of forgetting OR Judgment* of forgetting OR JOF*)] in the search process. It is well known that more restrictions on search terms will reduce the number of studies returned from the electronic databases, bringing the risk of missing eligible studies.

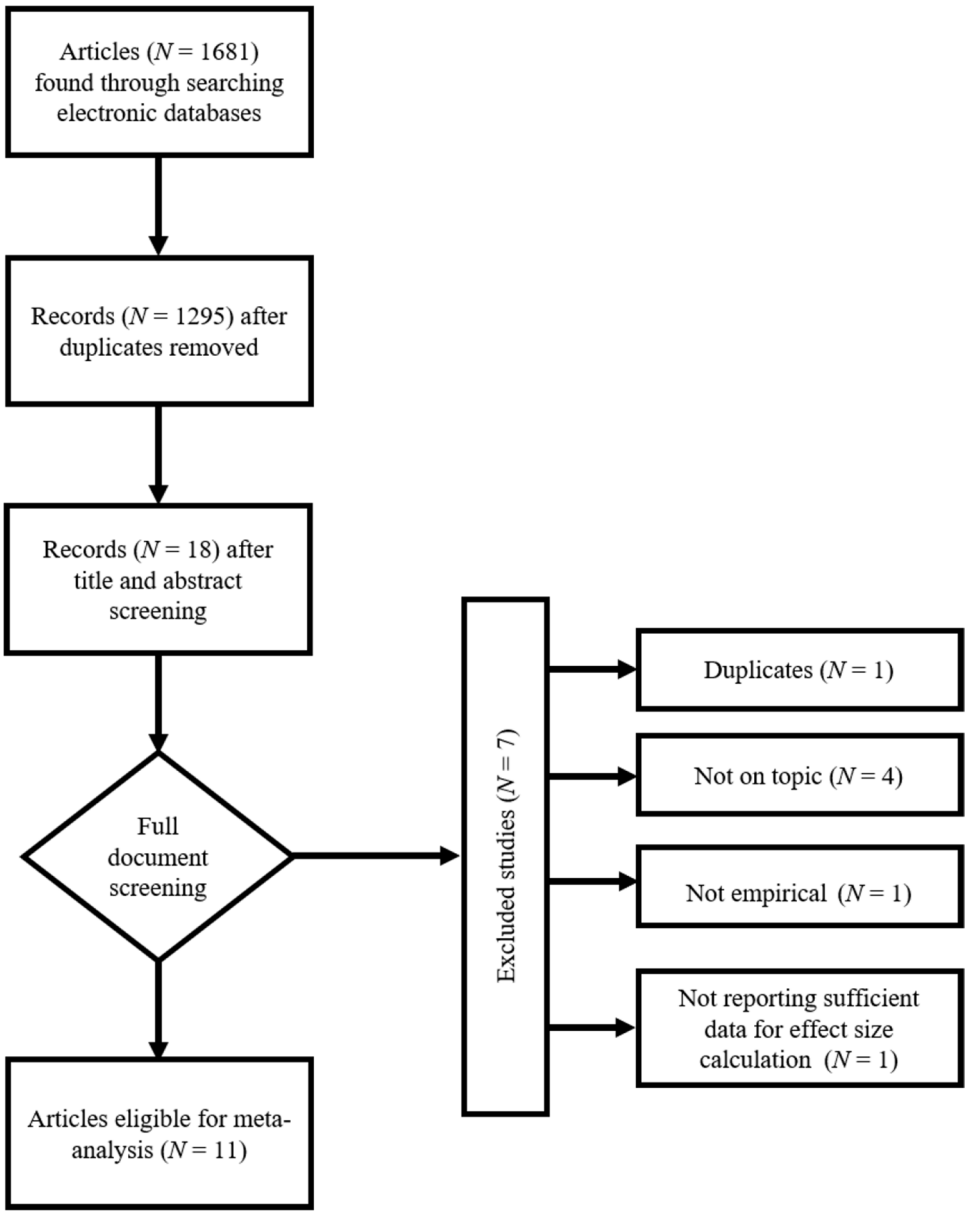

Fig. 4 Flowchart depicting article screening results

-0.251, $p = .802$. Heterogeneity amongst the remaining 30 effects was not significant, $Q(29) = 34.266$, $p = .230$.

Figure 6 is a funnel plot depicting the relationship between the 33 effects and their corresponding standard errors (*SE*s). To test potential publication bias, a meta-regression test (Egger et al., 1997) was conducted, which showed that the asymmetry of the funnel plot was non-significant, slope coefficient of Hedges' *g*s on *SE*s = -2.063 [-5.104, 0.979], $Z = -1.329$, $p = .184$, indicating that the risk of publication bias in the included effects was little.

Because the above results showed that these 33 effects were heterogeneous, sub-group meta-analyses were performed to explore their potential moderators. Previous studies found that material type significantly moderates the reactivity effect of JOLs (Double et al., 2018; Dunlosky et al., 2002; Thiede & Dunlosky, 1999; C. Yang et al., 2021). For instance, Double et al. (2018) found that making JOLs significantly enhances retention of related word pairs and single words, but it has minimal influence on retention of unrelated word pairs. Therefore, a sub-group meta-analysis was conducted to explore whether material type moderated the difference in test performance between JOL and JOF items. The results showed that material type did not significantly moderate the difference in test performance, $Q(3) = 0.173$, $p = .982$. Table 2 lists the results for each material type.

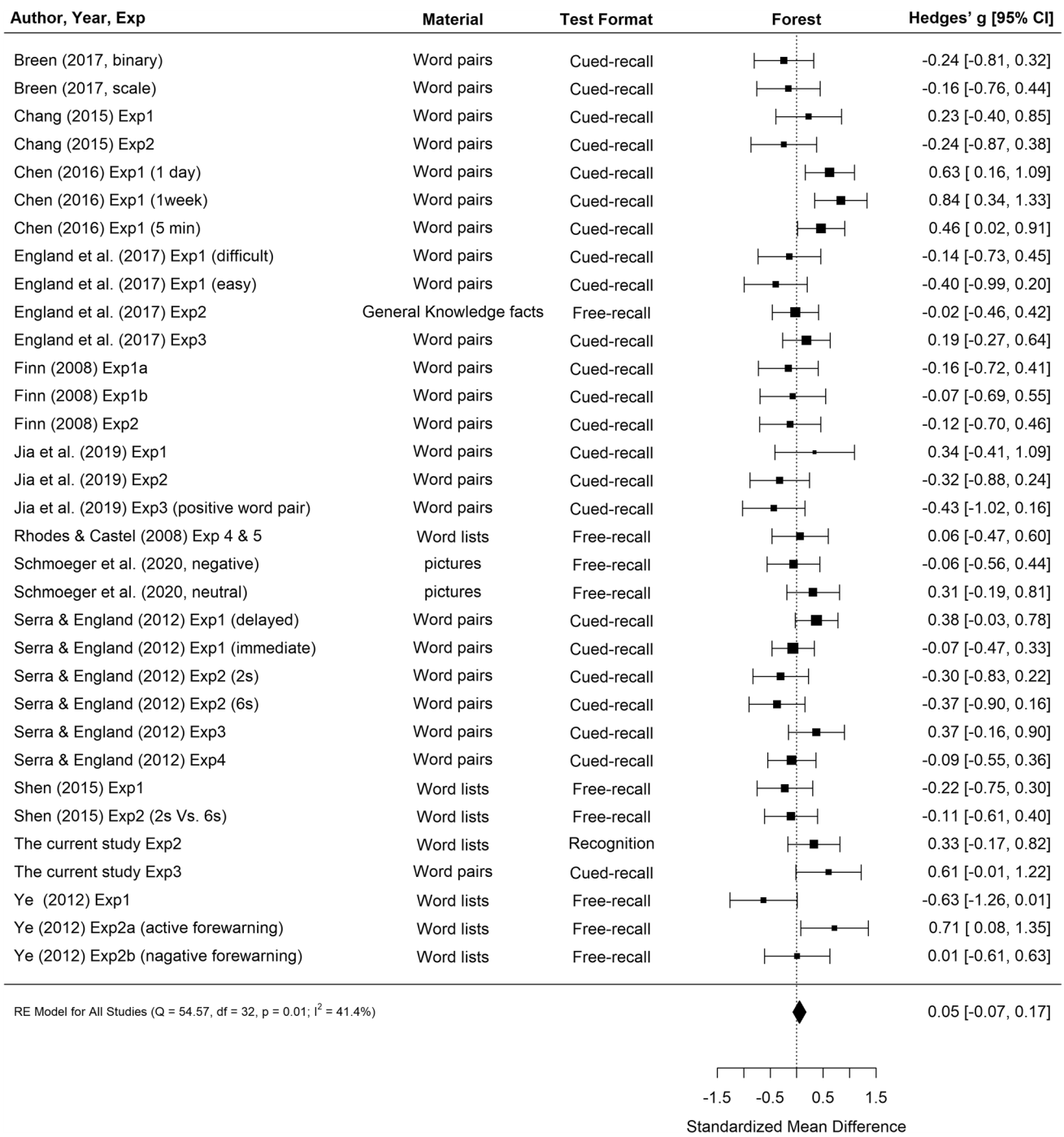

Fig. 5 Forest plot summarizing the 33 effect sizes (Hedges' *g*s), their experimental characteristics (material types and test formats), and the random-effects (RE) meta-analysis results. Error bars represent 95% CI

Myers et al. (2020) found that the reactivity effect of JOLs is modulated by test format. Thus, another sub-group meta-analysis was conducted to investigate whether test format moderated the difference in test performance between JOL and JOF items. The answer was negative, $Q(2) = 0.790$, $p = .674$ (see Table 2 for detailed results).

In summary, the current meta-analysis found little difference in test performance between JOL and JOF items, which is consistent with the findings documented in Experiments 2 and 3.

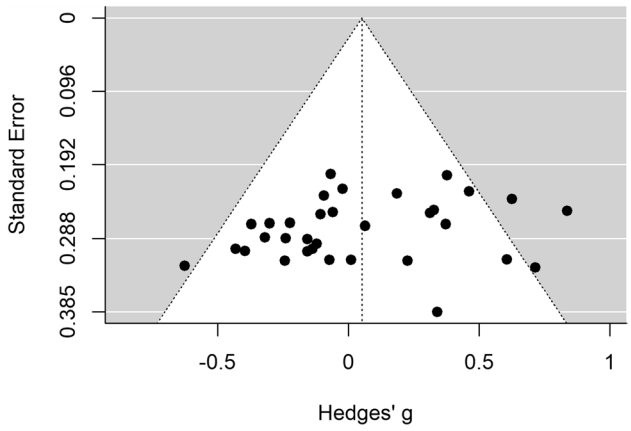

Fig. 6 Funnel plot depicting the relationship between Hedges' *g*s and their corresponding *SE*s

Table 2 Moderator analysis results

Moderators	*k*	*g*	95% CI	Q_B	*p*
Material type				0.173	. 982
General knowledge facts	1	-0.022	[-0.686, 0.642]		.948
Pictures	2	0.125	[-0.374, 0.624]		.623
Word pairs	23	0.023	[-0.261, 0.306]		.876
Word List	7	-0.055	[-0.098, 0.208]		.482
Test format				0.790	.674
Cued recall	23	0.056	[-0.092, 0.205]		.459
Free recall	9	0.004	[-0.233, 0.241]		.973
Recognition	1	0.327	[-0.351, 1.005]		.345

General discussion

The current study was conducted to (1) explore whether making JOFs reactively potentiates memory, (2) investigate whether JOLs and JOFs have different reactivity effects on memory, and (3) test the positive reactivity theory and the framing explanation. The principal findings documented in the current study are that making JOFs, similar to making JOLs, significantly enhanced retention of single words (Experiment 1), and there was minimal difference between the reactivity effects of JOFs and JOLs on learning of single words (Experiment 2) and word pairs (Experiment 3). Furthermore, the meta-analysis, which integrated results across 33 effects, demonstrated minimal difference in test performance between JOL and JOF items. These consistent findings jointly imply little or no difference in reactivity effects between JOLs and JOFs.

Although a set of recent studies has consistently demonstrated that making concurrent JOLs can reactively change memory itself (Double & Birney, 2018; Double et al., 2018; C. Yang et al., 2021), no research has been conducted to explore whether making concurrent JOFs can also change memory. The current study is the first to provide evidence justifying the reactivity effect of JOFs. The significant reactivity effect of JOFs is consistent with the positive reactivity theory, which attributes the reactivity effect to the fact that making item-by-item metamemory judgments enhances learning engagement and induces strategy changes and more elaborative processing. The positive reactivity effects of JOFs on learning of single words and word pairs suggest that making concurrent JOFs can be employed as a practical strategy to boost learning, at least for enhancing the learning of word lists and word pairs.

Although an emerging body of recent studies has been conducted to explore the difference between JOFs and JOLs, their research findings are largely inconsistent, and it remains largely unknown whether JOFs and JOLs are truly distinct forms of metamemory judgments (e.g., England et al., 2017; Finn, 2008; Koriat et al., 2004; Serra & England, 2012; Serra & England, 2019). To further explore this critical question, the current study acted as the first to explore if JOFs and JOLs have different reactivity effects on memory. The answer was overall negative. Experiments 2 and 3 consistently showed Bayesian evidence supporting the absence of a difference between the reactivity effects of JOFs and JOLs on learning of single words and word pairs. Even though the meta-analysis included over 1,700 participants' data, it still showed little difference in test performance. These consistent findings jointly support the claim that JOLs and JOFs differ minimally in their reactivity effects on memory.

The minimal difference between the reactivity effects of JOLs and JOFs is consistent with the positive reactivity theory (Mitchum et al., 2016; Rivers, 2018). For instance, asking participants to make metamemory judgments, regardless of whether the judgments were framed in terms of remembering (i.e., JOLs) or forgetting (i.e., JOFs), might enhance participants' engagement across the learning task because they had to focus on study items in order to make appropriate metamemory judgments. In addition, to provide appropriate judgments, participants had to search "diagnostic" cues to inform judgment formation, and the search process might in turn induce elaborative processing of study items, leading to enhanced retention.

The minimal difference between the reactivity effects of JOLs and JOFs is inconsistent with the framing explanation. Recall that the framing explanation proposes that forgetting frame, compared with remembering frame, reduces memory confidence and motivates individuals to expand greater study effort (Finn, 2008), which should lead to a larger reactivity effect of JOFs compared to that of JOLs. However, the

documented findings do not support this assumption. It is also worth noting that Serra and England (2012) observed that forgetting frame, relative to remembering frame, did not reduce memory confidence. Hence, findings from the current study and those from Serra and England (2012) jointly challenge the framing explanation.

Limitations and future research directions

It should be acknowledged that the current study suffers from three limitations. First, our Experiments 1–3 found a significantly positive reactivity effect of JOFs for young college students. It is unknown whether the reactivity effect of JOFs is generalizable to other populations. Future research can favorably explore this question and further test whether JOLs and JOFs have different reactivity effects in different populations. For instance, although many studies showed that making concurrent JOLs can reactively enhance retention of word pairs for young college students, Tauber and Witherby (2019) found that this positive reactivity effect of JOLs fails to generalize to older adults. Future research can explore if the positive reactivity effect of JOFs, different from the reactivity effect of JOLs, is able to benefit older adults' memory.

Second, the current study found positive reactivity effects of JOFs on learning of single words and word pairs, suggesting that making concurrent JOFs can act as a practical strategy to enhance learning of such materials. However, obviously, such materials are not highly representative of real educational materials, such as text passages and lecture videos. Future studies can usefully employ real educational materials to test the reactivity effect of JOFs, which is of practical importance.

In addition, exploring the reactivity effect of JOFs on learning of other materials can also be used to compare the reactivity effects of JOFs and JOLs. For instance, although previous studies found that making JOLs enhances retention of single words and word pairs, Ariel et al. (2021) found that making JOLs fails to enhance memory of text passages. Future research can explore whether making JOFs, different from making JOLs, is able to enhance retention of text passages. C. Yang et al. (2021) found that making JOLs is able to alter retention of general knowledge facts, and hence future research is encouraged to explore whether the reactivity effect of JOFs is also generalizable to learning of general knowledge facts.

Finally, the current meta-analysis took the difference in test performance between JOL and JOF items as an approximate measure of the difference in reactivity effects between JOLs and JOFs. Even though our Experiments 2 and 3 consistently observed minimal difference in test performance between no-JOL and no-JOF items, the difference in test performance between JOL and JOF items is an imperfect measure of the difference in reactivity effects because the calculation of this measure does not involve test performance of no-JOL and no-JOF items. Further research on the difference in reactivity effects between JOLs and JOFs is required, and a new meta-analysis should be implemented to directly assess the difference in reactivity effects when more eligible studies are available.

Conclusion

Making concurrent JOFs can reactively enhance retention of single words and word pairs. There is minimal difference between the reactivity effects of JOFs and JOLs. The positive reactivity theory is an available explanation to account for the reactivity effect, and the framing explanation garners less support.

Appendix 1

Experiment 1

JOFs Participants successfully provided item-by-item JOFs to 98.4% (SD = 1.9%) of the words in the JOF lists. To make JOFs and JOLs comparable in Experiments 2 and 3, we reverse-scored JOFs (i.e., 100 – JOF) in Experiments 1-3, following precedents (England et al., 2017; Finn, 2008; Serra & England, 2012, 2019; Tauber & Rhodes, 2012). The average of reversed JOFs was 61.830 (SD = 12.839).

Relative accuracy For each participant, a Gamma (G) correlation was calculated to measure relative accuracy of JOFs. To calculate intra-individual Gs, JOFs were reverse-scored. The averaged G across participants were 0.164 (SD = 0.217, 95% CI [.083, .245]), which is significantly greater than 0, $t(29) = 4.140$, $p < .001$, Cohen's $d = 0.756$, $BF_{10} = 105.044$.

Experiment 2

JOFs & JOLs For the JOF group, participants successfully provided item-by-item JOFs to 98.6% (SD = 2.2%) of the words in the JOF lists. For the JOL group, participants successfully provided item-by-item JOLs to 97.6% (SD = 1.0%) of the words in the JOL lists. An independent t-test showed a significant difference in proportions of words for which participants successfully made a judgment during the required time-window between the JOF and JOL groups, difference = 1.0% [0.1%, 1.9%], $t(62) = 2.403$, $p = .019$, $d = 0.601$, $BF_{10} = 2.784$. This might be a sample error because (1) this significant difference was not replicated in Experiment 3 and (2) there is little reason to expect that the JOF group would successfully

make item-by-item judgments to more words than the JOL group.

As in Experiment 1, we reverse-scored JOFs to make them comparable with JOLs. An independent sample *t*-test showed no significant difference between reversed JOFs ($M = 59.133$, $SD = 11.854$) and JOLs ($M = 63.558$, $SD = 12.748$), difference = -4.424 [-10.576, 1.727], $t(62) = -1.438$, $p = .156$, $d = -0.359$, $BF_{10} = 0.610$ (for related findings, see Breen, 2017; Serra & England, 2012).

Relative accuracy We calculated *G*s to measure relative accuracy of reversed JOFs (*M* of *G*s = 0.239, *SD* = 0.222) and JOLs (*M* of *G*s = 0.196, *SD* = 0.259). An independent *t*-test showed no significant difference in relative accuracy between reversed JOFs and JOLs, difference = 0.043 [-0.078, 0.164], $t(62) = 0.710$, $p = .480$, $d = 0.178$, $BF_{10} = 0.316$ (for related findings, see England et al., 2017; Serra & England, 2012; Serra & England, 2019; Tauber & Rhodes, 2012).

Experiment 3

JOFs & JOLs For the JOF group, participants successfully provided item-by-item JOFs to 98.5% (*SD* = 2.3%) of the word pairs in the JOF lists. For the JOL group, participants successfully provided item-by-item JOLs to 97.7% (*SD* = 1.9%) of the word pairs in the JOL lists. An independent *t*-test showed no significant difference in proportions of word pairs for which participants successfully made a judgment during the required time-window between the JOF and JOL groups, difference = 0.7% [-2.0%, 0.6%], $t(40) = -1.092$, $p = .281$, $d = -0.337$, $BF_{10} = 0.487$.

As in Experiment 1, we reverse-scored JOFs to make them comparable with JOLs. An independent sample *t*-test showed a significant difference between reversed JOFs ($M = 59.344$, $SD = 3.185$) and JOLs ($M = 43.220$, $SD = 10.946$), difference = 16.124 [11.097, 21.152], $t(40) = 6.482$, $p < .001$, $d = 2.000$, $BF_{10} = 1.030e+5$, which was similar to the findings documented in previous research (Serra & England, 2019).

Relative accuracy *G*s were calculated to quantify relative accuracy. Four participants (two in the JOL group and two in the JOF group) were excluded from analyses because of constant judgments provided to all items. An independent *t*-test showed a significant difference in relative accuracy between reversed JOFs ($M = 0.191$, $SD = 0.279$) and JOLs ($M = -0.191$, $SD = 0.374$), difference = 0.381 [0.164, 0.599], $t(36) = 3.561$, $p = .001$, $d = 1.155$, $BF_{10} = 29.651$.

Appendix 2

Experiment 2

We conducted a Bayesian mixed ANOVA via JASP, with judgment type as the between-subjects factor, and study method as the within-subjects factor. We used the default prior options for the effects (e.g., $r = 0.5$ for the fixed effects). The "Model Comparison" table, presented below, reports model comparison results. The column of BF_{10} indicates how many times the data are more likely under the model with only the main effect of study method than under other models.

The "Analysis of Effects" table lists *BF*s for inclusion of each effect. For study method, there is extremely strong evidence in favor of its inclusion ($BF_{incl} = 1.532e+7$). For judgment type, there is weak evidence against its inclusion ($BF_{incl} = 0.654$). And for the interaction between study method and judgment type, there is some evidence against its inclusion ($BF_{incl} = 0.390$).

The result of Bayesian mixed ANOVA in Experiment 2.

Model Comparison

Models	$P_{(M)}$	$P_{(M\|data)}$	BF_M	BF_{10}	error %
study method	0.200	0.524	4.399	1.000	
study method + judgment type	0.200	0.343	2.085	0.654	16.079
study method + judgment type + study method * judgment type	0.200	0.134	0.617	0.255	4.175
Null model (incl. subject)	0.200	3.879e-8	1.552e-7	7.406e-8	3.527
judgment type	0.200	1.775e-8	7.101e-8	3.389e-8	3.565

Analysis of Effects

Effects	$P_{(incl)}$	$P_{(incl\|data)}$	BF_{incl}
study method	0.400	0.866	1.532e+7
judgment type	0.400	0.343	0.654
judgment type * study method	0.200	0.134	0.390

Experiment 3

Similar to Experiment 2, we executed a 2×2 Bayesian mixed ANOVA. There is strong evidence in favor of inclusion of study method ($BF_{incl} = 159.518$), weak evidence against inclusion of judgment type ($BF_{incl} = 0.308$), and some evidence against inclusion of interaction between study method and judgment type ($BF_{incl} = 0.308$).

The result of Bayesian mixed ANOVA in Experiment 3.

Model Comparison

Models	$P_{(M)}$	$P_{(M\|data)}$	BF_M	BF_{10}	error %
study method	0.200	0.593	5.831	1.000	
study method + judgment type	0.200	0.307	1.771	0.517	6.059

Models	$P_{(M)}$	$P_{(M\|data)}$	BF_M	BF_{10}	error %
study method + judgment type + study method * judgment type	0.200	0.095	0.418	0.160	10.886
Null model (incl. subject)	0.200	0.004	0.015	0.006	1.032
judgment type	0.200	0.002	0.007	0.003	1.488

Analysis of Effects

Effects	$P_{(incl)}$	$P_{(incl\|data)}$	BF_{incl}
study method	0.400	0.900	169.518
judgment type	0.400	0.309	0.517
judgment type * study method	0.200	0.095	0.308

Acknowledgements This research was supported by the Natural Science Foundation of China (32000742; 32171045) and the Fundamental Research Funds for the Central Universities (2019NTSS28).

References

References marked with an asterisk (*) indicate studies included in the meta-analysis.

Ariel, R., Karpicke, J. D., Witherby, A. E., & Tauber, S. K. (2021). Do judgments of learning directly enhance learning of educational materials? *Educational Psychology Review, 33*, 693-712. https://doi.org/10.1007/s10648-020-09556-8

Banks, W. P. (1970). Signal detection theory and human memory. *Psychological Bulletin, 74*, 81-99. https://doi.org/10.1037/h0029531

Barchard, K. A. (2015). Null hypothesis significance testing does not show equivalence. *Analyses of Social Issues and Public Policy, 15*, 418-421. https://doi.org/10.1111/asap.12095

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). Converting among effect sizes. In U. Chichester (Ed.), *Introduction to meta-analysis* (pp. 45-49). Wiley.

*Breen, R. (2017). *Measuring metacognition: the effects of framing and scale type on metacognitive accuracy.* University of Tasmania, Retrieved from https://eprints.utas.edu.au/31272/

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition, 2*, 16. https://doi.org/10.5334/joc.72

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLoS ONE, 5*, e10729. https://doi.org/10.1371/journal.pone.0010729

Chan, J. C., Meissner, C. A., & Davis, S. D. (2018). Retrieval potentiates new learning: A theoretical and meta-analytic review. *Psychological Bulletin, 144*, 1111–1146. https://doi.org/10.1371/journal.pone.0010729.s002

*Chang, X. (2015). *The influence of framing effect under different material difficulty on judgement of learning.* (Master), Inner Mongolia Normal University, Retrieved from https://kns.cnki.net/kcms/detail/detail.aspx?dbcode=CMFD&dbname=CMFD201601&filename=1015432233.nh&v=6%25mmd2Bvgxt%25mmd2BePBMUY3R4FBlltO2w9loOu0kHjeES9RG0cWUIy3HfDc4LQpbqLaGGi6sM

*Chen, G., Qiao, F., & Zhao, J. (2016). The influence of retention interval and cue types on judgments of forgetting. *Studies of Psychology and Behavior, 14*, 433-437. Retrieved from http://www.en.cnki.com.cn/Article_en/CJFDTotal-CLXW201604001.htm

Cumming, G. (2013). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.

Double, K. S., & Birney, D. P. (2018). Reactivity to confidence ratings in older individuals performing the latin square task. *Metacognition and Learning, 13*, 309-326. https://doi.org/10.1007/s11409-018-9186-5

Double, K. S., & Birney, D. P. (2019). Reactivity to measures of metacognition. *Frontiers in Psychology, 10*, 2755. https://doi.org/10.3389/fpsyg.2019.02755

Double, K. S., Birney, D. P., & Walker, S. A. (2018). A meta-analysis and systematic review of reactivity to judgements of learning. *Memory, 26*, 741-750. https://doi.org/10.1080/09658211.2017.1404111

Dougherty, M. R., Robey, A. M., & Buttaccio, D. (2018). Do metacognitive judgments alter memory performance beyond the benefits of retrieval practice? A comment on and replication attempt of Dougherty, Scheck, Nelson, and Narens (2005). *Memory & Cognition, 46*, 558-565. https://doi.org/10.3758/s13421-018-0791-y

Dunlosky, J., & Hertzog, C. (1997). Older and younger adults use a functionally identical algorithm to select items for restudy during multitrial learning. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 52*, P178-P186. https://doi.org/10.1093/geronb/52B.4.P178

Dunlosky, J., Rawson, K. A., & McDonald, S. L. (2002). Influence of practice tests on the accuracy of predicting memory performance for paired associates, sentences, and text material. In *Applied metacognition.* (pp. 68-92). Cambridge University Press.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ, 315*, 629-634. https://doi.org/10.1136/bmj.315.7109.629

*England, B. D., Ortegren, F. R., & Serra, M. J. (2017). Framing affects scale usage for judgments of learning, not confidence in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*, 1898-1908. https://doi.org/10.1037/xlm0000420

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175-191. https://doi.org/10.3758/bf03193146

*Finn, B. (2008). Framing effects on metacognitive monitoring and control. *Memory & Cognition, 36*, 813-821. https://doi.org/10.3758/mc.36.4.813

Hedges, L. V. (1982). Estimation of effect size from a series of independent experiments. *Psychological Bulletin, 92*, 490-499. https://doi.org/10.1037/0033-2909.92.2.490

Hu, X., Liu, Z., Li, T., & Luo, L. (2016). Influence of cue word perceptual information on metamemory accuracy in judgement of learning. *Memory, 24*, 383-398. https://doi.org/10.1080/09658211.2015.1009470

Janes, J. L., Rivers, M. L., & Dunlosky, J. (2018). The influence of making judgments of learning on memory performance: Positive, negative, or both? *Psychonomic Bulletin & Review, 25*, 2356-2364. https://doi.org/10.3758/s13423-018-1463-4

*Jia, N., Wei, L., & Dai, J. (2019). Framing effects in judgments of learning: The role of processing fluency. *Studies of Psychology and Behavior, 17*, 729-735. Retrieved from http://www.xml-data.org/XYYXWYJ/html/2019/6/2019-6-729.htm

Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in Psychtoolbox-3? *Perception 36 ECVP*. Abstract Supplement.

Koriat, A., Bjork, R. A., Sheffer, L., & Bar, S. K. (2004). Predicting one's own forgetting: The role of experience-based and theory-based processes. *Journal of Experimental Psychology: General, 133*, 643-656. https://doi.org/10.1037/0096-3445.133.4.643

Mitchum, A. L., Kelley, C. M., & Fox, M. C. (2016). When asking the question changes the ultimate answer: Metamemory judgments change memory. *Journal of Experimental Psychology: General, 145*, 200-219. https://doi.org/10.1037/a0039923

Mulder, J., & Wagenmakers, E.-J. (2016). Bayes factors for testing hypotheses in psychological research: Practical relevance and new developments. *Journal of Mathematical Psychology, 72*, 1-5. https://doi.org/10.1016/j.jmp.2016.01.002

Myers, S. J., Rhodes, M. G., & Hausman, H. E. (2020). Judgments of learning (JOLs) selectively improve memory depending on the type of test. *Memory & Cognition, 48*, 745-758. https://doi.org/10.3758/s13421-020-01025-5

Nelson, T., & Leonesio, R. J. (1996). Consciousness and metacognition. *American Psychologist, 51*, 102-116. https://doi.org/10.1037/0003-066X.51.2.102

Nelson, T., & Narens, L. (1994). Why investigate metacognition. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing.* (pp. 1-25). The MIT Press.

Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological Bulletin, 144*, 710-756. https://doi.org/10.1037/bul0000151

*Rhodes, M. G., & Castel, A. D. (2008). Memory predictions are influenced by perceptual information: Evidence for metacognitive illusions. *Journal of Experimental Psychology: General, 137*, 615-625. https://doi.org/10.1037/a0013684

Rivers, M. L. (2018). *Investigating memory reactivity with a within-participant manuipulation of judgements of learning.* (Doctoral dissertation), Kent State University, Retrieved from http://rave.ohiolink.edu/etdc/view?acc_num=kent1536928272520919

Sahakyan, L., Delaney, P. F., & Kelley, C. M. (2004). Self-evaluation as a moderating factor of strategy change in directed forgetting benefits. *Psychonomic Bulletin & Review, 11*, 131-136. https://doi.org/10.3758/BF03206472

*Schmoeger, M., Deckert, M., Loos, E., & Willinger, U. (2020). How influenceable is our metamemory for pictorial material? The impact of framing and emotionality on metamemory judgments. *Cognition, 195*, 104112. https://doi.org/10.1016/j.cognition.2019.104112

*Serra, M. J., & England, B. D. (2012). Magnitude and accuracy differences between judgements of remembering and forgetting. *The Quarterly Journal of Experimental Psychology, 65*, 2231-2257. https://doi.org/10.1080/17470218.2012.685081

Serra, M. J., & England, B. D. (2019). Forget framing might involve the assumption of mastery, but probably does not activate the notion of forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 46*, 2384–2396. https://doi.org/10.1037/xlm0000804

*Shen, D. (2015). *A study of framing effects on the judgement of learning.* (Master), Zhejiang Normal University, Retrieved from https://kns.cnki.net/kcms/detail/detail.aspx?dbcode=CMFD&dbname=CMFD201601&filename=1015647343.nh&v=ZfgNe%25mmd2FiLE1G7RmjljE%25mmd2BFmQWLaxu5XGfblg%25mmd2BZxNSzCOmSpArd0VdD2fa8xZjqIERL

Soderstrom, N. C., Clark, C. T., Halamish, V., & Bjork, E. L. (2015). Judgments of learning as memory modifiers. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*, 553-558. https://doi.org/10.1037/a0038388

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers, 31*, 137-149. https://doi.org/10.3758/BF03207704

Tauber, S. K., & Rhodes, M. G. (2012). Measuring memory monitoring with judgements of retention (JORs). *Quarterly Journal of Experimental Psychology (Hove), 65*, 1376-1396. https://doi.org/10.1080/17470218.2012.656665

Tauber, S. K., & Witherby, A. E. (2019). Do judgments of learning modify older adults' actual learning? *Psychology and Aging, 34*, 836-847. https://doi.org/10.1037/pag0000376

Tekin, E., & Roediger, H. L. (2020). Reactivity of judgments of learning in a levels-of-processing paradigm. *Zeitschrift für Psychologie, 228*, 278-290. https://doi.org/10.1027/2151-2604/a000425

Thiede, K. W., & Dunlosky, J. (1999). Toward a general model of self-regulated study: An analysis of selection of items for study and self-paced study time. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 1024-1037. https://doi.org/10.1037/0278-7393.25.4.1024

Verkoeijen, P. P. J. L., Rikers, R. M. J. P., & Schmidt, H. G. (2005). The effects of prior knowledge on study-time allocation and free recall: Investigating the discrepancy reduction model. *The Journal of Psychology, 139*, 67-79. https://doi.org/10.3200/JRLP.139.1.67-79

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of statistical software, 36*, 1-48. https://doi.org/10.18637/jss.v036.i03

Winograd, E., & Vom Saal, W. (1966). Discriminability of association value in recognition memory. *Journal of Experimental Psychology, 72*, 328-334. https://doi.org/10.1037/h0023649

Yang, H., Cai, Y., Liu, Q., Zhao, X., Wang, Q., Chen, C., & Xue, G. (2015). Differential neural correlates underlie judgment of learning and subsequent memory performance. *Frontiers in Psychology, 6*, 1699. https://doi.org/10.3389/fpsyg.2015.01699

Yang, C., Huang, J., Li, B., Yu, R., Luo, L., & Shanks, D. R. (2021). Learning difficulty determines whether concurrent metamemory judgments enhance or impair learning outcomes: Meta-analytic and empirical tests. *Submitted for publication.*

*Ye, J. (2012). *Influence about the framing effect on practice with under confidence with practice effect.* (Master), Zhejiang Normal University, Retrieved from https://kns.cnki.net/kcms/detail/detail.aspx?dbcode=CMFD&dbname=CMFD201301&filename=1012486079.nh&v=W9auXH4ZSFOIgTxzd5H5DVKtNeau927Lnnt0DC7qElH1FQrzrA46wEflxBJ%25mmd2Fsqr5

Zhao, J. (2015). *The Influence of the factor for forgetting metacognitive monitoring.* (Master), University of Jinan, Retrieved from https://kns.cnki.net/kcms/detail/detail.aspx?dbcode=CMFD&dbname=CMFD201601&filename=1015438084.nh&v=aYX4viLJukgWSxLxQh6lRPLSd1yjSwMp6K0%25mmd2BHaR3TVGvC35m%25mmd2B%25mmd2Fv%25mmd2B9aTwlRXIu%25mmd2FPt. Available from Cnki

Open practices statement The data contained in this project are publicly available at the Open Science Framework: https://osf.io/6j9xf

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