Journal of Memory and Language 136 (2024) 104506

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Do changed learning goals explain why metamemory judgments reactively affect memory?

Baike Li[a,b], David R. Shanks[c], Wenbo Zhao[d], Xiao Hu[e,f], Liang Luo[b,e], Chunliang Yang[b,e,*]

[a] School of Psychology, Liaoning Normal University, Dalian, China
[b] Institute of Developmental Psychology, Faculty of Psychology, Beijing Normal University, Beijing, China
[c] Division of Psychology and Language Sciences, University College London, London, UK
[d] School of Social Development and Public Policy, Beijing Normal University, Beijing, China
[e] Beijing Key Laboratory of Applied Experimental Psychology, National Demonstration Center for Experimental Psychology Education, Beijing Normal University, Beijing, China
[f] Faculty of Psychology, Beijing Normal University, Beijing, China

ARTICLE INFO

Keywords:
Reactivity
Changed goal
Immediate JOLs
Pre-study JOLs
Transfer

ABSTRACT

Measurement of mental processes is the bedrock of cognitive psychology, but the interpretation of such measurements is profoundly undermined by evidence that many mental processes are changed by (are reactive to) the act of being observed and measured. The current article is concerned with one particular type of reactivity, namely changes in memory performance when individuals are asked to concurrently monitor their learning via judgments of learning (JOLs). One explanation for memory reactivity is that the requirement to engage in metamemory monitoring changes learners' goals, shifting them towards greater prioritization of mastering easy items and de-prioritization of memorizing difficult ones. This hypothesis is tested in 5 experiments (2 of which were pre-registered), which varied item difficulty by contrasting related (e.g., *computer – keyboard*) and unrelated (e.g., *book – shoe*) word pairs. While the experiments find robust evidence that recall is affected by the requirement to make immediate JOLs (reactivity), two key predictions of the goal-change account are not supported. The observed findings suggest that a change in the learner's goal is not the main mechanism underlying JOL reactivity. Alternative explanations for why memory is reactive to metamemory judgments are discussed.

Introduction

Subjective judgments (e.g., confidence ratings) have been widely used as a tool for measuring metacognition in a variety of cognitive domains, including learning and memory (Rhodes & Tauber, 2011; Yang et al., 2021), decision-making (Fleming, Weil, Nagy, Dolan, & Rees, 2010; Hu, Yang, & Luo, 2023), and deductive reasoning (Block, 2008; Shynkaruk & Thompson, 2006). Numerous behavioral and neuroimaging studies have measured individuals' metacognitive ability by asking them to make item-by-item judgments, such as a confidence rating after each perceptual (Hu et al., 2023) or memory (Hu, Yang, & Luo, 2022) decision. Most of these studies implicitly assumed that such metacognitive judgments provide neutral assessments of the cognitive processes they measure and have no impact on those processes or accompanying task performance (Spellman & Bjork, 1992).

Research going back several decades, however, shows that in many situations target processes and behaviors are reactively affected by these metacognitive judgments (Arbuckle & Cuddy, 1969; King, Zechmeister, & Shaughnessy, 1980; Li, Zhao, et al., 2023; Li et al., 2022; Mitchum, Kelley, & Fox, 2016; Shi et al., 2023; Soderstrom, Clark, Halamish, & Bjork, 2015; Zechmeister & Shaughnessy, 1980; W. B. Zhao et al., 2022; 2023; W. L. Zhao et al., 2022), a phenomenon known as the *reactivity effect* and which suggests that metacognitive judgments are not passive measures but can themselves alter "reality" (for reviews, see Double & Birney, 2019; Double, Birney, & Walker, 2018). For instance, many studies found that soliciting confidence ratings reactively enhances perceptual decision accuracy (Bonder & Gopher, 2019), slows down decision speed (Lei et al., 2020; Li, Hu, et al., 2023), improves reasoning performance in high-confidence individuals (Double & Birney, 2017), facilitates recognition memory (W. L. Zhao et al., 2022), and so on

* Corresponding author at: 19 Xinjiekou Wai Street, Beijing 100875, China.
E-mail address: chunliang.yang@bnu.edu.cn (C. Yang).

https://doi.org/10.1016/j.jml.2024.104506
Received 19 March 2023; Received in revised form 12 January 2024; Accepted 12 January 2024
Available online 1 February 2024

(Double & Birney, 2018). Besides soliciting explicit metacognitive judgments, classic research established that asking individuals to explain what they are doing or thinking (Nisbett & Ross, 1980; Nisbett & Wilson, 1977) can trigger meta-awareness (Winkielman & Schooler, 2011), which in turn reactively alters their task performance (i.e., the reactivity effect of concurrent verbalization). The recent work on reactivity induced by metacognitive judgments extends earlier assessments of reactivity induced by concurrent verbalization (for a review, see Fox, Ericsson, & Best, 2011).

One form of reactivity induced by metacognitive judgments, which has attracted substantial research interest in recent years, is the effect of judgments of learning (JOLs; metacognitive estimates about the likelihood of remembering a studied item on a later memory test). Research has observed that many learning outcomes are reactively changed as a consequence of making a JOL while or after studying each item (e.g., Ariel, Karpicke, Witherby, & Tauber, 2021; Chang & Brainerd, 2023; Double et al., 2018; Janes, Rivers, & Dunlosky, 2018; Li et al., 2022; Mitchum et al., 2016; Myers, Rhodes, & Hausman, 2020; Rivers, Janes, & Dunlosky, 2021; Senkova & Otani, 2021; Soderstrom et al., 2015; Tauber & Witherby, 2019; Tekin & Roediger, 2020; Witherby & Tauber, 2017b; W. B. Zhao et al., 2022; W. L. Zhao et al., 2022). In addition, many studies have documented the reactivity of memory to JOLs with different types of materials, such as word lists (Li et al., 2022; Maxwell & Huff, 2022), pure lists of related word pairs (Li et al., 2022; Rivers et al., 2021; Witherby & Tauber, 2017b; W. L. Zhao et al., 2022), visual images (Shi et al., 2023), and inter-item relations (W. B. Zhao et al., 2022).

Potential Mechanisms Underlying the Enhanced-Relatedness Effect

Why does soliciting metacognitive judgments reactively temper the entity being monitored? A possible explanation, proposed by Mitchum et al. (2016), is that monitoring ongoing mental processes during a cognitive task changes individuals' task goals, in turn leading to a reactive influence on performance. Below, we term this explanation the *changed-goal hypothesis*, following Mitchum et al. (2016). It is well-known that overtly soliciting metacognitive judgments enhances individuals' awareness of any discrepancy between their current level of mastery and their desired goals, which then guides subsequent metacognitive control processes (e.g., adjusting task goals, changing task strategies; Finn, 2008; Thiede, Anderson, & Therriault, 2003; Yang, Potts, & Shanks, 2017). This might explain why metacognitive judgments can affect the very processes being judged.

Regarding memory reactivity to JOLs, Mitchum et al. (2016) speculated that when the difference in learning difficulty among list items is obvious (e.g., when studying a mixed list of related and unrelated word pairs), soliciting JOLs encourages participants to consider that some items are more memorable than others. To prevent "labor in vain" (i.e., exerting effort toward remembering difficult items produces little improvement; Nelson & Leonesio, 1988), participants change their study goal from mastering all items to prioritizing easy ones, with a sacrifice in learning difficult ones. This leads to positive reactivity for easy items and negative reactivity for difficult ones. In essence, the use of a JOL scale with extreme values at 0 and 100 communicates to participants that some items are not memorable at all whereas others are highly memorable. This encourages participants to be more selective in their learning goals, switching away from their normal "mastery" mindset.

Mitchum et al. (2016) provided a clear demonstration of the memory reactivity effect of JOLs and of the possible role of goal-change in this effect. In their Experiments 1–3, Mitchum et al. instructed two groups (JOL *vs.* no-JOL) of participants to study a mixed list of related (e.g., *computer* – *keyboard*) and unrelated (e.g., *book* – *shoe*) word pairs, and they were allowed to spend as much time as they wanted to study each pair. The only difference between the two groups was that the JOL group was required to make a JOL after studying each pair, whereas the no-JOL group was not. Mitchum et al. found that, although both the JOL and no-JOL groups spent longer studying unrelated than related pairs, the JOL group did so to a lesser extent than the no-JOL group. Stated differently, the correlation between study time and cue-target relatedness (an index of learning difficulty) was less negative in the JOL than in the no-JOL group. This pattern is consistent with a shift in the JOL group towards prioritizing encoding of related (easy) pairs and sacrificing unrelated (difficult) ones.

What about actual recall? Mitchum et al. found that recall of related pairs was numerically enhanced by making JOLs while recall of unrelated ones was significantly impaired. In other words, they found that the difference in recall between related and unrelated pairs (i.e., the relatedness effect) was significantly larger in the JOL than in the no-JOL group, a phenomenon we term the *enhanced relatedness effect* (Janes et al., 2018). Clearly, the enhanced relatedness effect is also consistent with the changed-goal hypothesis.

It is worth noting that enhanced relatedness effects were not only observed in self-paced (Mitchum et al., 2016, Experiments 1–4) but also in experimenter-paced study conditions (Mitchum et al., 2016, Experiment 5; Janes et al., 2018). For instance, in Mitchum et al.'s Experiment 5, participants were instructed to study a mixed list of related and unrelated pairs in an experimenter-paced study procedure (i.e., with each pair presented for 5 s for participants to study). Participants in the JOL group made item-by-item JOLs after studying each word pair, whereas those in the no-JOL group did not. Strikingly, the results again showed an enhanced relatedness effect. Mitchum et al. proposed that, in experimenter-paced study conditions, the limited study time creates pressure on word pair learning, leading to a shift from a mastery orientation (i.e., mastering all items) toward a concentration on easy pairs (Metcalfe & Kornell, 2003). Indeed, Mitchum et al. noted that they heard participants spontaneously exclaim "I will never remember that" during the study phase, suggesting that they might simply wait out the item presentation screen without attempting to encode the study items when they perceived them to be too difficult to remember. Hence, Mitchum et al. claimed that the enhanced relatedness effect in experimenter-paced conditions is again consistent with the changed-goal hypothesis.

Janes et al. conducted two experiments with an experimenter-paced procedure to further test the changed-goal hypothesis (Janes et al., 2018, Experiments 2 and 3). Participants studied either a mixed list or two pure lists of related and unrelated word pairs. In the mixed list condition, participants studied 60 word pairs in total, composed of 30 related and 30 unrelated pairs. By contrast, in the pure list condition, they only studied a pure list of 30 related or a pure list of 30 unrelated pairs.

Based on the changed-goal hypothesis, Janes et al. predicted that the enhanced relatedness effect should only occur in the mixed list but not in the pure list condition, because pure lists of related and unrelated word pairs lack variation in item difficulty. That is, the essential driver of goal-change – the presence of items of widely varying difficulty – is absent in the pure list condition. Consistent with this prediction, Janes et al. observed an enhanced relatedness effect in the mixed list but not in the pure list condition. Accordingly, Janes et al. concluded that their study provides "support for the changed-goal hypothesis" (p. 2361).

Before moving forward, it should be noted that, besides the changed-goal hypothesis, there is another available theoretical account for the enhanced relatedness effect, that is, the *dual-mechanism hypothesis* (Mitchum et al., 2016; Janes et al., 2018). The dual-mechanism hypothesis asserts that the enhanced relatedness effect is caused by two separate mechanisms: 1) cue-strengthening (inducing positive reactivity for related pairs; Soderstrom et al., 2015), and 2) dual-task costs (inducing negative reactivity for unrelated pairs; Mitchum et al., 2016). Specifically, Soderstrom et al. (2015) developed a cue-strengthening explanation to explain positive reactivity for related pairs. They assumed that when a participant attempts to make a reasonable JOL for a word pair, they search for "diagnostic" cues (e.g., relatedness strength between the cue and target, mediators between the cue and target) to guide JOL formation. The cues activated by the requirement of making JOLs in turn strengthen the cue-target relation for related pairs, leading to positive reactivity. By contrast, because there is no pre-existing relation between the cue and target for unrelated pairs, making JOLs fails to enhance their memory.

Mitchum et al. (2016) proposed the dual-task costs hypothesis to explain why the reactivity effect is negative in some situations. They assumed that the additional requirement of making JOLs may borrow limited resources (e.g., study time, working memory capacity) from the primary learning task (Delaney & Verkoeijen, 2009; Griffin, Wiley, & Thiede, 2008; Rundus, 1971). In addition, frequent task mode switching between encoding (i.e., studying word pairs) and monitoring (i.e., making JOLs) may bring further dual-task costs (Davis, Chan, & Wilford, 2017; Doherty et al., 2019; Griffin et al., 2008; Karaca, Kurpad, Wilford, & Davis, 2020; Yang, Potts, & Shanks, 2018). Hence, making JOLs may reactively impair memory, especially when the learning task itself is highly challenging.

In summary, the dual-mechanism hypothesis asserts that generating JOLs reactively enhances memory of related pairs through inducing cue-strengthening, and impairs memory of unrelated pairs through inducing dual-task costs, in combination leading to the enhanced relatedness effect (Janes et al., 2018). The current study primarily focuses on assessing the validity of the changed-goal hypothesis. We further elaborate on the dual-mechanism hypothesis in the General Discussion.

Conflicting evidence on the changed-goal hypothesis

Previous findings about the role of goal-change in JOL reactivity are conflicting and inconclusive. For instance, although the findings provided by Mitchum et al. (2016) support the changed-goal hypothesis, Janes et al. (2018) failed to replicate Mitchum et al.'s self-paced learning findings. Specifically, Janes et al. (2018, Experiment 1) observed no difference in correlation between study time and relatedness strength between the JOL and no-JOL groups in self-paced study conditions. Furthermore, Janes et al. observed that, when the study procedure was self-paced, there was no statistically detectable difference in the relatedness effect between JOL and no-JOL groups. The inconsistent findings provided by Janes et al. (2018) and Mitchum et al. (2016) suggest that further tests on the changed-goal hypothesis are needed.

Additionally, Halamish and Undorf (2023) recently provided further evidence challenging the changed-goal hypothesis. In this study, two (JOL and no-JOL) groups of participants studied a mixed list of unrelated (e.g., *raft-kiss*), related (e.g., *lips-kiss*), and identical (e.g., *kiss-kiss*) word pairs. Participants in the JOL group gave highest JOLs for identical pairs, medium JOLs for related pairs, and lowest JOLs for unrelated pairs, reflecting that they believed that identical pairs were easiest and unrelated ones were most difficult to memorize. According to the changed-goal hypothesis, making JOLs should induce a larger positive reactivity effect for identical than for related pairs, because the former should be prioritized on the basis of their perceived ease of learning. However, in contrast to this prediction, Halamish and Undorf observed that the effect size of positive reactivity for related pairs (Cohen's $d = 0.57$) was over double that for identical pairs ($d = 0.25$).

Overall, previous findings regarding the changed-goal hypothesis of JOL reactivity are conflicting, with some supporting it (Janes et al., 2018, Experiment 2; Mitchum et al., 2016), and some challenging it (Halamish & Undorf, 2023; Janes et al., 2018, Experiment 1). It remains unclear whether making JOLs reactively alters memory through changing study goals. Besides the changed-goal hypothesis, the dual-mechanism hypothesis can also readily account for the enhanced relatedness effect (Janes et al., 2018), which will be further elaborated in the General Discussion.

The current study aims to test the changed-goal hypothesis in two ways. First, it investigates whether pre-study JOLs (i.e., JOLs made before studying each item) enhance the relatedness effect as much as immediate JOLs (i.e., JOLs made after studying each item). Second, it explores whether the enhanced relatedness effect transfers to no-JOL pairs when JOL and no-JOL pairs are studied in an interleaved manner (see below for details).

Pre-study JOLs vs. immediate JOLs

Previous studies mainly explored the reactivity effect of immediate JOLs on memory (Janes et al., 2018; Mitchum et al., 2016). Another widely studied form comprises pre-study JOLs (Castel, 2008; Price & Harrison, 2017; Witherby & Tauber, 2017a; Yang et al., 2021). Unlike immediate JOLs which are formed on the basis of a combination of processing experience (i.e., online experience obtained from the learning task, such as ease of processing) and beliefs about memory (i.e., beliefs about how a given factor, such as word frequency, affects memory), pre-study JOLs are solely based on metamemory beliefs because they are made before participants receive and study each item (Yang et al., 2021). [For discussion of the difference between immediate and pre-study JOLs, see Price and Harrison (2017).]

Pre-study JOLs have been frequently employed to measure people's beliefs about how a given factor (e.g., concreteness, relatedness, word frequency, font size, emotion, age, and so on) affects memory (Jia et al., 2015; Mueller, Dunlosky, Tauber, & Rhodes, 2014; Witherby & Tauber, 2017a; Yang, Huang, & Shanks, 2018). Directly related to the current study are the findings from Mueller and Dunlosky (2017). In this study, participants in a pre-study JOL group were informed whether the next word pair would be a related or unrelated one and made a pre-study JOL to predict the likelihood that they would remember the next pair in a later memory test. By contrast, participants in an immediate JOL group made a JOL *after* studying each pair. Mueller et al. (2017) found that participants in both the immediate JOL and pre-study JOL groups provided substantially higher JOLs for related than for unrelated pairs (for connected findings, see Price & Harrison, 2017).

Mueller et al.'s (2017) findings suggest that both pre-study and immediate JOLs encourage participants to compare the relative memorability of related and unrelated pairs (Mueller & Dunlosky, 2017; Price & Harrison, 2017). Therefore, according to the changed-goal hypothesis, pre-study JOLs, similar to immediate JOLs, should change participants' study goals and induce an enhanced relatedness effect (Mitchum et al., 2016). More specifically, the changed-goal hypothesis assumes that making immediate JOLs enhances awareness of the difference in learning difficulty between related and unrelated pairs, which then induces participants to prioritize memorizing easy related pairs and sacrifice difficult unrelated pairs. Given that making pre-study JOLs can also enhance awareness of the difference in learning difficulty (as reflected by the substantial difference in pre-study JOLs between related and unrelated pairs), pre-study JOLs should also change participants' study goals and induce an enhanced relatedness effect (Mitchum et al., 2016). It is even reasonable to expect that pre-study JOLs would induce a larger enhanced relatedness effect than immediate JOLs, because pre-study JOLs are made before participants observe each word pair (Castel, 2008). Knowing in advance whether the next pair will be easy or difficult to remember, they can then accordingly prepare to allocate more or fewer resources toward studying it.

Overall, the changed-goal hypothesis generates two predictions: (1) Pre-study JOLs should reactively enhance the relatedness effect, and (2) the enhanced relatedness effect of pre-study JOLs should be equal to (or even larger than) the effect of immediate JOLs. Experiments 1–4 were conducted to test the changed-goal hypothesis through assessing these two theoretical predictions.

Transfer of reactivity

Experiment 5 tested another prediction of the changed-goal hypothesis, concerning the transfer of reactivity from JOL pairs to no-JOL ones when they are intermixed. In the introduction to Experiment 5, we explain how the changed-goal hypothesis can be tested through investigating reactivity transfer.

Experiment 1

Experiment 1 was primarily conducted to test the changed-goal hypothesis through exploring (1) whether pre-study JOLs reactively enhance the relatedness effect, and (2) whether the enhanced relatedness effect of pre-study JOLs is equal to (or even larger than) the effect of immediate JOLs.

As discussed above, Janes et al. (2018, Experiment 1) and Mitchum et al. (2016, Experiments 1-3) observed inconsistent findings about the existence or absence of the enhanced relatedness effect in self-paced study conditions. Their results were also inconsistent about whether the requirement of making immediate JOLs reduces the negative correlation between study time and cue-target relatedness. Therefore, another aim of Experiment 1 is to employ a self-paced study procedure to further test the replicability of Mitchum et al.'s self-paced study findings.

Data availability

The data and analysis scripts for all experiments reported here are publicly available on OSF (https://osf.io/z9cfs).

Method

Design and participants

Experiment 1 involved a 3 (study method: immediate JOL *vs.* no-JOL *vs.* pre-study JOL) × 2 (relatedness: related *vs.* unrelated) mixed design, with study method as a between-subjects variable and relatedness as a within-subjects variable.

A pilot study was conducted to determine the required sample size. The pilot study recruited 10 participants in each of the immediate JOL, no-JOL, and pre-study JOL groups. The procedure and experimental stimuli in the pilot study were the same as in the formal experiment. The pilot results showed a negligible effect size for the interaction between study method and relatedness, $\eta_p^2 = .006$. A power analysis, conducted via G*Power (Faul, Erdfelder, Lang, & Buchner, 2007), showed that 2,100 participants (i.e., 700 in each group) were required to observe a significant ($\alpha = .05$) interaction at .90 power.

Given that the effect size for the interaction observed in the pilot study was so small and unlikely to be meaningful, we did not determine the sample size according to the pilot results. Instead, we decided to recruit 90 participants in total, with 30 in each group. This sample size was similar to those in Mitchum et al.'s Experiments 2–5. To mitigate potential concerns about the sample size, we conduct Bayesian analyses to determine if the obtained results support the existence or absence of a study method by relatedness interaction. Furthermore, as discussed below, a meta-analysis was performed to integrate results across Experiments 2–4 to further mitigate potential worry about statistical power issues.

Finally, 90 participants (M age = 21.79, SD = 2.17; 72 female) were recruited from the Beijing Normal University (BNU) participant pool, and randomly allocated to the three groups, with 30 in each group.[1] They signed informed consent, were tested individually in a sound-proofed cubicle, and received financial remuneration. The protocol was approved by the Ethics Committee of BNU Faculty of Psychology.

Materials

Eighty Chinese word pairs were selected from Hu, Liu, Li, and Luo (2016). Hu and colleagues asked participants to rate the relatedness of each word pair on a scale ranging from 1 (*completely unrelated*) to 4 (*strongly related*). Half the word pairs (e.g., *doctor – nurse*) were semantically related (M relatedness rating = 3.663, SD = 0.160), and the other half (e.g., *road – table*) were unrelated (M relatedness rating = 1.203, SD = 0.125). Because the related and unrelated pairs were associated with different target words, we also compared the differences in the target words' characteristics. The results showed no detectable differences in log-transformed word frequency ($p = .171$, $BF_{10} = 0.530$), concreteness ($p = .169$, $BF_{10} = 0.658$), familiarity ($p = .413$, $BF_{10} = 0.397$), or the number of strokes ($p = .121$, $BF_{10} = 0.670$) between related and unrelated pairs (for details, see Table S1).

Sixty-four (i.e., 32 related and 32 unrelated) pairs were used in the formal experiment, with the other 16 used for practice. All stimuli were presented via Matlab *Psychtoolbox* (Kleiner, Brainard, & Pelli, 2007).

Procedure

Participants in all three groups were informed that they would study 64 word pairs in preparation for a later cued recall test. Before the experiment, participants completed a practice task to familiarize themselves with the experimental procedure.

The study procedure is depicted in Fig. 1. During the study phase, the 64 (32 related and 32 unrelated) word pairs were presented one-by-one, in random order. For the no-JOL group, before the presentation of each pair, a cross sign was shown at the center of the screen for 0.5 s to mark the inter-stimulus interval (ISI). Then, a (first) material type prompt ("Related" or "Unrelated") was shown on screen for 2 s to inform participants whether the next word pair would be a related or unrelated pair. Next, a word pair was presented on screen and participants were allowed to spend as much time as they wanted to study it. They clicked the mouse when they finished studying it. Then the word pair disappeared and the same (second) material type prompt was shown again for 2 s to inform them whether the just-studied word pair was a related or unrelated pair. This cycle repeated until participants studied all 64 pairs.

The procedure in the immediate JOL group was the same as in the no-JOL group, but with one difference. That is, after participants studied each pair, the second material type prompt ("related" or "unrelated") was presented on screen for 2 s, with a slider scale, ranging from 0 (*Sure I will not remember it*) to 100 (*Sure I will remember it*), shown below it (see Fig. 1). Participants had 2 s to drag and click the scale to make a JOL.[2] If they successfully made a JOL during the 2 s time-window, the prompt and scale remained on screen for the remainder of the 2 s to ensure that the total task duration was roughly equal among the three groups. If they did not make a JOL, a message box appeared to remind them to carefully make memory predictions for subsequent pairs. Participants clicked the mouse to remove the message box, and then the next trial started automatically.

The procedure in the pre-study JOL group was the same as in the immediate JOL group, except that the JOL slider was shown below the first instead of the second material type prompt. The pre-study JOL group had 2 s to make a JOL before studying each pair.

After participants studied all 64 pairs, they undertook a distractor task in which they solved as many mathematics problems (e.g., *7 + 45=__?*) as they could in 5 mins. Then all participants completed a cued

[1] Gender split in the current study was uneven because this study was conducted in Beijing Normal University, in which most students are female. By integrating data across Experiments 2–4, we found that gender did not affect overall memory performance, $p = .223$, $BF_{10} = 0.846$. More importantly, gender did not moderate the magnitude of the enhanced relatedness effect, $p = .367$, $BF_{10} = 0.501$, suggesting minimal difference in the enhanced relatedness effect between male and female participants. Hence, below we do not further discuss gender.

[2] We restricted the time for making JOLs at 2 s to control the total task duration among the three groups. According to other studies conducted in our laboratory, most participants can make a JOL within 2 s even when there is no time pressure. As reported in Appendix A, across Experiments 1–5, participants successfully made JOLs to most (≥ 96.8%) word pairs within 2 s.

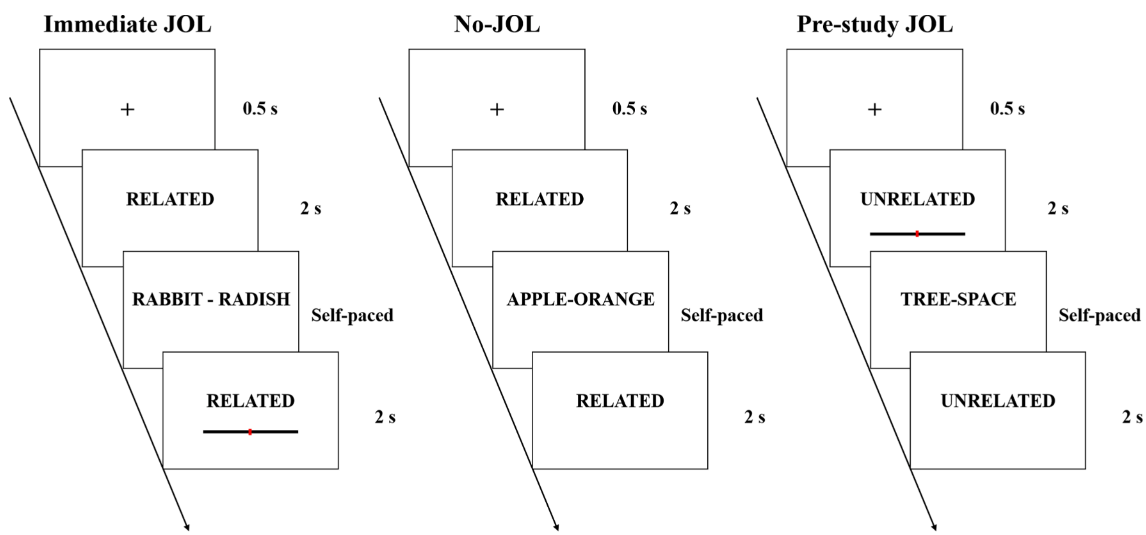

Fig. 1. Flow chart depicting the study procedure in Experiment 1. The only difference among the three groups was that participants in the no-JOL group did not make JOLs, those in the pre-study JOL group made a JOL before studying each pair, and those in the immediate JOL group made a JOL after studying each pair.

recall test (e.g., *doctor* -___), in which the 64 cue words were presented one-by-one, in random order. Participants were required to recall the target corresponding to each cue and typed their answer into a blank box. There was no time pressure and no feedback in the cued recall test.

Results

Results regarding item-by-item JOLs and study time were not a major research interest. These results are reported in Appendix A. It is worth noting that both immediate and pre-study JOLs were substantially lower for unrelated than for related pairs, suggesting that participants in both the immediate JOL and pre-study JOL groups explicitly realized that unrelated pairs were more difficult to remember than related ones. The same patterns were also observed in Experiments 2–4. For the sake of concision, we do not discuss JOL results in the following experiments.

In Experiments 1–5, test performance data were analyzed by both frequentist and Bayesian analyses of variance (ANOVAs).[3] The Bayesian ANOVAs were performed to assess whether the documented findings favor the null (H_0; e.g., absence of the interaction between study method and relatedness) or the alternative (H_1; e.g., existence of the interaction) hypothesis. The Bayes factor (BF_{10}) represents the relative strength of evidence favoring the alternative over the null hypothesis (Hoijtink, Mulder, van Lissa, & Gu, 2019; Keysers, Gazzola, & Wagenmakers, 2020). For instance, $BF_{10} = 3$ indicates that the alternative hypothesis is 3 times as likely to be true as the null hypothesis, and $BF_{10} = 0.33$ indicates that the null hypothesis is about 3 (=1/0.33) times as likely to be true as the alternative hypothesis (Barchard, 2015; Mulder & Wagenmakers, 2016; Tendeiro & Kiers, 2019).

To make it easy for readers to compare ANOVA results across experiments, we summarize them in Table 1. All ANOVAs found a main effect of relatedness (with better recall for related than for unrelated pairs) and no main effect of study method. Hence, below we do not repeatedly discuss the main effects of relatedness and study method. Instead, we focus on the interaction between relatedness and study method (i.e., the difference in the relatedness effect among different study methods). Given that all ANOVA results are already summarized in Table 1, below we only report *p* values and Bayes factors for ANOVAs.

Frequentist and Bayesian mixed ANOVAs were conducted, with relatedness (related *vs.* unrelated) as a within-subjects variable, study method (immediate JOLs *vs.* no-JOL *vs.* pre-study JOLs) as a between-subjects variable, and recall performance as the dependent variable. The results showed no interaction between relatedness and study method, $p = .619$, $BF_{10} = 0.143$ (see Fig. 2), implying strong evidence for no difference in the relatedness effect among the three groups.

Group comparison of the relatedness effect

Three mixed Bayesian ANOVAs were performed to compare the relatedness effect between each pair of groups. The first ANOVA analyzed data from the immediate JOL and no-JOL groups to test if immediate JOLs reactively enhance the relatedness effect in self-paced study conditions. Recall that Mitchum et al. (2016) and Janes et al. (2018) reported inconsistent findings regarding this question. Aligning with Janes et al.'s data, the results showed no relatedness by study method interaction, $p = .430$, $BF_{10} = 0.321$, reflecting no enhanced relatedness effect of immediate JOLs in self-paced study conditions.

The second ANOVA analyzed data from the pre-study JOL and no-JOL groups to test whether pre-study JOLs reactively enhance the relatedness effect, as predicted by the changed-goal hypothesis. The results showed no relatedness by study method interaction, $p = .381$, $BF_{10} = 0.355$, reflecting no enhanced relatedness effect of pre-study JOLs under self-paced study.

The third ANOVA analyzed data from the immediate JOL and pre-study JOL groups to establish whether there is any difference in the relatedness effect between the immediate JOL and pre-study JOL groups. The results again showed no relatedness by study method interaction, $p = .943$, $BF_{10} = 0.249$, reflecting no difference in the relatedness effect between the immediate JOL and pre-study JOL groups.

Recall of related and unrelated pairs

In each of Experiments 1–5, Bayesian *t* tests were conducted to compare recall of related and unrelated pairs between each pair of groups (Experiments 1–4) or study conditions (Experiment 5). The *t* test results are summarized in Tables 2 and 3, respectively.

As shown in Table 2, there was no statistically detectable difference in recall of related pairs, $ps \geq .451$, $BF_{10}s \leq 0.334$. In addition, as shown in Table 3, there was no difference in recall of unrelated pairs, $ps \geq .324$, $BF_{10}s \leq 0.397$.

[3] Frequentist and Bayesian analyses were performed via JASP 0.15.0.0, with all parameters set at default. Interested readers can find the default parameters in the OSF file (https://osf.io/z9cfs/).

Table 1
Frequentist and Bayesian ANOVA results in Experiments 1–5.

	F	p	η_p^2	BF_{10}
Experiment 1				
Immediate JOL vs. pre-study JOL vs. no-JOL				
Relatedness	173.021	< .001	.665	1.031×10^{20}
Study method	0.493	.612	.011	0.208
Relatedness × Study method	0.482	.619	.011	0.143
Immediate JOL vs. no-JOL				
Relatedness	116.865	< .001	.668	1.215×10^{13}
Study method	0.460	.500	.008	0.400
Relatedness × Study method	0.631	.430	.011	0.321
Pre-study JOL vs. no-JOL				
Relatedness	120.553	< .001	.675	2.929×10^{13}
Study method	0.963	.331	.016	0.450
Relatedness × Study method	0.779	.381	.013	0.355
Immediate JOL vs. pre-study JOL				
Relatedness	108.713	< .001	.652	3.041×10^{12}
Study method	0.071	.791	.001	0.343
Relatedness × Study method	0.005	.943	<.001	0.249
Experiment 2				
Immediate JOL vs. pre-study JOL vs. no-JOL				
Relatedness	610.348	< .001	.874	5.705×10^{37}
Study method	0.410	.869	.006	0.145
Relatedness × Study method	23.159	< .001	.345	1.211×10^{6}
Immediate JOL vs. no-JOL				
Relatedness	521.966	< .001	.898	2.467×10^{26}
Study method	0.238	.628	.004	0.318
Relatedness × Study method	45.309	< .001	.434	8.379×10^{5}
Pre-study JOL vs. no-JOL				
Relatedness	219.749	< .001	.791	1.350×10^{19}
Study method	0.001	.974	< .001	0.355
Relatedness × Study method	1.047	.310	.018	0.409
Immediate JOL vs. pre-study JOL				
Relatedness	566.593	< .001	.906	1.557×10^{31}
Study method	0.258	.614	.004	0.302
Relatedness × Study method	30.048	< .001	.337	1.260×10^{4}
Experiment 3				
Immediate JOL vs. pre-study JOL vs. no-JOL				
Relatedness	318.027	< .001	.785	2.49×10^{25}
Study method	0.217	.805	.005	0.158
Relatedness × Study method	15.137	< .001	.258	6826.416
Immediate JOL vs. no-JOL				
Relatedness	222.299	< .001	.793	3.90×10^{16}
Study method	0.424	.518	.007	0.345
Relatedness × Study method	27.176	< .001	.319	5090.983
Pre-study JOL vs. no-JOL				
Relatedness	127.233	< .001	.687	1.62×10^{13}
Study method	0.118	.733	.002	0.400
Relatedness × Study method	2.345	.131	.039	0.655
Immediate JOL vs. pre-study JOL				
Relatedness	314.469	< .001	.844	8.06×10^{22}
Study method	0.101	.751	.002	0.303
Relatedness × Study method	15.855	< .001	.215	112.285
Experiment 4				
Immediate JOL vs. pre-study JOL vs. no-JOL				
Relatedness	379.158	< .001	.812	3.733×10^{29}
Study method	0.410	.665	.009	0.169
Relatedness × Study method	13.875	< .001	.240	3057.103
Immediate JOL vs. no-JOL				
Relatedness	294.440	< .001	.833	2.231×10^{21}
Study method	0.918	.342	.015	0.417
Relatedness × Study method	24.755	< .001	.296	2427.438
Pre-study JOL vs. no-JOL				
Relatedness	155.849	< .001	.725	8.380×10^{15}
Study method	0.248	.620	.004	0.331
Relatedness × Study method	0.647	.424	.011	0.338
Immediate JOL vs. pre-study JOL				
Relatedness	333.614	< .001	.852	1.029×10^{23}
Study method	0.138	.712	.002	0.341
Relatedness × Study method	17.617	< .001	.233	212.456
Experiment 5				
JOL vs. control				
Relatedness	296.866	< .001	.814	3.601×10^{22}
Condition	0.320	.573	.005	0.331
Relatedness × Study method	21.061	< .001	.236	843.887
No-JOL vs. control				
Relatedness	161.246	< .001	.703	7.131×10^{16}
Condition	0.649	.423	.009	0.424
Relatedness × Study method	0.877	.352	.013	0.342
JOL vs. no-JOL				
Relatedness	137.871	< .001	.802	2.176×10^{26}
Condition	0.631	.432	.018	0.214
Relatedness × Study method	26.840	< .001	.441	35.381

Correlation between study time and strength of cue-target relatedness

Following Mitchum et al. (2016) and Janes et al. (2018), for each participant, we calculated a Spearman rank correlation coefficient (r_s) between study time and cue-target relatedness, that is, ratings of cue-target relatedness strength taken from Hu et al. (2016). Bayesian and frequentist t-tests found little difference in r_s scores between the immediate JOL (M = -0.420, SD = 0.191) and no-JOL (M = -0.379, SD = 0.231) groups, difference = -0.041, 95 % CI [-0.150, 0.069], $t(58)$ = -0.747, p = .458, d = -0.193, BF_{10} = 0.332. In the same manner, there was minimal difference in r_s scores between the immediate JOL and pre-study JOL (M = -0.389, SD = 0.176) groups, difference = -0.031, 95 % CI [-0.126, 0.064], $t(58)$ = -0.660, p = .512, d = -0.170, BF_{10} = 0.315. Lastly, there was negligible difference in r_s scores between the no-JOL and pre-study JOL groups, difference = 0.010, 95 % CI [-0.096, 0.116], $t(58)$ = 0.181, p = .857, d = 0.047, BF_{10} = 0.266. Overall, these results indicate that there is minimal difference in the patterns of study time allocation among the three groups, inconsistent with the changed-goal hypothesis.

Discussion

Consistent with Janes et al. (2018) but inconsistent with Mitchum et al. (2016), Experiment 1 found that immediate JOLs do not reactively enhance the relatedness effect under self-paced study conditions. Furthermore, pre-study JOLs have minimal reactive impact on the relatedness effect under these conditions. Additionally, consistent with Janes et al. (2018), Experiment 1 failed to replicate Mitchum et al.'s (2016) study time findings by showing no difference in correlations between study time and relatedness strength across the three groups. These findings jointly challenge the changed-goal hypothesis. In the General Discussion, we discuss why immediate JOLs do not enhance the relatedness effect in self-paced study conditions.

Experiment 2

Mitchum et al., (2016, Experiment 5) observed an enhanced relatedness effect with an experimenter-paced study procedure. Similarly, Janes et al. (2018) consistently detected an enhanced relatedness effect of immediate JOLs in their Experiments 2–3, in which the study procedure was experimenter-paced (for related findings, see Myers et al., 2020; Rivers et al., 2021; Soderstrom et al., 2015). Mitchum et al. (2016) proposed that, in experimenter-paced study conditions, learners cannot fully memorize all items in a limited study period, creating pressure on word pair learning and leading to a shift from a mastery orientation (i.e., mastering all items) toward a concentration on easy related pairs (Metcalfe & Kornell, 2003). This might be the reason why the requirement of making immediate JOLs produces an enhanced relatedness effect in experimenter-paced conditions.

Following prior studies (Janes et al., 2018; Mitchum et al., 2016), Experiment 2 employed an experimenter-paced study procedure to further test the changed-goal hypothesis.

Method

Design and participants

Experiment 2 involved the same 3 × 2 mixed design as Experiment 1.

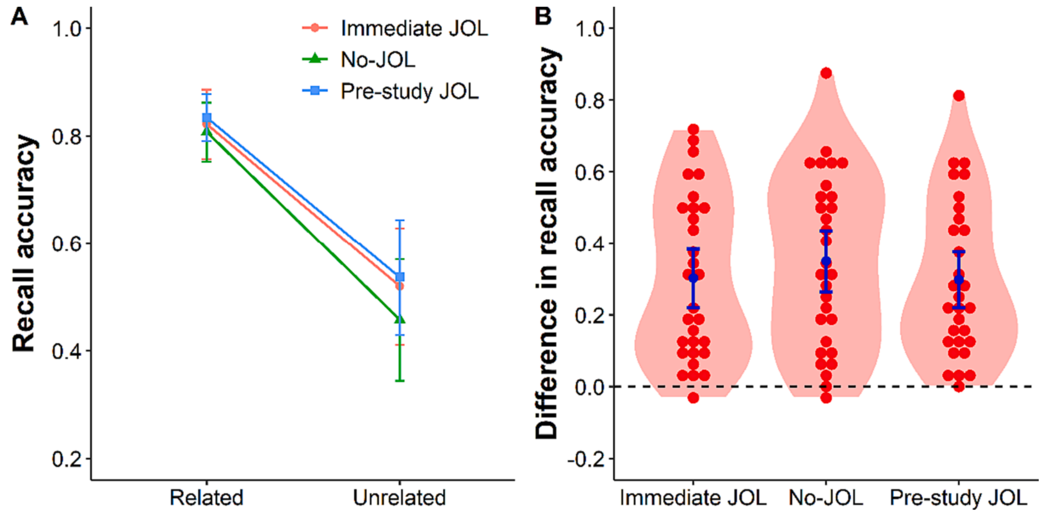

Fig. 2. A: Recall accuracy as a function of study method and relatedness in Experiment 1. B: Difference in recall accuracy between related and unrelated word pairs (i.e., the relatedness effect) as a function of study method. In Panel B, red points represent the difference score in recall between related and unrelated word pairs for each participant, and the blue points represent group averages. Error bars represent 95% CI. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Difference in recall of related pairs between pairs of groups or conditions in Experiments 1–5.

	Difference [95 % CI]	t	p	d	BF_{10}
Experiment 1					
Immediate JOL *vs.* no-JOL	0.015 [-0.072, 0.101]	0.339	.736	0.088	0.275
Pre-study JOL *vs.* no-JOL	−0.027 [-0.098, 0.044]	−0.760	.451	−0.196	0.334
Immediate JOL *vs.* pre-study JOL	−0.013 [-0.092, 0.067]	−0.315	.754	−0.081	0.274
Experiment 2					
Immediate JOL *vs.* no-JOL	0.107 [0.038, 0.175]	3.118	.003	0.798	13.073
Pre-study JOL *vs.* no-JOL	−0.021 [-0.097, 0.056]	−0.545	.588	−0.141	0.297
Immediate JOL *vs.* pre-study JOL	0.086 [0.028, 0.144]	2.961	.004	0.758	9.059
Experiment 3					
Immediate JOL *vs.* no-JOL	0.095 [0.027, 0.163]	2.797	.007	0.722	6.273
Pre-study JOL *vs.* no-JOL	0.019 [-0.063, 0.100]	0.460	.647	0.119	0.287
Immediate JOL *vs.* pre-study JOL	0.076 [0.010, 0.143]	2.288	.026	0.591	2.255
Experiment 4					
Immediate JOL *vs.* no-JOL	0.074 [0.011, 0.137]	2.338	.023	0.599	2.472
Pre-study JOL *vs.* no-JOL	0.004 [-0.068, 0.076]	0.114	.909	0.029	0.262
Immediate JOL *vs.* pre-study JOL	0.078 [-0.008, 0.148]	2.238	.029	0.578	2.062
Experiment 5					
JOL *vs.* control	0.077 [0.005, 0.149]	2.124	.037	0.508	1.647
No-JOL *vs.* control	−0.018 [-0.101, 0.065]	−0.428	.670	−0.102	0.266
JOL *vs.* no-JOL	0.095 [0.042, 0.148]	3.633	$<$ .001	0.614	34.384

Table 3
Difference in recall of unrelated pairs between pairs of groups or conditions in Experiments 1–5.

	Difference [95 % CI]	t	p	d	BF_{10}
Experiment 1					
Immediate JOL *vs.* no-JOL	0.063 [-0.098, 0.223]	0.780	.439	0.201	0.339
Pre-study JOL *vs.* no-JOL	0.079 [-0.080, 0.238]	0.995	.324	0.257	0.397
Immediate JOL *vs.* pre-study JOL	−0.017 [-0.173, 0.140]	−0.213	.832	−0.055	0.267
Experiment 2					
Immediate JOL *vs.* no-JOL	−0.146 [-0.251, −0.041]	−2.772	.007	−0.710	5.994
Pre-study JOL *vs.* no-JOL	−0.024 [-0.102, 0.150]	−0.379	.706	−0.098	0.279
Immediate JOL *vs.* pre-study JOL	−0.122 [-0.220, 00.024]	−2.486	.016	−0.637	3.283
Experiment 3					
Immediate JOL *vs.* no-JOL	−0.158 [-0.297, −0.020]	−0.372	.026	−0.590	2.250
Pre-study JOL *vs.* no-JOL	−0.055 [-0.199, 0.088]	−0.770	.444	−0.199	0.337
Immediate JOL *vs.* pre-study JOL	−0.103 [-0.222, 0.016]	−1.736	.088	−0.448	0.917
Experiment 4					
Immediate JOL *vs.* no-JOL	−0.154 [-0.273, −0.035]	−2.591	.012	−0.664	4.062
Pre-study JOL *vs.* no-JOL	−0.043 [-0.088, 0.173]	−0.653	.516	−0.167	0.312
Immediate JOL *vs.* pre-study JOL	−0.111 [-0.235, 0.012]	−1.801	.077	−0.465	1.008
Experiment 5					
JOL *vs.* control	−0.127 [-0.246, −0.007]	−2.120	.038	−0.507	1.633
No-JOL *vs.* control	−0.063 [-0.194, 0.069]	−0.945	.348	−0.226	0.360
JOL *vs.* no-JOL	−0.064 [-0.111, −0.018]	−2.814	.008	−0.476	5.129

A pilot study, with 10 participants in each group, showed that the effect size for the study method by relatedness interaction was $\eta_p^2 = .205$. A power analysis found that 18 participants in each group were required to observe a significant ($\alpha = .05$) interaction at .90 power. To be more conservative, we decided to increase the sample size to 30 in each group – the same as in Experiment 1.

In total, 95 participants were recruited from the BNU participant pool. Data from four participants were lost due to computer failure, leaving final data from 91 participants (*M* age = 20.50, *SD* = 2.32; 83 female). They were randomly allocated to the three groups, with 30 in the no-JOL, 31 in the immediate JOL, and 30 in the pre-study JOL group. They gave informed consent, were tested individually in a sound-proofed cubicle, and received financial remuneration.

Materials and procedure

The materials and experimental procedure were the same as in Experiment 1, except that the study procedure was changed to experimenter-paced. The presentation duration for each pair was fixed at 6 s during the study phase.

Results

Test performance data were analyzed in the same way as in Experiment 1. As revealed in Table 1 and Fig. 3, across the three groups, there was a substantial interaction between relatedness and study method, $p < .001$, $BF_{10} = 1.211 \times 10^6$.

Group comparison of the relatedness effect

As shown in Table 1 and Fig. 3, the relatedness effect was greater in the immediate JOL than in the no-JOL group, $p < .001$, $BF_{10} = 8.379 \times 10^5$, reflecting an enhanced relatedness effect of immediate JOLs and replicating the findings of Janes et al. (2018, Experiments 2 and 3). However, there was no statistically detectable difference in the relatedness effect between the pre-study JOL and no-JOL groups, $p = .310$, $BF_{10} = 0.409$, suggesting that pre-study JOLs tend not to enhance the relatedness effect and running counter to the changed-goal hypothesis. Furthermore, the relatedness effect was substantially greater in the immediate JOL than in the pre-study JOL group, $p < .001$, $BF_{10} = 1.260 \times 10^4$, again challenging the changed-goal hypothesis.

Recall of related and unrelated pairs

The data in Table 2 indicate that recall of related pairs was significantly greater in the immediate JOL than in both the no-JOL ($p = .003$, $BF_{10} = 13.073$) and pre-study JOL ($p = .004$, $BF_{10} = 9.059$) groups. By contrast, there was minimal difference between the no-JOL and pre-study JOL groups ($p = .558$, $BF_{10} = 0.297$).

As shown in Table 3, recall of unrelated pairs was poorer in the immediate JOL than in the no-JOL group ($p = .007$, $BF_{10} = 5.944$), and poorer than in the pre-study JOL group ($p = .016$, $BF_{10} = 3.283$). By contrast, there was little difference between the pre-study JOL and no-JOL groups ($p = .706$, $BF_{10} = 0.279$).

Discussion

Experiment 2 found that, when the study procedure is experimenter-paced, only immediate JOLs, but not pre-study JOLs, reactively enhance the relatedness effect, facilitate recall of related pairs, and impair recall of unrelated ones. These findings contradict the changed-goal hypothesis's prediction that pre-study JOLs should also induce an enhanced relatedness effect. The results also indicate that immediate JOLs are associated with a substantially larger relatedness effect than pre-study JOLs, which challenges the prediction of the changed-goal hypothesis that the enhanced relatedness effect of pre-study JOLs should be equal to (or even larger than) the effect of immediate JOLs. Overall, the findings of Experiment 2 jointly challenge the changed-goal hypothesis.

Experiment 3

Experiment 3 was conducted to conceptually replicate the main findings of Experiment 2. Additionally, it aimed to address a potential limitation of that experiment. In Experiment 2, the no-JOL group received material type prompts ("Related" or "Unrelated") before and after studying each pair. Such material type prompts might enhance awareness of the difference in learning difficulty between related and unrelated pairs, leading to a change of study goal. This might be the reason why Experiment 2 observed no enhanced relatedness effect of pre-study JOLs. On the other hand, Experiment 2 did observe a reliable enhanced relatedness effect of immediate JOLs.

To mitigate concerns about this potential limitation of Experiment 2, Experiment 3 removed the material type prompts in the no-JOL group. In addition, the pre-study JOL group only received material type prompts before studying each pair while the immediate JOL group only received prompts after studying each pair (see below for details).

Method

Design and participants

Experiment 3 involved the same 3 × 2 mixed design as Experiments 1 and 2. According to Experiment 2, the effect size for the study method by relatedness interaction was $\eta_p^2 = .345$. A power analysis showed that 10 participants in each group were required to observe a significant ($\alpha = .05$) interaction at .90 power. To be more conservative, we decided to increase the sample size to 30 in each group – the same as in Experiments 1 and 2.

In total, 90 participants (M age = 21.033, SD = 2.052; 86 female)

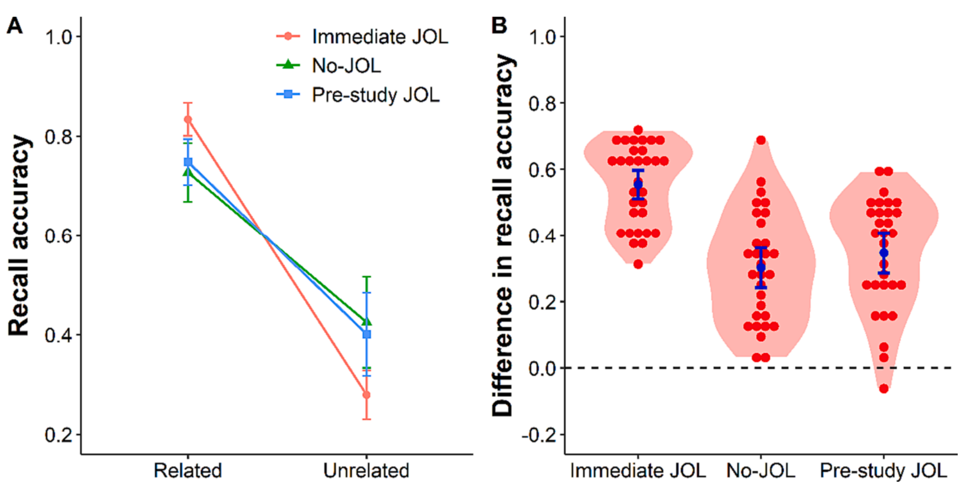

Fig. 3. A: Recall accuracy as a function of study method and relatedness in Experiment 2. B: Difference in recall accuracy between related and unrelated word pairs (i.e., the relatedness effect) as a function of study method. In Panel B, red points represent the difference score in recall between related and unrelated word pairs for each participant, and the blue points represent group averages. Error bars represent 95% CI. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

were recruited from the BNU participant pool and were randomly allocated to the three groups, with 30 in each group. They gave informed consent, were tested individually in a sound-proofed cubicle, and received financial remuneration.

Materials and procedure

The materials and procedure were the same as in Experiment 2, but with several minor changes. Specifically, in the no-JOL group, the first and second material type prompts ("Related" or "Unrelated") were replaced by a mask prompt ("##"), which means that participants in the no-JOL group were not informed of the material type before and after studying each pair. In the immediate JOL group, the first material type prompt was replaced by a mask ("##") and participants only received the second prompt ("Related" or "Unrelated") after studying each pair. By contrast, the pre-study JOL group only received the first material type prompt ("Related" or "Unrelated") before studying each pair, with the second prompt replaced by a mask ("##").

Results

The major results were the same as in Experiment 2. Across the three groups, there was a reliable interaction between relatedness and study method, $p < .001$, $BF_{10} = 6{,}826$.

Group comparison of the relatedness effect

As shown in Table 1 and Fig. 4, the relatedness effect was larger in the immediate JOL than in the no-JOL group, $p < .001$, $BF_{10} = 5{,}090$, reflecting an enhanced relatedness effect of immediate JOLs and replicating the findings of Experiment 2. However, there was minimal difference in the relatedness effect between the pre-study JOL and no-JOL groups, $p = .131$, $BF_{10} = 0.655$, reflecting that pre-study JOLs do not enhance the relatedness effect and running counter to the changed-goal hypothesis. Furthermore, the relatedness effect was substantially larger in the immediate JOL than in the pre-study JOL group, $p < .001$, $BF_{10} = 112.285$, again challenging the changed-goal hypothesis.

Recall of related and unrelated pairs

Table 2 reveals that recall of related pairs was greater in the immediate JOL than in both the no-JOL ($p = .007$, $BF_{10} = 6.273$) and pre-study JOL ($p = .026$, $BF_{10} = 2.255$) groups. By contrast, there was little difference in recall of related pairs between the no-JOL and pre-study JOL groups ($p = .647$, $BF_{10} = 0.287$).

As shown in Table 3, recall of unrelated pairs was poorer in the immediate JOL than in the no-JOL group ($p = .026$, $BF_{10} = 2.250$), and marginally poorer than in the pre-study JOL group ($p = .088$, $BF_{10} = 0.917$). However, there was little difference in recall of unrelated pairs between the pre-study JOL and no-JOL groups ($p = .444$, $BF_{10} = 0.337$).

Discussion

Experiment 3 successfully replicated the main findings of Experiment 2 by showing that, with experimenter-paced study, only immediate JOLs, but not pre-study JOLs, reactively enhance the relatedness effect, facilitate recall of related pairs, and impair recall of unrelated ones. These findings again disprove the two reactivity predictions about pre-study JOLs generated by the changed-goal hypothesis.

Experiment 4

Experiment 4 was pre-registered to conceptually replicate the main findings of Experiments 2 and 3. Registration information is available at https://osf.io/qx5f9.

In Experiments 2 and 3, participants in the pre-study JOL group were informed in advance about whether the next pair would be a "related" or "unrelated" pair. Such material type prompts might be too implicit to enhance awareness of the difference in learning difficulty between related and unrelated pairs, which might be the reason why pre-study JOLs did not enhance the relatedness effect in these two experiments. However, participants did provide substantially higher pre-study and immediate JOLs for related than for unrelated pairs (see Appendix A), which means that, in both groups, participants did explicitly realize that unrelated pairs were much more difficult to remember than related ones.

To mitigate concerns about this potential concern in Experiments 2 and 3, Experiment 4 replaced the implicit material type prompts ("Related" or "Unrelated") with explicit ones ("Easy" or "Difficult"). Specifically, in Experiment 4, participants in the pre-study JOL group were overtly informed whether the next word pair was "Easy" or "Difficult" to remember before studying each pair. According to the changed-goal hypothesis, such overt prompts plus pre-study JOLs should jointly change participants' study goal and elicit an enhanced relatedness effect.

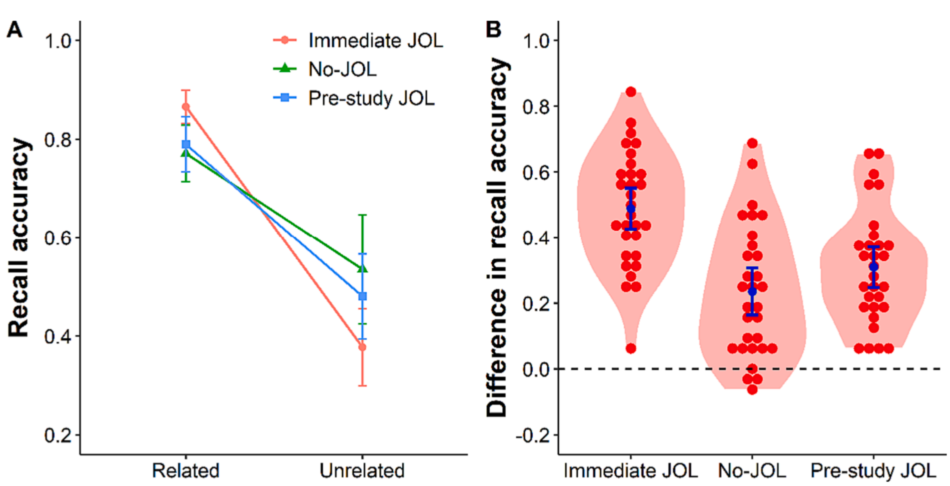

Fig. 4. A: Recall accuracy as a function of study method and relatedness in Experiment 3. B: Difference in recall accuracy between related and unrelated word pairs (i.e., the relatedness effect) as a function of study method. In Panel B, red points represent the difference score in recall between related and unrelated word pairs for each participant, and the blue points represent group averages. Error bars represent 95% CI. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Method

Design and participants

Experiment 4 involved the same 3 × 2 mixed design as Experiments 1–3. According to Experiment 3, the interaction between study method and relatedness was $\eta_p^2 = .258$. A power analysis showed that 14 participants in each group were required to observe a significant ($\alpha = .05$) interaction at .90 power. To be more conservative, we decided to increase the sample size to 30 in each group – the same as in Experiments 1–3.

In total, 93 participants were recruited from the BNU participant pool. Data from two participants were lost due to computer failure, leaving final data from 91 participants (M age = 21.923, SD = 2.339; 85 female). They were randomly allocated to the three groups, with 31 in the no-JOL, 30 in the pre-study JOL, and 30 in the immediate JOL group. All participants gave informed consent, were tested individually in a sound-proofed cubicle, and received financial remuneration.

Materials and procedure

The materials and procedure were the same as in Experiment 3, except that the "Related" and "Unrelated" prompts were replaced by "Easy" and "Difficult" prompts. Participants in the pre-study JOL group were pre-informed whether the next word pair was easy or difficult to remember before studying each pair, those in the immediate JOL group were told whether the just-studied pair was easy or difficult to remember after studying each pair, and those in the no-JOL group received mask prompts ("##").

Results

There are no discrepancies between the pre-registered and reported data analyses. All results revealed the same patterns as in Experiments 2 and 3. Among the three groups, there was a reliable interaction between relatedness and study method, $p < .001$, $BF_{10} = 3{,}057$.

Group comparison of the relatedness effect

As shown in Table 1 and Fig. 5, the relatedness effect was greater in the immediate JOL than in the pre-study JOL group, $p < .001$, $BF_{10} =$ 2,427, reflecting an enhanced relatedness effect of immediate JOLs and replicating the findings of Experiments 2 and 3. However, there was minimal difference in the relatedness effect between the pre-study JOL and no-JOL groups, $p = .424$, $BF_{10} = 0.338$, reflecting that pre-study JOLs do not enhance the relatedness effect and running counter to the changed-goal hypothesis. Furthermore, the relatedness effect was larger in the immediate JOL than in the pre-study JOL group, $p < .001$, $BF_{10} =$ 212.456, again challenging the changed-goal hypothesis.

Recall of related and unrelated pairs

Table 2 reveals that recall of related pairs was greater in the immediate JOL than in both the no-JOL ($p = .023$, $BF_{10} = 2.472$), and pre-study JOL ($p = .029$, $BF_{10} = 2.062$) groups. There was little difference between the no-JOL and pre-study JOL groups ($p = .909$, $BF_{10} = 0.262$).

As shown in Table 3, recall of unrelated pairs was poorer in the immediate JOL than in the no-JOL group ($p = .012$, $BF_{10} = 4.062$), and marginally poorer than in the pre-study JOL group ($p = .077$, $BF_{10} =$ 1.008). By contrast, there was little difference between the no-JOL and pre-study JOL groups ($p = .516$, $BF_{10} = 0.312$).

Discussion

Experiment 4 adopted overt material type prompts ("Easy" or "Difficult") and successfully replicated the main findings of Experiments 2 and 3. The observed findings again run counter to the changed-goal hypothesis.

Meta-analyses of Experiments 2–4

It is reasonable to assume that the null findings of the enhanced relatedness effect in the pre-study JOL groups in Experiment 2–4 might derive from inadequate statistical power, given that the sample sizes in these experiments were relatively small (i.e., about 30 participants per group). Although Experiments 2–4 consistently found Bayesian evidence of no enhanced relatedness effect of pre-study JOLs and the sample sizes in these three experiments were pre-determined according to power analyses, we conducted three meta-analyses to integrate data from 272 participants across Experiments 2–4 to further mitigate potential worry about statistical power.

For each participant in Experiments 2–4, we calculated the mean difference in cued recall performance between related and unrelated pairs to represent the magnitude of the relatedness effect. Next, for each experiment, a mean difference in the relatedness effect between every pair of groups and its associated variance was calculated, and these were then submitted to the JASP software to perform Bayesian model-averaged meta-analyses (Berkhout, Haaf, Gronau, Heck, & Wagenmakers, 2023; Gronau, Heck, Berkhout, Haaf, & Wagenmakers, 2021). Bayesian model-averaged meta-analysis considers the evidence for four relevant models (including the fixed-effects null hypothesis, the fixed-effects alternative hypothesis, the random-effects null hypothesis, and

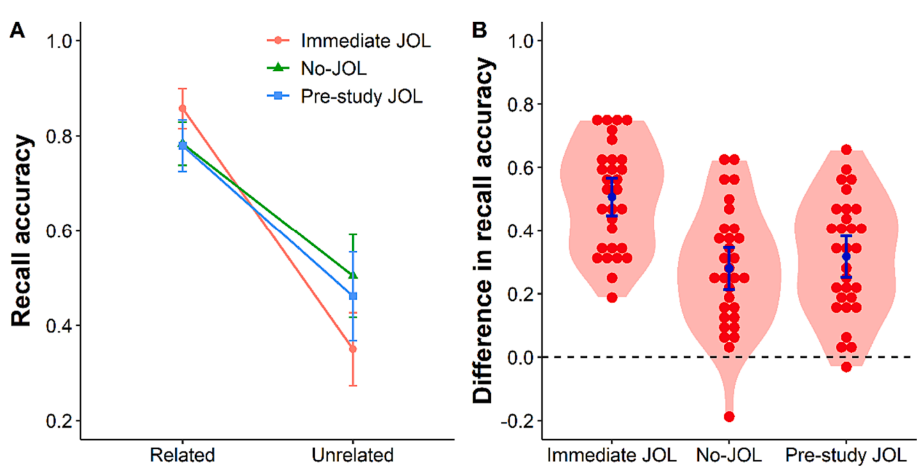

Fig. 5. A: Recall accuracy as a function of study method and relatedness in Experiment 4. B: Difference in recall accuracy between related and unrelated word pairs (i.e., the relatedness effect) as a function of study method. In Panel B, red points represent the difference score in recall between related and unrelated word pairs for each participant, and the blue points represent group averages. Error bars represent 95% CI. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the random-effects alternative hypothesis), and provides overall estimates for the effect size by aggregating across the four models weighted by their plausibility.

The results showed that the relatedness effect in the immediate JOL groups was substantially greater than that in the no-JOL groups, difference = 0.244 [0.188, 0.301], BF_{10} = 58.794. However, critically, there was minimal difference in the relatedness effect between the pre-study JOL and no-JOL groups, difference = 0.052 [-0.009, 0.113], BF_{10} = 0.179, contrary to the changed-goal hypothesis. Furthermore, the relatedness effect was substantially greater in the immediate JOL than in the pre-study JOL groups, difference = 0.193 [0.137, 0.248], BF_{10} = 28.321, again challenging the changed-goal hypothesis. Overall, these meta-analytic results are consistent with those from the ANOVAs reported above for Experiments 2–4.

Before moving forward, it should be highlighted that the difference (0.244, BF_{10} = 58.794) in the relatedness effect between the immediate JOL and no-JOL groups was substantially larger than the difference (0.052, BF_{10} = 0.179) between the pre-study JOL and no-JOL groups, which suggests that, even if goal change does contribute to the enhanced relatedness effect of immediate JOLs, this contribution is small. Put differently, goal change cannot be the main mechanism underlying the enhanced relatedness effect of immediate JOLs.

Experiment 5

Experiments 1–4 tested the changed-goal hypothesis by contrasting the enhanced relatedness effect between immediate and pre-study JOLs. Another approach to test the changed-goal hypothesis is to investigate the transfer of the enhanced relatedness effect of immediate JOLs. Imagine that participants study a mixed list of related and unrelated pairs and are required to provide immediate JOLs for half the pairs (i.e., JOL pairs) but not for the other half (i.e., no-JOL pairs), with JOL and no-JOL pairs studied in a randomly interleaved manner (e.g., a JOL pair, a no-JOL pair, a no-JOL pair, a JOL pair, a no-JOL pair…). Before studying each pair, they are explicitly informed whether the next pair will be easy or difficult to remember. Hence, participants explicitly know the difficulty of each pair regardless of whether they need to make a JOL for that pair or not. In such a situation, the changed-goal hypothesis generates two predictions regarding reactivity and transfer: (1) Immediate JOLs should induce an enhanced relatedness effect for JOL pairs, for the reasons already explained (Janes et al., 2018; Mitchum et al., 2016), and (2) the enhanced relatedness effect should transfer to no-JOL pairs.

The logic behind the transfer prediction is straightforward. If immediate JOLs truly change study goals and induce a shift in study strategy (i.e., allocating more resources toward remembering related pairs with a commensurate sacrifice for unrelated ones), participants should continue using the shifted strategy to study no-JOL pairs, because (1) they explicitly know whether a given no-JOL pair is easy or difficult to remember even when they do not need to make a JOL, and (2) according to the changed-goal hypothesis, they believe that allocating more resources toward studying related pairs at the cost of unrelated ones will be beneficial for memory performance overall. Numerous studies have established that, when a learner believes that a given strategy is beneficial for their memory and actively adopts this strategy for studying some items, they will continue using this strategy to study other items even when they are not explicitly prompted to do so and even after a long delay (Hui, de Bruin, Donkers, & van Merriënboer, 2021; Sahakyan, Delaney, & Kelley, 2004; Sun et al., 2022).

Overall, the changed-goal hypothesis predicts that the enhanced relatedness effect of immediate JOLs should transfer to no-JOL pairs when JOL and no-JOL pairs are interleaved. Experiment 5 was pre-registered to test the changed-goal hypothesis by assessing this transfer prediction. The pre-registration is available at https://osf.io/fr6uj.

Before moving forward, it should be noted that Rivers et al. (2021) recently adopted a similar interleaving procedure. Their purpose in doing so was to explore whether the reactivity effect survives when JOL and no-JOL pairs are intermixed within-subjects. In their study, participants made JOLs for half of the pairs but not for the other half. Rivers et al. (2021, Experiment 3) observed a larger relatedness effect for JOL than for no-JOL pairs, suggesting that JOL reactivity survives in within-subjects design experiments. However, because Rivers et al. (2021) did not compare the relatedness effect in no-JOL pairs to a baseline of the effect in the control group (in which participants did not make JOLs for any pairs), their results do not shed light on the current question, namely whether the enhanced relatedness effect transfers to no-JOL pairs.

Experiment 5 was specifically designed to test the transfer prediction of the changed-goal hypothesis. To achieve this aim, it included a control group in which participants did not make JOLs to any pairs. The relatedness effect in the control group was set as a benchmark to determine whether the enhanced relatedness effect transfers to no-JOL pairs (that is, whether the relatedness effect for no-JOL pairs in the experimental group is larger than the equivalent effect in the control group).

Method

Design and participants

Experiment 5 involved a 2 (group: experimental *vs.* control) × 2 (relatedness: related *vs.* unrelated) mixed design, with group as a between-subjects variable and relatedness as a within-subjects variable. Additionally, the experimental group involved a within-subjects design (study condition: JOL *vs.* no-JOL). That is, in the experimental group, half the pairs were studied with immediate JOLs, with the other half studied without immediate JOLs. JOL and no-JOL pairs were presented in an interleaved order. In the control group, all pairs were studied without immediate JOLs.

A pilot study, with 10 participants in each group, showed a larger relatedness effect in the experimental group's JOL condition than in the control group, $\eta_p^2 = .134$. A power analysis indicated that 35 participants in each group were required to observe a significant ($\alpha = .05$) difference in the relatedness effect at .90 power.

The pilot study showed a negligible difference in the relatedness effect between the no-JOL condition in the experimental group and the control group, $\eta_p^2 < .001$. Because this effect size was so small, we did not determine the sample size according to this effect size.

The pilot study also showed a larger relatedness effect in the JOL than no-JOL condition in the experimental group, $\eta_p^2 = .467$. A power analysis suggested that 16 participants in the experimental group were required to observe a significant ($\alpha = .05$) difference in the relatedness effect between these two conditions at .90 power.

Based on these power analysis results, our pre-registered plan was to recruit 70 participants in total, with 35 in each group. Finally, 70 participants (*M* age = 21.79, *SD* = 2.35; 65 female) were recruited from the BNU participant pool and were randomly allocated to the two groups, with 35 in each group. They gave informed consent, were tested individually in a sound-proofed cubicle, and received financial remuneration.

Materials

The study stimuli were the same as in Experiments 1–4. To avoid any item-selection effects, for each participant the computer randomly selected 32 (16 related and 16 unrelated) word pairs to present in red, with the other 32 shown in blue. The presentation sequence of all 64 pairs was randomly decided by the computer.

Design and procedure

Participants in both groups were informed that they would study 64 word pairs in preparation for a later memory test, with 32 pairs presented in red and the other 32 shown in blue. Half the participants in the experimental group were told that they needed to make memory predictions for red but not for blue pairs. The remaining participants were told to make predictions for blue but not for red pairs. Participants in the control group did not receive these memory prediction instructions. In addition, both groups were informed that they should endeavor to memorize all pairs regardless of whether they were shown in red or blue because all of them would be eventually tested. Before the experiment, participants completed a practice task to familiarize themselves with the procedure.

The study procedure is depicted in Fig. 6. The 64 word pairs were presented one-by-one, in random order. For each trial in the control group, a cross sign was shown on the screen for 0.5 s. Next, a mask prompt ("##") was shown for 2 s. Then, a word pair was presented for 6 s, in either red or blue. Then the mask prompt was shown again for 2 s. This cycle repeated until participants studied all 64 pairs.

The procedure in the experimental group was identical to that in the control group, but with two differences. First, the mask prompts were replaced by overt material type prompts ("Easy" or "Difficult"). These prompts were included to ensure that participants in the experimental group explicitly knew whether a given word pair was easy or difficult to remember regardless of whether they needed to make a JOL for that pair.

The second difference was that, for half the word pairs (e.g., red pairs) in the experimental group, a JOL slider was shown below the second material type prompt, and participants were instructed to make an immediate JOL within 2 s. For the other pairs (e.g., blue pairs), the JOL slider was not shown, and participants did not need to make an immediate JOL.

After participants studied all 64 pairs, they undertook a distractor task and a cued recall test, which were identical to those in Experiments 1–4.

Results

There are no discrepancies between the pre-registered and reported data analyses. For the sake of brevity, below we term the experimental group's JOL condition as the JOL condition, the experimental group's no-JOL condition as the no-JOL condition, and the control group as the control condition. Three Bayesian ANOVAs were conducted to compare the relatedness effect between each pair of conditions and the ANOVA results are summarized in Table 1.

As shown in Table 1 and Fig. 7, the relatedness effect was larger in the JOL than in the control condition, $p < .001$, $BF_{10} = 843.887$, replicating the enhanced relatedness effect of immediate JOLs. However, there was minimal difference in the relatedness effect between the no-JOL and control conditions, $p = .352$, $BF_{10} = 0.342$, implying minimal transfer of the enhanced relatedness effect and challenging the changed-goal hypothesis. Furthermore, the relatedness effect was substantially larger in the JOL than in the no-JOL condition, $p < .001$, $BF_{10} = 35.381$, again challenging the changed-goal hypothesis.

Recall of related and unrelated pairs

Pre-planned Bayesian t tests were performed to investigate the difference in recall of related and unrelated pairs between conditions. The detailed results are summarized in Tables 2 and 3, respectively.

The data in Table 2 reveal that recall of related pairs was greater in the JOL than in both the control ($p = .037$, $BF_{10} = 1.647$, even though the Bayesian evidence for this difference is relatively weak), and no-JOL conditions ($p < .001$, $BF_{10} = 34.384$). However, there was little

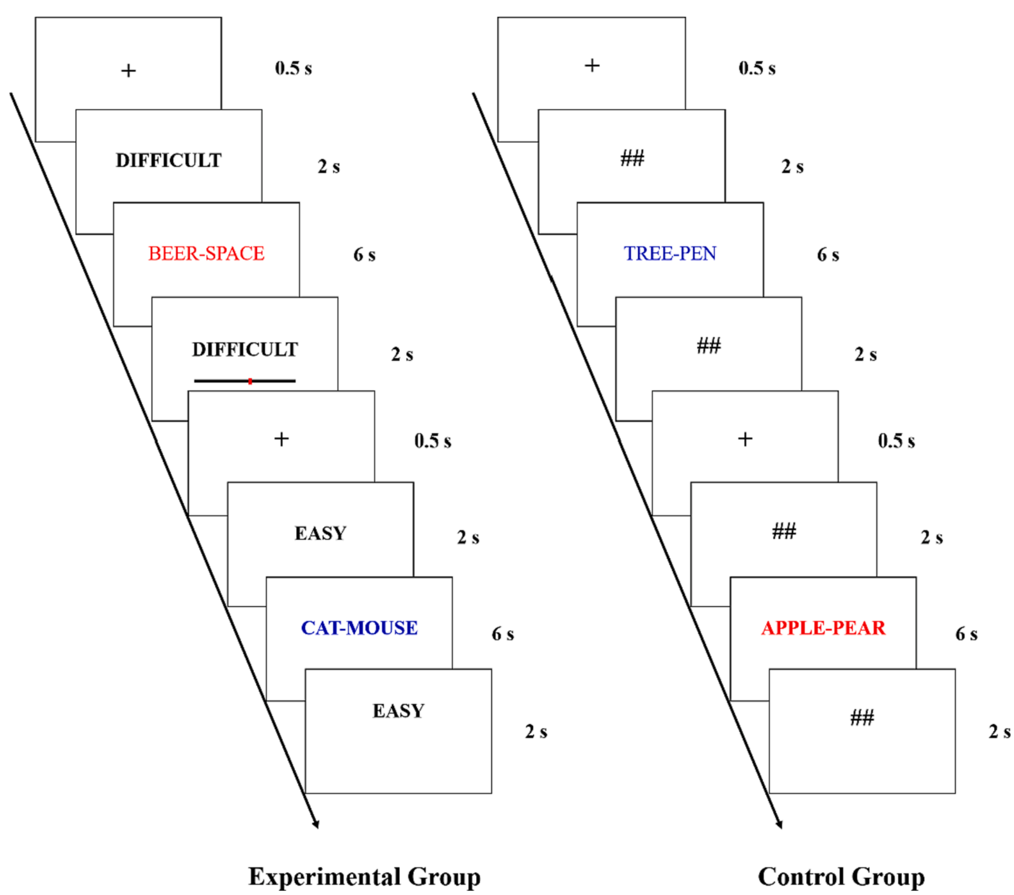

Fig. 6. Flow chart depicting the study procedure in the experimental (left) and control (right) groups in Experiment 5. Participants in the experimental JOL group made JOLs for half the word pairs but not for the other half, whereas those in the control group did not make JOLs for any pairs. The experimental group received material type prompts before and after studying each pair, whereas the control group received mask prompts.

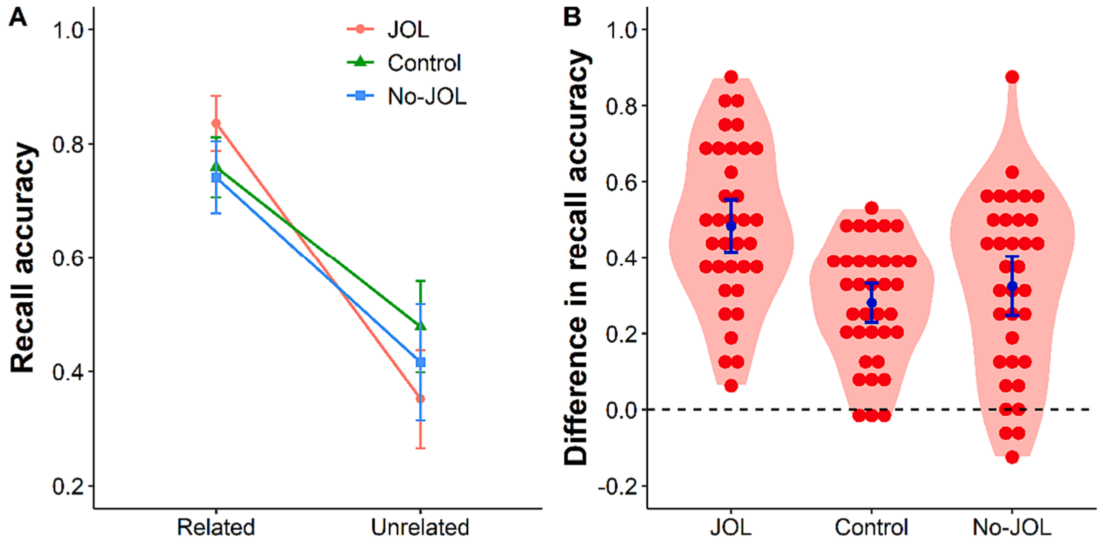

Fig. 7. A: Recall accuracy as a function of condition and relatedness in Experiment 5. The JOL and no-JOL conditions refer to trials with and without JOLs, respectively, in the experimental group. B: Difference in recall accuracy between related and unrelated word pairs (i.e., the relatedness effect) as a function of condition. In Panel B, red points represent the difference score in recall between related and unrelated word pairs for each participant, and the blue points represent group averages. Error bars represent 95% CI. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

difference between the control and no-JOL conditions (p = .670, BF_{10} = 0.266) for the recall of related pairs.

As reported in Table 3, recall of unrelated pairs was poorer in the JOL than in both the control (p = .038, BF_{10} = 1.633, even though the Bayesian evidence for this difference is relatively weak), and no-JOL conditions (p = .008, BF_{10} = 5.129). However, there was little difference between the control and no-JOL conditions (p = .348, BF_{10} = 0.360) for the recall of unrelated pairs.

Discussion

Experiment 5 demonstrated that the enhanced relatedness effect, positive reactivity for related pairs, and negative reactivity for unrelated pairs do not transfer to no-JOL pairs when JOL and no-JOL pairs are interleaved. These findings challenge the transfer prediction of the changed-goal hypothesis.

General Discussion

The current study conducted five experiments to test the changed-goal hypothesis of reactivity, proposed by Mitchum et al. (2016). Below we briefly summarize the main findings of Experiments 1–5 and discuss their theoretical implications.

Given that Mitchum et al. (2016) and Janes et al. (2018) obtained inconsistent findings about the enhanced relatedness effect of immediate JOLs in self-paced study conditions, Experiment 1 was conducted to further test the replicability of the effect. Consistent with the results reported by Janes et al. (2018) but inconsistent with those of Mitchum et al. (2016), Experiment 1 found no enhanced relatedness effect of immediate JOLs when the study procedure was self-paced. Furthermore, again consistent with Janes et al. (2018) but not with Mitchum et al. (2016), Experiment 1 observed that immediate JOLs did not change participants' study time allocation (as reflected by no difference in correlation between study time and cue-target relatedness strength), running counter to the changed-goal hypothesis.

Further findings challenging the changed-goal hypothesis come from Experiments 2–4, in which the study procedure was changed to experimenter-paced. The results showed that under these conditions only immediate JOLs, but not pre-study JOLs, enhanced the relatedness effect, regardless of whether the control group received material type prompts or not, and regardless of whether the prompts were implicit ("Related" and "Unrelated") or explicit ("Easy" or "Difficult"). Furthermore, these three experiments consistently showed that the relatedness effect was larger in the immediate JOL than in the pre-study JOL group. These findings were further confirmed by Bayesian model-averaged meta-analyses, which integrated results across Experiments 2–4 to increase statistical power. Specifically, the meta-analytic results showed that immediate JOLs substantially enhanced the relatedness effect (difference = 0.244, BF_{10} = 58.794), whereas pre-study JOLs did not (difference = 0.052, BF_{10} = 0.179). Furthermore, the relatedness effect was substantially larger in the immediate JOL than in the pre-study JOL groups (difference = 0.193, BF_{10} = 28.321). Overall, these findings jointly challenge the predictions of the changed-goal hypothesis that (1) pre-study JOLs should reactively enhance the relatedness effect, and (2) the effect of pre-study JOLs should be equal to or even larger than the effect of immediate JOLs.

Finally, Experiment 5 demonstrated minimal transfer of the enhanced relatedness effect to no-JOL pairs when JOL and no-JOL pairs were studied in a randomly interleaved order. This finding runs counter to the transfer prediction of the changed-goal hypothesis. Hence, even if goal change does contribute to the enhanced relatedness effect of immediate JOLs, it cannot be the main mechanism underlying the effect.

In summary, the findings from all five experiments and meta-analyses converge on the conclusion that goal change seems to play a minimal role in the enhanced relatedness effect of immediate JOLs.

Other explanations of JOL reactivity

As discussed in the Introduction, besides the changed-goal hypothesis, another viable account for the enhanced relatedness effect of immediate JOLs is the *dual-mechanism* hypothesis, proposed by Janes et al. (2018). This hypothesis assumes that the enhanced relatedness effect of immediate JOLs is caused by two separate mechanisms: (1) cue-strengthening (Soderstrom et al., 2015) and (2) dual-task costs (Mitchum et al., 2016). This hypothesis can readily account for all findings observed here (Janes et al., 2018).

First, it can explain the absence of the enhanced relatedness effect of immediate JOLs observed in Experiment 1, in which the study procedure was self-paced. Because participants had unlimited time to study each pair, cued recall of related pairs was almost at ceiling (M =.81) in the no-

JOL group, leaving little room for positive reactivity (i.e., cue-strengthening) to occur in the immediate JOL group. Additionally, dual-task costs are less likely to occur in self-paced study conditions because participants had unlimited time to perform the encoding and monitoring tasks (Finn & Roediger, 2013). Hence, the additional requirement of making JOLs produced little dual-task costs (i.e., negative reactivity) for unrelated pairs.

Second, the dual-mechanism hypothesis can explain the findings observed in Experiments 2–4, in which the study procedure was changed to experimenter-paced. In the immediate JOL group, participants needed to closely analyze the word pairs in order to identify "diagnostic" cues to guide JOL formation (Mueller & Dunlosky, 2017; Price & Harrison, 2017; Yang et al., 2021). The cues activated for making immediate JOLs in turn strengthened the cue-target relation for related pairs, leading to positive reactivity. Because pre-study JOLs were made before the presentation of each pair, no item-specific cues (e.g., relatedness strength, mediators between the cue and target) could be activated to inform pre-study JOLs (Price & Harrison, 2017). Hence, pre-study JOLs did not yield positive reactivity for related pairs.

Supporting evidence for this explanation comes from the difference in relative accuracy between immediate and pre-study JOLs (see Appendix A). Relative JOL accuracy (measured as the intra-individual correlation between JOLs and recall performance) is an index of the extent to which participants can metacognitively discriminate well-remembered items from less-well remembered ones (Rhodes & Tauber, 2011). Experiments 1–4 consistently found that relative accuracy of pre-study JOLs was substantially poorer than that of immediate JOLs (for connected findings, see Price & Harrison, 2017). This poorer relative accuracy of pre-study JOLs resulted from the fact that, before seeing the next word pair, participants could not identify item-specific cues (except for the material type, related/unrelated) to inform their pre-study JOLs.

According to the dual-task costs explanation of the dual-mechanism hypothesis (Mitchum et al., 2016), the additional requirement of making immediate JOLs might borrow limited study time from the primary learning task. For instance, when participants were studying the word pairs, they might concurrently search for informative cues in preparation to make immediate JOLs, leading to negative reactivity for unrelated pairs. By contrast, pre-study JOLs were made before the presentation of each pair, which means that the monitoring process could not borrow time from the subsequently presented word pair. Hence, only immediate JOLs, but not pre-study ones, produced negative reactivity for unrelated pairs.

Third, the dual-mechanism hypothesis provides an insightful explanation for the limited transfer finding observed in Experiment 5. Specifically, the dual-mechanism hypothesis assumes that positive reactivity for related pairs (caused by cue-strengthening) and negative reactivity for unrelated pairs (induced by dual-task costs) are task-specific phenomena: positive and negative reactivity effects are induced by the specific task requirement of making immediate JOLs, and when there is no such task requirement, these positive and negative effects disappear. Indeed, Experiment 5 only observed positive and negative reactivity for JOL and not for no-JOL pairs.

Overall, in contrast to the changed-goal hypothesis, the dual-mechanism hypothesis can provide reasonable explanations for all findings observed in Experiments 1–5.

Limitations and future research directions

As discussed above, the findings documented across Experiments 1–5 jointly challenge the changed-goal hypothesis and suggest that this mechanism plays a minimal role in the enhanced relatedness effect of immediate JOLs. In contrast, the dual-mechanism hypothesis can explain all findings observed here. However, it should be explicitly acknowledged that the current study was primarily conducted to test the changed-goal rather than the dual-mechanism hypothesis. It would be unreasonable to claim that the findings observed in the current study directly support the dual-mechanism hypothesis. Direct tests of this hypothesis are called for.

The current study focused particularly on the role of goal change in the enhanced relatedness effect, in which we employed a mixed list of related and unrelated word pairs as study stimuli and assessed memory performance using cued recall tests. Besides the reactivity effect on memory of a mixed list of related and unrelated word pairs (Janes et al., 2018; Maxwell & Huff, 2022; Mitchum et al., 2016; Soderstrom et al., 2015), recent studies found that making JOLs also reactively alters memory of word lists (Li, Zhao, et al., 2023; W. B. Zhao et al., 2023; W. L. Zhao et al., 2022), visual images (Shi et al., 2023), identical pairs (Halamish & Undorf, 2023), and inter-item relations (W. B. Zhao et al., 2022; 2023). Furthermore, the reactivity effect tends to be modulated by material type (Ariel et al., 2021; Schäfer & Undorf, 2023) and test format (Myers et al., 2020). It is possible that the mechanisms underlying reactivity effects on memory for different types of materials and measured by different test formats are distinct in nature. Future research needs to explore the mechanisms underlying reactivity effects on memory of other types of materials and the effects measured by other test formats.

A methodological limitation of the current study that should be highlighted is that the related and unrelated word pairs employed here were associated with different targets. Specifically, in the cued recall test, the targets that participants needed to recall for related and unrelated pairs were different. As in most previous studies (Halamish & Undorf, 2023; Janes et al., 2018; Li, Zhao, et al., 2023; Maxwell & Huff, 2022; Soderstrom et al., 2015; Tauber & Witherby, 2019; Witherby & Tauber, 2017b), we constructed distinct lists of related and unrelated word pairs, meaning that the relatedness manipulation is potentially confounded with any uncontrolled properties of the list items. As discussed in the Materials section of Experiment 1, we ensured that the targets were matched for log-transformed word frequency, concreteness, familiarity, and the number of strokes. Also, none of our main conclusions hinges on the absolute magnitude of the relatedness effect. Instead, our focus is on the extent to which the independent variables (e.g., pre-study JOLs *vs*. immediate JOLs) interact with the relatedness effect, and different groups studied exactly the same materials. Nevertheless, future research should follow Myers et al. (2020) in assigning both a related (e.g., *computer*) and an unrelated (e.g., *chair*) cue to the same target (e.g., *keyboard*), and counterbalance the assignment of a given target to the related and unrelated condition. In this way, researchers can better match targets between related and unrelated pairs and appropriately avoid potential confounding effects induced by differences in materials.

Concluding Remarks

Neither immediate nor pre-study JOLs reactively enhance the relatedness effect in self-paced study conditions. In experimenter-paced conditions, only immediate JOLs, but not pre-study JOLs, enhance the relatedness effect. The enhanced relatedness effect of immediate JOLs shows minimal transfer to no-JOL pairs when JOL and no-JOL pairs are interleaved. Goal change plays a small or possibly even no role in the enhanced relatedness effect of immediate JOLs.

CRediT authorship contribution statement

Baike Li: Writing – review & editing, Writing – original draft, Project administration, Methodology, Formal analysis, Data curation, Conceptualization. **David R. Shanks:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization. **Wenbo Zhao:** Visualization, Data curation. **Xiao Hu:** Software, Methodology, Data curation. **Liang Luo:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization. **Chunliang Yang:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

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

Acknowledgments

This research was supported by the Natural Science Foundation of China (32371116; 32000742; 32171045), the Fundamental Research Funds for the Central Universities (1233200008), Tang Scholar Foundation, the Research Program Funds of the Collaborative Innovation Center of Assessment for Basic Education Quality at Beijing Normal University (2021-01-132-BZK01), the United Kingdom Economic and Social Research Council (ES/S014616/1; ES/Y002482/1), and the China Postdoctoral Science Foundation (2023 M740300).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jml.2024.104506.

References

Arbuckle, T. Y., & Cuddy, L. L. (1969). Discrimination of item strength at time of presentation. *Journal of Experimental Psychology, 81*, 126–131. https://doi.org/10.1037/h0027455

Ariel, R., Karpicke, J. D., Witherby, A. E., & Tauber, S. (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

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

Berkhout, S. W., Haaf, J. M., Gronau, Q. F., Heck, D. W., & Wagenmakers, E.-J. (2023). A tutorial on Bayesian model-averaged meta-analysis in JASP. *Behavior Research Methods*. https://doi.org/10.3758/s13428-023-02093-6

Block, G. F. (2008). *Problem-solving and metacognitive judgments in the Mastermind deductive reasoning game.* (Ph.D.), University of California, Irvine, Ann Arbor. Retrieved from https://www.proquest.com/dissertations-theses/problem-solving-metacognitive-judgments/docview/304648258/se-2.

Bonder, T., & Gopher, D. (2019). The effect of confidence rating on a primary visual task. *Frontiers in Psychology, 10*. https://doi.org/10.3389/fpsyg.2019.02674

Castel, A. D. (2008). Metacognition and learning about primacy and recency effects in free recall: The utilization of intrinsic and extrinsic cues when making judgments of learning. *Memory & Cognition, 36*, 429–437. https://doi.org/10.3758/MC.36.2.429

Chang, M., & Brainerd, C. J. (2023). Changed-goal or cue-strengthening? Examining the reactivity of judgments of learning with the dual-retrieval model. *Metacognition and Learning, 18*, 183–217. https://doi.org/10.1007/s11409-022-09321-y

Davis, S. D., Chan, J. C., & Wilford, M. M. (2017). The dark side of interpolated testing: Frequent switching between retrieval and encoding impairs new learning. *Journal of Applied Research in Memory and Cognition, 6*, 434–441. https://doi.org/10.1016/j.jarmac.2017.07.002

Delaney, P. F., & Verkoeijen, P. P. J. L. (2009). Rehearsal strategies can enlarge or diminish the spacing effect: Pure versus mixed lists and encoding strategy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 1148–1161. https://doi.org/10.1037/a0016380

Doherty, J. M., Belletier, C., Rhodes, S., Jaroslawska, A., Barrouillet, P., Camos, V., & Logie, R. H. (2019). Dual-task costs in working memory: An adversarial collaboration. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 45*, 1529–1551. https://doi.org/10.1037/xlm0000668

Double, K. S., & Birney, D. P. (2017). Are you sure about that? Eliciting confidence ratings may influence performance on Raven's progressive matrices. *Thinking & Reasoning, 23*, 190–206. https://doi.org/10.1080/13546783.2017.1289121

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. (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

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

Finn, B., & Roediger, H. L., 3rd. (2013). Interfering effects of retrieval in learning new information. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*, 1665–1681. https://doi.org/10.1037/a0032377

Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., & Rees, G. (2010). Relating introspective accuracy to individual differences in brain structure. *Science, 329*, 1541–1543. https://doi.org/10.1126/science.1191883

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

Griffin, T. D., Wiley, J., & Thiede, K. W. (2008). Individual differences, rereading, and self-explanation: Concurrent processing and cue validity as constraints on metacomprehension accuracy. *Memory & Cognition, 36*, 93–103. https://doi.org/10.3758/MC.36.1.93

Gronau, Q. F., Heck, D. W., Berkhout, S. W., Haaf, J. M., & Wagenmakers, E.-J. (2021). A primer on Bayesian model-averaged meta-analysis. *Advances in Methods and Practices in Psychological Science, 4*, 25152459211031256. https://doi.org/10.1177/25152459211031256.

Halamish, V., & Undorf, M. (2023). Why do judgments of learning modify memory? Evidence from identical pairs and relatedness judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 49*, 547. https://doi.org/10.1037/xlm0001174

Hoijtink, H., Mulder, J., van Lissa, C., & Gu, X. (2019). A tutorial on testing hypotheses using the Bayes factor. *Psychological Methods, 24*, 539–556. https://doi.org/10.1037/met0000201

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

Hu, X., Yang, C., & Luo, L. (2022). Retrospective confidence rating about memory performance is affected by both retrieval fluency and non-decision time. *Metacognition and Learning, 17*, 651–681. https://doi.org/10.1007/s11409-022-09303-0

Hu, X., Yang, C., & Luo, L. (2023). Are the contributions of processing experience and prior beliefs to confidence ratings domain-general or domain-specific? *Journal of Experimental Psychology: General, 152*, 28–44. https://doi.org/10.1037/xge0001257

Hui, L., de Bruin, A. B. H., Donkers, J., & van Merriënboer, J. J. G. (2021). Does individual performance feedback increase the use of retrieval practice? *Educational Psychology Review, 33*, 1835–1857. https://doi.org/10.1007/s10648-021-09604-x

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, X., Li, P., Li, X., Zhang, Y., Cao, W., Cao, L., & Li, W. (2015). The effect of word frequency on judgments of learning: Contributions of beliefs and processing fluency. *Frontiers in Psychollogy, 6*, 1995. https://doi.org/10.3389/fpsyg.2015.01995

Karaca, M., Kurpad, N., Wilford, M. M., & Davis, S. D. (2020). Too much of a good thing: Frequent retrieval can impair immediate new learning. *Memory, 28*, 1181–1190. https://doi.org/10.1080/09658211.2020.1826526

Keysers, C., Gazzola, V., & Wagenmakers, E.-J. (2020). Using Bayes factor hypothesis testing in neuroscience to establish evidence of absence. *Nature Neuroscience, 23*, 788–799. https://doi.org/10.1038/s41593-020-0660-4

King, J. F., Zechmeister, E. B., & Shaughnessy, J. J. (1980). Judgments of knowing: The influence of retrieval practice. *The American Journal of Psychology, 93*, 329–343. https://doi.org/10.2307/1422236

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

Lei, W., Chen, J., Yang, C., Guo, Y., Feng, P., Feng, T., & Li, H. (2020). Metacognition-related regions modulate the reactivity effect of confidence ratings on perceptual decision-making. *Neuropsychologia, 144*, Article 107502. https://doi.org/10.1016/j.neuropsychologia.2020.107502

Li, B., Hu, X., Shanks, D. R., Su, N., Zhao, W. B., Meng, L., … Yang, C. (2023). Confidence ratings increase response thresholds in decision making. *Psychonomic Bulletin & Review, 1–10*. https://doi.org/10.3758/s13423-023-02380-5

Li, B., Zhao, W. B., Shi, A., Zhong, Y., Hu, X., Liu, M., & Yang, C. (2023). Does the reactivity effect of judgments of learning transfer to learning of new information? *Memory, 31*, 918–930. https://doi.org/10.1080/09658211.2023.2208792

Li, B., Zhao, W. B., Zheng, J., Hu, X., Su, N., Fan, T., & Luo, L. (2022). Soliciting judgments of forgetting reactively enhances memory as well as making judgments of learning: Empirical and meta-analytic tests. *Memory & Cognition, 50*, 1061–1077. https://doi.org/10.3758/s13421-021-01258-y

Maxwell, N. P., & Huff, M. J. (2022). Reactivity from judgments of learning is not only due to memory forecasting: Evidence from associative memory and frequency judgments. *Metacognition and Learning, 17*, 589–625. https://doi.org/10.1007/s11409-022-09301-2

Metcalfe, J., & Kornell, N. (2003). The dynamics of learning and allocation of study time to a region of proximal learning. *Journal of Experimental Psychology: General, 132*, 530–542. https://doi.org/10.1037/0096-3445.132.4.530

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

Mueller, M. L., & Dunlosky, J. (2017). How beliefs can impact judgments of learning: Evaluating analytic processing theory with beliefs about fluency. *Journal of Memory and Language, 93*, 245–258. https://doi.org/10.1016/j.jml.2016.10.008

Mueller, M. L., Dunlosky, J., Tauber, S. K., & Rhodes, M. G. (2014). The font-size effect on judgments of learning: Does it exemplify fluency effects or reflect people's beliefs about memory? *Journal of Memory and Language, 70*, 1–12. https://doi.org/10.1016/j.jml.2013.09.007

Mulder, J., & Wagenmakers, E.-J. (2016). Editors' introduction to the special issue "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. (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. O., & Leonesio, R. J. (1988). Allocation of self-paced study time and the "labor-in-vain effect". *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 676–686. https://doi.org/10.1037//0278-7393.14.4.676

Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment: Englewood Cliffs*. NJ, USA: Prentice-Hall.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*, 231–259. https://doi.org/10.1037/0033-295X.84.3.231

Price, J., & Harrison, A. (2017). Examining what prestudy and immediate judgments of learning reveal about the bases of metamemory judgments. *Journal of Memory and Language, 94*, 177–194. https://doi.org/10.1016/j.jml.2016.12.003

Rhodes, M. G., & Tauber, S. K. (2011). The influence of delaying judgments of learning on metacognitive accuracy: A meta-analytic review. *Psychological Bulletin, 137*, 131–148. https://doi.org/10.1037/a0021705

Rivers, M. L., Janes, J. L., & Dunlosky, J. (2021). Investigating memory reactivity with a within-participant manipulation of judgments of learning: Support for the cue-strengthening hypothesis. *Memory, 29*, 1342–1353. https://doi.org/10.1080/09658211.2021.1985143

Rundus, D. (1971). Analysis of rehearsal processes in free recall. *Journal of Experimental Psychology, 89*, 63–77. https://doi.org/10.1037/h0031185

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

Schäfer, F., & Undorf, M. (2023). On the educational relevance of immediate judgment of learning reactivity: No effects of predicting one's memory for general knowledge facts. *Journal of Applied Research in Memory and Cognition, Advance online publication.*. https://doi.org/10.1037/mac0000113

Senkova, O., & Otani, H. (2021). Making judgments of learning enhances memory by inducing item-specific processing. *Memory & Cognition, 49*, 955–967. https://doi.org/10.3758/s13421-020-01133-2

Shi, A., Xu, C., Zhao, W. B., Shanks, D. R., Hu, X., Luo, L., & Yang, C. (2023). Judgments of learning reactively facilitate visual memory by enhancing learning engagement. *Psychonomic Bulletin & Review, 30*, 676–687. https://doi.org/10.3758/s13423-022-02174-1

Shynkaruk, J. M., & Thompson, V. A. (2006). Confidence and accuracy in deductive reasoning. *Memory & Cognition, 34*, 619–632. https://doi.org/10.3758/BF03193584

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

Spellman, B. A., & Bjork, R. A. (1992). When predictions create reality: Judgments of learning may alter what they are intended to assess. *Psychological Science, 3*, 315–317. https://doi.org/10.1111/j.1467-9280.1992.tb00680.x

Sun, Y., Shi, A., Zhao, W. B., Yang, Y., Li, B., Hu, X., … Luo, L. (2022). Long-lasting effects of an instructional intervention on interleaving preference in inductive learning and transfer. *Educational Psychology Review, 34*, 1679–1707. https://doi.org/10.1007/s10648-022-09666-5

Tauber, S. K., & Witherby, A. E. (2019). Do judgments of learning modify older adults' actual learning? *Psychology & 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

Tendeiro, J. N., & Kiers, H. A. L. (2019). A review of issues about null hypothesis Bayesian testing. *Psychological Methods, 24*, 774–795. https://doi.org/10.1037/met0000221

Thiede, K. W., Anderson, M. C., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology, 95*, 66–73. https://doi.org/10.1037/0022-0663.95.1.66

Winkielman, P., & Schooler, J. W. (2011). Splitting consciousness: Unconscious, conscious, and metaconscious processes in social cognition. *European Review of Social Psychology, 22*, 1–35. https://doi.org/10.1080/10463283.2011.576580

Witherby, A. E., & Tauber, S. K. (2017a). The concreteness effect on judgments of learning: Evaluating the contributions of fluency and beliefs. *Memory & Cognition, 45*, 639–650. https://doi.org/10.3758/s13421-016-0681-0

Witherby, A. E., & Tauber, S. K. (2017b). The influence of judgments of learning on long-term learning and short-term performance. *Journal of Applied Research in Memory and Cognition, 6*, 496–503. https://doi.org/10.1016/j.jarmac.2017.08.004

Yang, C., Huang, T. S. T., & Shanks, D. R. (2018). Perceptual fluency affects judgments of learning: The font size effect. *Journal of Memory and Language, 99*, 99–110. https://doi.org/10.1016/j.jml.2017.11.005

Yang, C., Potts, R., & Shanks, D. R. (2017). Metacognitive unawareness of the errorful generation benefit and its effects on self-regulated learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*, 1073–1092. https://doi.org/10.1037/xlm0000363

Yang, C., Potts, R., & Shanks, D. R. (2018). Enhancing learning and retrieval of new information: A review of the forward testing effect. *npj Science of Learning, 3*, 8. https://doi.org/10.1038/s41539-018-0024-y

Yang, C., Yu, R., Hu, X., Luo, L., Huang, T., & Shanks, D. R. (2021). How to assess the contributions of processing fluency and beliefs to the formation of judgments of learning: Methods and pitfalls. *Metacognition and Learning, 16*, 319–343. https://doi.org/10.1007/s11409-020-09254-4

Zechmeister, E. B., & Shaughnessy, J. J. (1980). When you know that you know and when you think that you know but you don't. *Bulletin of the Psychonomic Society, 15*, 41–44. https://doi.org/10.3758/BF03329756

Zhao, W. B., Li, J., Shanks, D. R., Li, B., Hu, X., Yang, C., & Luo, L. (2022). Metamemory judgments have dissociable reactivity effects on item and inter-item relational memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 49*, 557–574. https://doi.org/10.1037/xlm0001160

Zhao, W. B., Yin, Y., Hu, X., Shanks, D. R., Yang, C., & Luo, L. (2023). Memory for inter-item relations is reactively disrupted by metamemory judgments. *Metacognition and Learning, 18*, 549–566. https://doi.org/10.1007/s11409-023-09340-3

Zhao, W. L., Li, B., Shanks, D. R., Zhao, W. B., Zheng, J., Hu, X., & Yang, C. (2022). When judging what you know changes what you really know: Soliciting metamemory judgments reactively enhances children's learning. *Child Development, 93*, 405–417. https://doi.org/10.1111/cdev.13689