Metamemory Judgments Have Dissociable Reactivity Effects on Item and Inter-Item Relational Memory

Wenbo, Zhao[1], David R. Shanks[2], Baike Li[3], Xiao Hu[3,4], Chunliang Yang[3,4], Liang Luo[1,3]

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

[2] Division of Psychology and Language Sciences, University College London, London, the UK.

[3] Institute of Developmental Psychology, Faculty of Psychology, Beijing Normal University, Beijing, China.

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

Author Note

The data contained in this project and the experimental instructions for each experiment are publicly available at Open Science Framework (https://osf.io/zg3vr/). Correspondence concerning this article should be addressed to Liang Luo (luoliang@bnu.edu.cn) or Chunliang Yang (chunliang.yang@bnu.edu.cn), 19 Xinjiekou Wai Street, Beijing 100875, China.

Acknowledgments

This research was supported by the Natural Science Foundation of China (32000742; 32171045), the Fundamental Research Funds for the Central Universities (2019NTSS28), and the United Kingdom Economic and Social Research Council (ES/S014616/1).

Abstract

Making metamemory judgments reactively changes item memory itself. Here we report the first investigation of reactive influences of making judgments of learning (JOLs) on inter-item relational memory – specifically, temporal (serial) order memory. Experiment 1 found that making JOLs impaired order reconstruction. Experiment 2 observed minimal reactivity on free recall and negative reactivity on temporal clustering. Experiment 3 demonstrated a positive reactivity effect on recognition memory, and Experiment 4 detected dissociable effects of making JOLs on order reconstruction (negative) and forced-choice recognition (positive) by using the same participants and stimuli. Finally, a meta-analysis was conducted to explore reactivity effects on word list learning and to investigate whether test format moderates these effects. The results show a negative reactivity effect on inter-item relational memory (order reconstruction), a modest positive effect on free recall, and a medium-to-large positive effect on recognition. Overall, these findings imply that even though making metacognitive judgments facilitates item-specific processing, it disrupts relational processing, supporting the *item-order account* of the reactivity effect on word list learning.

Keywords: Judgments of learning; reactivity effect; test format; meta-analysis; item-order account.

Accurately monitoring on-going learning and memory status is critical for being a successful learner. Individuals frequently regulate their study activities (e.g., restudy decisions) according to their metamemory monitoring, which in turn influences learning efficiency (Finn, 2008; Thiede, Anderson, & Therriault, 2003; C. Yang, Potts, & Shanks, 2017). Over the last half-century, hundreds of studies have been conducted to evaluate the extent to which people are able to accurately monitor their memory status, and to investigate the mechanisms underlying the monitoring process (for reviews, see Dunlosky & Tauber, 2016; Rhodes & Tauber, 2011; C. Yang, Yu, et al., 2021). In these studies, participants are prompted to make a judgment of learning (JOL; i.e., a metacognitive estimate of the likelihood of remembering an item on a future memory occasion) during or after they study each item.

About 30 years ago, Spellman and Bjork (1992) speculated that making item-by-item JOLs might fail to measure what they intend to evaluate because the act of providing JOLs might reactively alter memory itself. Indeed, an emerging body of recent studies consistently demonstrated that making JOLs (at least in some situations) reactively changes memory itself, a phenomenon termed the *reactivity effect* on memory (for a review, see Double, Birney, & Walker, 2018).

Below we briefly summarize previous findings of the reactivity effect, and discuss the rationale of the current study (why it is important to explore the reactivity effect on inter-item relational memory). Next, we explain that all existing theories have difficulty in explaining the effect on word list learning, and then introduce a new theory to explain this effect – the *item-order account* (McDaniel & Bugg, 2008), which predicts that making JOLs enhances

item memory but disrupts inter-item relational memory. Finally, we provide an overview of the current study.

Reactivity effects

In a recent meta-analysis, Double et al. (2018) found that making item-by-item JOLs significantly enhances memory for at least some types of materials.[1] For instance, previous studies found that eliciting JOLs facilitates retention of related word pairs and word lists (Soderstrom, Clark, Halamish, & Bjork, 2015; Zhao et al., 2021), while it has minimal influence on memory for unrelated word pairs (Double et al., 2018) or text passages (Ariel, Karpicke, Witherby, & Tauber, 2021). Other studies found that the reactivity effect occurs in young children (Zhao et al., 2021) as well as young adults, although making JOLs fails to facilitate older adults' memory (Tauber & Witherby, 2019). Myers, Rhodes, and Hausman (2020) recently found that the reactivity effect on learning of word pairs is moderated by test format, with a positive reactivity effect emerging in a recognition test but absent in a free recall test.

Overall, although a set of studies has begun to investigate the reactivity effect and its moderating factors, all have focused on the reactive effect of making JOLs on item memory (e.g., recall or recognition of specific items). Thus far, no studies have explored whether

[1] Several studies have explored the effect of delayed JOLs on memory for word pairs (e.g., Tekin & Roediger, 2021). In these studies, after initial study of all word pairs (e.g., *apple - banana*), the cue words (e.g., *apple - __*) were presented one-by-one, and participants were asked to make delayed JOLs to predict the possibility that they could successfully recall the targets when prompted with the cues in a later cued recall test. The enhancing effect of delayed JOLs on memory may result from the fact that participants tend to retrieve the targets when making delayed JOLs (Tekin & Roediger, 2021). The current study instead focused on the reactivity effect of immediate JOLs (i.e., JOLs made while or immediately after studying). The effect of delayed JOLs is unrelated to the current study, and we hence do not discuss it further.

making JOLs has reactive influences on inter-item relational memory (i.e., memory for inter-item relations).[2]

An important form of inter-item relational memory is memory for temporal information (i.e., retention of information about the temporal order of studied items or encountered events), which is routinely encoded into episodic memory. Tulving (1972) suggested that episodic memory is composed of both its contents and its underlying organization (e.g., temporal organization). Previous studies found that list items encoded nearby in time are likely to be recalled consecutively in the order they were studied, reflecting a *temporal contiguity effect* (Kahana, 1996). Cognitive models of free recall, such as the Context Maintenance and Retrieval model (Polyn, Norman, & Kahana, 2009), assume that study items are associated with temporal contexts, such that items studied nearby in time share stronger temporal similarity (Farrell, 2012). Furthermore, previous studies established that, in free recall tests, temporal order memory guides output sequence, promotes output organization and facilitates recall performance (Farrell, 2012). Indeed, numerous studies have found that superior temporal clustering is associated with better free recall (Solway, Murdock, & Kahana, 2012; C. Yang, Zhao, et al., 2021).

Besides the important role of temporal order memory in guiding free recall observed in laboratory studies, temporal order memory also plays an important role in daily life, especially for autobiographical memory. For instance, when individuals are prompted to retrieve autobiographical events, their recall is typically organized in the order the events

[2] A few studies have used word pairs as stimuli to explore the reactivity effect on intra-item relational memory (e.g., memory of cue-target relations) (e.g., Soderstrom et al., 2015). It should be highlighted that intra- and inter-item relational memory are distinct in nature. For instance, Peterson and Mulligan (2013) provided a clear dissociation of these two forms of relational memory by showing that practice testing enhances intra-item relational memory but concurrently disrupts inter-item relational memory.

were encountered (Diamond & Levine, 2020), suggesting that autobiographical memory represents experience in a temporally structured way (Jeunehomme, Folville, Stawarczyk, Van der Linden, & D'Argembeau, 2018). Furthermore, correlations between recall of autobiographical events and their temporal clustering are quite strong, with *r*s ranging from .47 to .72 (White, 2002).

Overall, inter-item relational memory (e.g., memory of temporal order) is a critical component of human memory, and it plays an important role in daily life. However, all previous studies focused on the reactivity effect on item memory, and the question of whether making JOLs reactively alters inter-item relational memory has not yet been investigated.[3] The first aim of the current study is to fill this gap. Specifically, the current study employs a word list learning task to explore the reactivity effect on inter-item relational memory.

Reactivity effect on word list learning and the item-order account

A set of previous studies has explored the reactivity effect on word list learning, and the documented findings suggest that the effect tends to be moderated by test format. For instance, previous studies showed that making JOLs substantially enhances recognition of word lists (Li et al., 2021; H. Yang et al., 2015; Zhao et al., 2021), whereas it has a minimal reactive influence on free recall (Stevens, 2019; Tauber & Rhodes, 2012).

Several theories have been proposed to account for the reactivity effect, such as the *cue-strengthening theory* (Soderstrom et al., 2015), the *changed-goal theory* (Mitchum, Kelley, & Fox, 2016), and the *positive reactivity theory* (Mitchum et al., 2016). All of them have

[3] Although Senkova and Otani (2021) used related word lists as study stimuli to explore the reactivity effect on free recall of the words, they did not measure the difference in semantic clustering between the JOL and no-JOL conditions, leaving it unclear whether making JOLs affects inter-item relational memory.

difficulty in explaining the reactivity effect on word list learning (see Appendix A for detailed discussion). Here we propose that the *item-order account* might provide a viable explanation for this effect (McDaniel & Bugg, 2008).

The item-order account was originally proposed by McDaniel and Bugg (2008), who hypothesized that information about a list of study items (e.g., a list of words) contains two components: (1) information about individual items (i.e., item-specific information) and (2) information about relations among list items (i.e., inter-item relational information). Item-specific information includes item characteristics (such as concreteness) that differentiate a given item from others in the list and increase its distinctiveness. Inter-item relational information, in contrast, refers to relations among list items studied across different study trials, such as whether Item A or B was studied first. Both item-specific and inter-item relational processing contribute importantly to free recall of list items (Hunt & McDaniel, 1993; Mulligan & Lozito, 2007; Peterson & Mulligan, 2013). For instance, relational processing provides structure to support the search for targets and to guide output order, while item-specific processing enables individuals to retrieve list targets and protects target recall against intrusions and interference.

Importantly, the item-order account hypothesizes that relational processing and item-specific processing can be functionally disassociated (McDaniel & Bugg, 2008). When a given encoding strategy enhances processing of item-specific information, inter-item relational processing may be correspondingly disrupted because more cognitive resources are allocated to item-specific processing and fewer are left for inter-item relational processing. Many studies have provided evidence supporting the item-order account. For instance,

previous studies found that words spoken aloud are remembered better than those read silently, a phenomenon termed the *production effect* (for a review, see MacLeod & Bodner, 2017), but concurrently speaking aloud hinders inter-item relational processing as reflected by poorer order reconstruction performance for spoken than for silent words (Jonker, Levene, & MacLeod, 2014).

Similarly, although previous studies found that testing, by comparison with passive restudying, significantly enhances retention of studied items – a phenomenon known as the *testing effect* (for reviews, see Roediger & Karpicke, 2006; Rowland, 2014; C. Yang, Luo, Vadillo, Yu, & Shanks, 2021), testing significantly disrupts retention of temporal order information (Karpicke & Zaromb, 2010). Even though active generation, by comparison with passive reading, facilitates recall of studied items – a phenomenon termed the *generation effect* (for reviews, see Bertsch, Pesta, Wiscott, & McDaniel, 2007; McCurdy, Viechtbauer, Sklenar, Frankenstein, & Leshikar, 2020), the active generation process has a negative influence on temporal order memory (Karpicke & Zaromb, 2010). As a final example, although drawing study items enhances memory overall – a phenomenon referred to as *the drawing effect* (for a review, see Fernandes, Wammes, & Meade, 2018), drawing impairs order reconstruction (Jonker, Wammes, & MacLeod, 2019). Many other studies have documented evidence supporting the item-order account (e.g., Forrin & MacLeod, 2016; Jonker & MacLeod, 2015). Overall, there are ample findings, deriving from different memory phenomena, supporting the main proposals of the item-order account.

It is reasonable to assume that, similar to speaking aloud, practice testing, generation, and drawing, making JOLs may reactively enhance item-specific processing and concurrently

disrupt inter-item relational processing. For instance, participants have to closely encode and analyze the current item in order to make an appropriate memory prediction for it, which improves item-specific processing and produces superior item memory (for related discussion, see Senkova & Otani, 2021). Concurrently, greater processing of item-specific information may detract from encoding of inter-item relational information (e.g., information about temporal order among list items), leading to poorer inter-item relational memory.

Going beyond the cue-strengthening, changed-goal, and positive reactivity theories, the item-order account can readily account for the reactivity effect on word list learning and the moderating role of test format. It is well-known that free recall performance is dependent on both item and inter-item relational memory (McDaniel & Bugg, 2008; Mulligan & Peterson, 2015). Because making JOLs concurrently improves item memory and impairs inter-item relational memory, these two effects may cancel each other out, leading to little reactivity on free recall of word lists. By contrast, in recognition tests, test items are typically presented in random order or the presentation order is controlled by the experimenter (rather than by participants themselves), hence recognition performance relies less on inter-item relational memory and more on item memory (Engelkamp, Biegelmann, & McDaniel, 1998; Guynn et al., 2014; Hunt & Einstein, 1981; Mickes, Wixted, Shapiro, & Scarff, 2009). Therefore, the reactivity effect on recognition is generally positive and larger than the effect on free recall.

Overview of the current study

As discussed above, all previous studies focused on the reactive influences of making JOLs on recall or recognition of specific items, and it remains unknown whether making JOLs reactively changes memory of inter-item relations. In addition, all existing theories (i.e.,

cue-strengthening, changed-goal, and positive reactivity) have difficulty in accounting for the reactivity effect on word list learning, and it is unknown whether the item-order account is a viable explanation of this effect. Accordingly, the current study was designed to explore (1) whether making JOLs reactively alters inter-item relational (temporal order) memory, and (2) whether the item-order account is a viable explanation of the reactivity effect on word list learning.

Experiment 1 instructed participants to study 16 lists of words. For eight JOL lists, participants were required to make concurrent JOLs while studying each word. By contrast, for the other eight no-JOL lists, they did not make such judgments. Immediately after they studied each list and completed a brief distractor task, they took an order reconstruction test, in which they ordered the just-studied words in the sequence they were studied. The reactivity effect on temporal order memory was quantified as the difference in order reconstruction accuracy between the JOL and no-JOL conditions. To foreshadow, Experiment 1 observed a significantly negative reactivity effect on order reconstruction. Experiment 2 employed a different test format – free recall – to test reactivity on temporal order memory by comparing temporal clustering scores (TCSs; see below for details) between the JOL and no-JOL conditions.

Experiment 3 employed a forced-choice recognition test to further investigate whether making item-by-item JOLs reactively enhances item memory. Relative to free recall, recognition places more demands on item memory and fewer on temporal order memory (Engelkamp et al., 1998; Guynn et al., 2014; Hunt & Einstein, 1981; Mickes et al., 2009). To foreshadow, Experiment 3 demonstrated that making concurrent JOLs enhanced recognition

performance, suggesting a positive reactivity effect on item memory. Experiment 4 demonstrated dissociable effects of making JOLs on order reconstruction (negative) and forced-choice recognition (positive) by using the same participants and stimuli. As an aid to readers, Table 1 summarizes the main results and conclusions from each experiment.

Finally, we conducted a meta-analysis to integrate results across studies to explore whether test format (i.e., order reconstruction, free recall, and recognition) moderates the reactivity effect on word list learning. Order reconstruction is directly related to temporal order memory, free recall is dependent on both item memory and temporal order memory, and recognition predominantly relies on item memory. Therefore, according to the item-order account, we expected to observe a negative reactivity effect in order reconstruction tests, a null or modest reactivity effect in free recall tests, and a positive effect in recognition tests.

Note that the four experiments and the meta-analysis reported in the current study were not pre-registered. We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the experiments.

Experiment 1

Experiment 1 employed an order reconstruction test to (1) measure the reactivity effect on temporal order memory, and (2) assess the item-order account's main proposal that making JOLs impairs processing of temporal information. The reactivity effect on item memory was not assessed in Experiment 1. Instead, it was investigated in Experiments 3 and 4 and the meta-analysis (see below for details).

Method

Participants

A pilot study was conducted to estimate the magnitude of the reactivity effect on temporal order memory (i.e., order reconstruction performance). This pilot study used the same procedure and stimuli as the formal experiment, with 10 participants recruited from the same participant pool. The results showed a medium-to-large (Cohen's d = -0.650) reactivity effect on order reconstruction. A power analysis, conducted via G*Power (Faul, Erdfelder, Lang, & Buchner, 2007), showed that approximately 21 participants were required to observe a significant (2-tailed, α = .050) reactivity effect at 0.80 power. To enhance statistical power, we decided to increase the sample size to 25.

Finally, 27 participants (15 female), with a mean age of 20.82 (SD = 2.19) years, were recruited from Beijing Normal University (BNU). They reported normal or corrected-to-normal vision, were tested individually in a sound-proofed cubicle, and received 50 RMB as compensation. The Ethics Committee at the Collaborative Innovation Center of Assessment for Basic Education Quality, BNU, approved Experiments 1-4.

Materials

Two hundred and sixteen two-character Chinese words were selected from the Chinese word database developed by Cai and Byrsbert (2010). Word frequency was between 2.98 and 51.33 per million. Twenty-four words were used for practice, and the remaining 192 words were used in the main experiment. To avoid any item-selection effects, for each participant, the computer randomly divided the 192 words into 16 lists, with 12 words in each list, and the 16 lists were randomly allocated into four blocks, with 4 lists in each block. Then, two blocks were randomly assigned to the JOL condition and the other two to the no-JOL condition. The presentation sequence of words in each list, the list sequence in each block, and the block

sequence were randomly decided by the computer for each participant. All stimuli were presented via the Matlab *Psychtoolbox* (Kleiner, Brainard, & Pelli, 2007) on a CRT display.

Design and procedure

The experiment involved a within-subjects design (study method: JOL *vs.* no-JOL). Participants were informed that they would study four blocks of words, with four lists in each block. After studying each list, they would complete a memory test. In two blocks, they would need to make a memory prediction while studying each word, but they did not need to make such predictions in the other two blocks. They were instructed to remember all words equally well regardless of whether they had to make memory predictions or not, because all words would be tested in a later order reconstruction test.

Before the experiment, participants practiced a JOL and a no-JOL list to familiarize themselves with the task. Following this practice, participants were asked if they understood the task requirements. If not, the experimenter re-explained the task, and they re-took the practice task. This cycle repeated until each participant fully understood the task requirements. Then, the main experiment began. Participants made concurrent JOLs in the JOL blocks but not in the no-JOL blocks. Before studying each block, the computer informed participants whether or not they would need to make memory predictions in the forthcoming block.

For lists in the no-JOL blocks, the 12 words were presented one-by-one, for 5 s each, in random order. After the presentation of each word, a cross sign appeared at the center of the screen for 0.5 s to mark the interstimulus interval. After studying each list, participants engaged in a 30 s distractor task in which the computer randomly generated a three-digit

number (e.g., *589*) and presented it at the center of the screen. Participants subtracted 3 from that digit in succession and wrote down the results on paper. There was no speed requirement for this calculation, and participants continued doing it until 30 s had elapsed.

Following the distractor task, an order reconstruction test was implemented. Specifically, the computer presented all 12 words on screen in a new random order. Participants were provided with a blank sheet. There were 12 lines on the sheet, with a digit (1-12) in front of each line. Participants were instructed to write the 12 words on the sheet in the order they were studied. For instance, if the first word was *"栏杆"* (*handrail*), they needed to write this word above the first line. There was no time pressure and no feedback in the order reconstruction task. Participants pressed a "Done" button to finish recall and trigger the study phase of the next list. This cycle repeated until all four lists in a block were studied and tested.

The procedure in the JOL blocks was the same as in the no-JOL blocks, but with one exception. Specifically, while participants studied each word, a 0-100 slider was presented below the word. Participants made a JOL during the 5 s time-window to predict the likelihood that they would remember it on a later memory test (0 = *Sure I will not remember it*; 100 = *Sure I will remember it*). If they did not make a JOL during this interval, a message box appeared to remind them to carefully make JOLs for the following words, and they clicked the mouse to trigger the next trial. If they successfully made a JOL, the word remained on screen for the remaining duration of the 5 s to ensure the total exposure duration of each word was the same between the JOL and no-JOL conditions. The whole experiment lasted about 50 minutes.

In summary, the only difference between the JOL and no-JOL conditions was that participants made concurrent JOLs while studying each word in the JOL condition.

Results

In Experiments 1-4, test performance results (i.e., the reactivity effect) were the major research interest, and hence are reported in the main text. Item-by-item JOLs (i.e., JOLs made for each word in the learning task) and their accuracy were not of substantive research interest, and therefore are reported in Appendix C.

For each of Experiments 1-4, we conducted a mixed analysis of variance (ANOVA), with block order as the between-subjects variable and study method as the within-subjects variable. The results showed no main effect of block order (ps $\geq$ .13) and no interaction between block order and study method (ps $\geq$ .44). Hence, below we do not discuss block order further.

Reactivity effect on order reconstruction

Following precedents (e.g., Jonker et al., 2014; Jonker & MacLeod, 2015; Mulligan & Lozito, 2007; Nairne, Riegler, & Serra, 1991), we used the antecedent scoring criterion to quantify order reconstruction performance. Specifically, items were only considered correct if they were exactly ordered in their original learning position. For instance, if "食谱" (*recipe*) was studied as Item 3, it would only be scored as correct if it was placed on the third line in the test sheet. The proportions of items correctly ordered were separately computed for the JOL and the no-JOL words for each participant.

Order reconstruction performance was significantly better for no-JOL (M = .88, SD = .17) than for JOL words (M = .79, SD = .16), difference = .095 [.041, .148], $t(26) = 3.66$, p

= .001, Cohen's d = 0.70 (see Figure 1A), reflecting a negative reactivity effect on order reconstruction.[4] As shown in Figure 1B, 22 participants showed a negative reactivity effect, 4 showed the converse pattern, and there was one tie.[5]

Discussion

Overall, the above results demonstrate a negative reactivity effect on order reconstruction, implying that making concurrent JOLs hinders processing of temporal information. These findings are consistent with the main proposal of the item-order account.

Experiment 2

Experiment 1 demonstrated a negative reactivity effect of making concurrent JOLs on order reconstruction, supporting the main proposal of the item-order account that making metacognitive judgments impairs processing of temporal information. Experiment 2 was conducted to replicate this negative reactivity effect on temporal order memory but employing a different test format – free recall. According to the item-order account, we predict that making JOLs will impair temporal clustering in a free recall test.

It is worth noting that previous findings of the reactivity effect on free recall performance (i.e., proportion of words recalled) have been somewhat inconsistent, with some studies showing a positive reactivity effect on free recall (e.g., Zechmeister & Shaughnessy,

[4] As shown in Figure 1B, there was an outlier. After excluding this outlier, order reconstruction performance remained significantly better for no-JOL (M = .88, SD = .18) than for JOL words (M = .80, SD = .15), difference = .077 [.037, .116], $t(25) = 3.96$, $p < .001$, Cohen's d = 0.78.

[5] As supplemental results, we also calculated absolute difference scores of serial positions. For instance, if a given word was studied at Position 7 and the participant assigned this word to Position 5 in the order reconstruction test, the absolute difference score for this word would be calculated as 2. The results for absolute distance scores showed the exact same pattern, with larger distance scores in the JOL (M = 0.52, SD = 0.53) than in the no-JOL condition (M = 0.26, SD = 0.45), difference = 0.265 [0.100, 0.430], $t(26) = 3.31$, p = .003, Cohen's d = 0.64, re-confirming that making JOLs impaired temporal order memory. Furthermore, for each word list, a Spearman rank correlation was conducted between study positions and the ordered positions in the test. For each participant, the correlation coefficients were averaged in the JOL and no-JOL conditions, respectively. The results also showed that the correlation was weaker in the JOL (M of r_s = .91, SD = .11) than in the no-JOL (M of r_s = .96, SD = .08) condition, difference = -.049 [-.084, -.015], $t(26) = -2.94$, p = .007, Cohen's d = -0.57.

1980) and others showing no reactivity (e.g., Tauber & Rhodes, 2012). Hence, the other goal of Experiment 2 was to further test the reactivity effect on free recall. According to the item-order account, we expect minimal reactivity on free recall, because making JOLs concurrently enhances item-specific processing and disrupts temporal clustering.

Method

Participants

Given that the primary aim of Experiment 2 was to explore the reactivity effect on inter-item relational (temporal order) memory, the sample size was set to 25, the same as in Experiment 1. Finally, 27 students (16 female), with a mean age of 21.04 (*SD* = 2.01) years, were recruited from BNU. They reported normal or corrected-to-normal vision, were tested individually in a sound-proofed cubicle, and received 50 RMB as compensation.

Materials, design, and procedure

The stimuli, design, and procedure were the same as in Experiment 1, except that the order-reconstruction tests were replaced by free recall tests. Specifically, participants were instructed to study the word lists in preparation for a free recall test. After studying each list and completing the distractor task, participants recalled as many words as they could from the just-studied list. They could recall the words in any order they liked, and they wrote their answers on a blank sheet. There was no time pressure and no feedback in the free recall tests. They clicked a “Done” button to finish recall and trigger the next list. The whole experiment lasted about 50 minutes.

Results

Reactivity effect on free recall

There was no statistically detectable difference in free recall between the JOL (M = .73, SD = .11) and no-JOL blocks (M = .72, SD = .15), difference = .012 [-.038, .062], $t(26)$ = .49, p = .63, Cohen's d = 0.09 (see Figure 2A), suggesting minimal reactivity on the number of items freely recalled. As shown in Figure 2B, 11 participants showed a negative reactivity effect, 15 showed the converse pattern, and there was one tie.

Reactivity effect on temporal clustering

Next, we examined the reactivity effect on temporal clustering. Temporal clustering was quantified as TCSs using the method developed by Polyn et al. (2009). TCSs index the degree to which items studied in neighboring serial positions in a list tend to be reported together during free recall (Lohnas, Polyn, & Kahana, 2011). TCSs range from 0 to 1, where a value of 0.5 indicates random clustering and a value of 1 indicates that the participant always chose the closest temporal associate. For example, if a participant recalled an item from Position 2 of the study list, the subsequent recall could be from a nearby position, such as Position 1 or 3, which would indicate better retention of temporal information, or it could be from a more distant position, such as Position 11 or 12, which would indicate poorer retention of temporal information.

To calculate TCSs, we determined the absolute temporal distances (measured by serial position from 1-12) between the positions of the just-recalled word and each of the not-yet-recalled ones. TCS between the just-recalled word and the subsequently recalled one was quantified as the proportion of all other possible absolute temporal distances that were greater than the observed distance. Put differently, TCS between the just-recalled and subsequently recalled word was computed as the proportion of all other possible temporal contiguities that

were lower than the observed one (for a detailed discussion of the calculation method, see Polyn et al., 2009; Sederberg, Miller, Howard, & Kahana, 2010; C. Yang, Zhao, et al., 2021).[6]

As shown in Figure 3A, TCSs were significantly greater in the no-JOL (M = .73, SD = .13) than in the JOL blocks (M = .69, SD = 0.11), difference = .042 [.005, .078], $t(26)$ = 2.36, p = .03, Cohen's d = 0.45, revealing that when freely recalling the word lists, participants were more likely to retain the relative order of studied items in the no-JOL than in the JOL conditions. As shown in Figure 3B, 17 participants showed a negative reactivity effect on TCSs, and 10 showed the converse patterns. Overall, these findings are again consistent with the item-order account's proposal that making JOLs impairs inter-item relational (serial) processing.

Correlation between free recall and TCSs

Previous studies established an association between temporal clustering and free recall performance, suggesting that retention of temporal information helps to guide output order and facilitates free recall (Dalezman, 1976; Forrin & MacLeod, 2016; Jonker & MacLeod, 2015; Mangels, 1997; C. Yang, Zhao, et al., 2021). The same finding was also observed here: across participants, there was a significantly positive correlation between TCSs and free recall in both the JOL (r = .63, 95% CI = [.33, .82], p < .001) and no-JOL (r = .45, 95% CI = [.08, .71], p = .019) conditions (see Figure 3C).

Discussion

Overall, the TCS results are consistent with the main findings from Experiment 1, reflecting a negative reactivity effect of soliciting JOLs on processing of temporal

[6] Interested readers can find a tutorial about the calculation method and the data-analysis scripts at https://memory.psych.upenn.edu/CRP_Tutorial

information, a result in keeping with the prediction from the item-order account that making JOLs impairs relational processing.

The TCS results may provide an explanation for the absence of positive reactivity on the number of items freely recalled. Specifically, we conjecture that making JOLs enhances item-specific processing, leading to a positive reactivity effect, and concurrently impairs processing of temporal information, leading to a negative effect. Because free recall is related to retention of both item-specific and inter-item relational (temporal order) information (Aka, Phan, & Kahana, 2021), it is possible that these positive and negative consequences might cancel out each other, causing an overall null reactivity effect of making JOLs on free recall performance. Future research could directly test this speculation.

Experiment 3

Experiments 1 and 2 demonstrated a negative reactivity effect on order reconstruction and temporal clustering. As discussed above, these negative reactivity effects might result from the fact that making item-by-item JOLs impaired processing of temporal information, as suggested by the item-order account. However, there is another possible explanation: they might derive from dual-task (encoding plus making JOLs) costs. Mitchum et al. (2016) proposed that making concurrent JOLs might borrow limited cognitive resources from the primary learning task, leading to a weaker encoding of temporal information. In addition, frequent task-switching between encoding and monitoring might also induce dual-task costs and lead to negative reactive influences (for related discussion, see Janes, Rivers, & Dunlosky, 2018; Zhao et al., 2021).

Indeed, there is an important divergence in experimental procedures between the current experiments and previous studies (Tauber & Rhodes, 2012; H. Yang et al., 2015). In many previous studies exploring reactivity effects on word list learning, item-by-item JOLs were made immediately *after* participants studied each word, and the total duration of word exposure in the JOL and no-JOL conditions were not perfectly matched. That is, in the JOL condition, participants had the opportunity to continue processing the just-studied item when making JOLs. By contrast, in the current experiments, item-by-item JOLs were made *while* participants studied each word, and the total duration in the JOL and no-JOL conditions were identical. This means that dual-task costs are more likely to occur in the current procedure. Hence, it is critical to determine whether the negative reactivity effects on order reconstruction and temporal clustering observed in Experiments 1 and 2 resulted from dual-task costs or from deleterious effects of making JOLs on processing of temporal information.

As articulated above, the dual-task costs explanation predicts a deleterious effect of making JOLs on both item-specific and relational encoding. The item-order account, in contrast, proposes that making JOLs enhances item-specific processing but disrupts relational processing. Experiment 3 was conducted to further explore whether making item-by-item JOLs while participants study each word can facilitate item memory. If so, there would be little need to worry about dual-task costs in this procedure.

To answer this question, we changed the test format to forced-choice recognition. It is obvious that recognition memory relies more on item memory with minimal involvement of temporal order memory (for related discussion, see Mulligan & Peterson, 2015), because items are presented in a random order in recognition tests, and no temporal order memory is

required to guide output sequence in recognition tests. Accordingly, the item-order account predicts a positive reactivity effect on recognition. By contrast, the dual-task costs explanation predicts a negative reactivity effect on recognition.

The difference in the reactivity effects between concurrent JOLs (i.e., making a JOL while studying each word) and immediate JOLs (i.e., making a JOL immediately after studying each word) is beyond the scope of the current study. Hence, Experiment 3 did not include a group of participants making immediate JOLs.[7] Instead, the primary aim of Experiment 3 was to assess whether making concurrent JOLs can induce a positive reactivity effect on recognition of list words. Recall that the dual-task costs explanation predicts a negative reactivity effect on recognition, whereas the item-order account expects a positive effect.

Method

Participants

Based on the effect size (Cohen's d = 1.33) of the reactivity effect on recognition recently documented by Zhao et al. (2021), a power analysis, conducted via G*power, showed that approximately 7 participants were required to observe a significant (2-tailed, α = .050) reactivity effect at 0.80 power. To be more conservative, we decided to increase the sample size to 25, the same as in Experiments 1 and 2. Finally, 25 participants (14 female), with a mean age of 22.16 (SD = 2.44) years, were recruited from BNU. They reported normal

[7] Future research could profitably employ concurrent (rather than immediate) JOLs to explore reactivity effects, because the positive effect of making immediate JOLs on memory might be confounded by additional processing (Soderstrom et al., 2015; Zechmeister & Shaughnessy, 1980; Zhao et al., 2021).

or corrected-to-normal vision, were tested individually in a sound-proofed cubicle, and received 50 RMB as compensation.

Materials

Four hundred and thirty-two Chinese words were selected from the Chinese word database (Cai & Byrsbert, 2010). For these 432 words (word frequency was 2.95 ~ 51.33 per million), 48 were used for practice and the remaining 384 were used in the formal experiment. For these 384 words, 192 were the same as those in Experiments 1 and 2, were studied in the learning phase, and served as "old" items in the forced-choice recognition test. The other 192 words, which were not used in Experiments 1 and 2, were used as "new" items (i.e., lures) in the recognition test.

As in Experiments 1 and 2, for each participant, the to-be-studied words were randomly divided into 16 lists, and these 16 lists were then randomly assigned to four blocks, with two blocks randomly selected as JOL blocks and the other two as no-JOL blocks. The presentation sequence of words in each list, the list sequence in each block, and the block sequence were randomly decided by the computer.

Design and procedure

The experiment involved a within-subjects design (study method: JOL *vs.* no-JOL). The procedure was largely similar to those in Experiments 1 and 2, but the list-by-list distractor tasks and free recall tests were removed. Instead, participants were instructed to study the words one-by-one, list-by-list, and block-by-block in preparation for a later recognition test. After studying all four blocks, participants engaged in an 8 min (= 30 s * 16) distractor task identical to that used in Experiments 1 and 2.

Finally, they took a forced-choice recognition test. For each participant, the 192 studied and 192 new words were randomly paired, with an "old" and a "new" word in each pair. The 192 pairs were presented one-by-one in a random order, with one word randomly presented on the left side of the screen and the other on the right side. Participants were instructed to decide which word was "old" (studied). There was no time pressure and no feedback in the forced-choice recognition test. For each trial in the recognition test, correct selection of the old word was scored as 1, and incorrect selection of the new one was scored as 0. For each participant, test performance in each condition was calculated as averages of those 0 and 1 scores. The whole experiment lasted about 50 minutes.

Note that Experiment 3 administered the forced-choice recognition test after participants studied all 16 lists rather than immediately after they studied each list as in Experiments 1 and 2. We delayed the recognition test to reduce any ceiling effect in recognition performance. In a pilot study, we found that participants could correctly recognize almost all words when recognition tests were administered immediately after they studied each list.

Results

Reactivity effect on recognition

Recognition performance for both the JOL and no-JOL words is shown in Figure 4A. Recognition performance was significantly better in the JOL (M = .87, SD = .11) than in the no-JOL blocks (M = .79, SD = .12), difference = .088 [.052, .124], $t(24) = 5.02$, $p < .001$, Cohen's d = 1.00, revealing a large positive reactivity effect of soliciting JOLs on recognition performance. As shown in Figure 4B, only 3 participants showed a negative reactivity effect, while the other 22 showed a positive effect.

Discussion

Overall, these recognition results reveal that making concurrent JOLs enhances item memory. This finding should allay concerns about dual-task costs induced by making concurrent JOLs, and provides further support to the main proposal of the item-order account: making concurrent JOLs facilitates item memory but impairs inter-item relational (temporal order) memory.

Experiment 4

In the above experiments, we observed dissociable effects of making JOLs on inter-item relational (order reconstruction and temporal clustering in Experiments 1 and 2) and item (recognition in Experiment 3) memory. In Experiment 4, we employed a within-subjects design to test whether these dissociable effects can be demonstrated simultaneously within the same participants and using the same stimuli. Recall that the item-order account predicts a negative reactivity effect on order reconstruction but a positive reactivity effect on recognition.

Method

Participants

Given that (1) the number of study trials in each test format condition was reduced by half and (2) the current experiment was designed to concurrently detect a negative effect on order reconstruction and a positive reactivity on recognition, we decided to double the sample size compared to previous experiments to approximately maintain statistical power. Finally, 60 participants (41 female), with a mean age of 21.00 (*SD* = 2.07) years, were recruited from

BNU. They reported normal or corrected-to-normal vision, were tested individually in a sound-proofed cubicle, and received 50 RMB as compensation.

Materials

Three hundred and twelve words were selected from the items used in Experiment 3. Specifically, 24 words were used for practice and the remaining 288 were used in the formal experiment. Among these 288 words, 192 served as study words, with the other 96 as "new" words presented in the forced-choice recognition test.

For each participant, the 192 words were randomly divided into 16 lists, with 12 words in each list, and these 16 lists were randomly assigned into 4 blocks. The computer randomly selected two blocks as the JOL blocks and the other two as the no-JOL blocks. The presentation sequence of words in each list, the list sequence in each block, and the block sequence were randomly determined by the computer.

Design and procedure

Experiment 4 involved a 2 (study method: JOL *vs.* no-JOL) × 2 (test format: order reconstruction *vs.* recognition) within-subjects design. The procedure in the learning phase was identical to that in Experiments 1 and 2 except that participants did not undertake the list-by-list interim tests.[8] Specifically, after participants studied each word list, they engaged in the same 30 s mathematical distractor task. The study phase of the next list commenced immediately afterward. This cycle repeated until all 16 lists had been studied.

[8] Another noteworthy point is that, in Experiment 4, participants were not informed that there were two types of memory tests. As explained below, whether a given word list would be presented in the order-reconstruction or forced-choice recognition test was randomly determined by the computer after participants studied all words. Informing participants that there were two types of tests might confuse them about how to make a JOL for a given word because they did not know in which format it would be finally tested. Therefore, in Experiment 4, we simply instructed participants to study the words in preparation for an unspecified memory test.

After studying all words, participants took two memory tests: order reconstruction and forced-choice recognition. The order of these two memory tests was randomized. The computer randomly selected 8 lists (i.e., 2 lists from each of the 4 blocks) and assigned them to the order reconstruction test condition, with the other 8 lists assigned to the forced-choice recognition test condition.

In the order reconstruction test condition, the 8 lists (i.e., 4 JOL lists and 4 no-JOL lists) were tested one-by-one in random order. For each list, the 12 words were presented in random order on the same screen, and participants were asked to re-order them in the order they appeared in that list.

In the forced-choice recognition test, 96 "old" (i.e., 12 words from each of the 8 lists) and 96 "new" words were randomly paired. The computer randomly presented the 96 word pairs one-by-one and list-by-list. The presentation sequence of words in each list and the list sequence was randomly decided by the computer. For each word pair, participants were asked to indicate which word was "old".

There was no time pressure or feedback in either memory test. The whole experiment lasted about 50 minutes.

Results

A mixed ANOVA was conducted with test order (order reconstruction test first *vs*. recognition test first) as the between-subjects variable, study method and test format as the within-subjects variables, and test performance as the dependent variable. The results showed no main effect of test order ($p = .19$), no interaction between test order and study method (p

= .15), no interaction between test order and test type (p = .29), and no three-way interaction (p = .18). Hence, below we do not further discuss test order.

Reactivity effects on order construction and recognition

A 2 × 2 repeated measures ANOVA assessed the effects of study method (JOL *vs.* no-JOL) and test format (order reconstruction *vs.* recognition) on test performance. The main effect of test format was significant, $F(1, 59) = 1772.34$, $p < .001$, $\eta_p^2 = .97$, with superior test performance in the forced-choice recognition than in the order reconstruction test (see Figure 5A). There was no main effect of study method, $F(1, 59) < 0.001$, $p = 1.000$, $\eta_p^2 < .001$.

Of critical interest, there was a significant study method by test format interaction, $F(1, 59) = 52.12$, $p < .001$, $\eta_p^2 = .47$. Consistent with Experiment 1, order reconstruction performance was significantly poorer in the JOL (M = .15, SD = .13) than in the no-JOL conditions (M = .21, SD = .14), difference = -.060 [-.089, -.031], $t(59) = -4.17$, $p < .001$, Cohen's d = -0.54, reflecting a negative reactivity effect on order reconstruction (see the left pair of bars in Figure 5A). As shown in Figure 5B, in the order reconstruction test, 38 participants showed a negative reactivity effect, 15 showed the converse pattern, and there were 7 ties.[9]

By contrast, but consistent with Experiment 3, recognition performance was significantly better in the JOL (M = .88, SD = .09) than in the no-JOL conditions (M = .82, SD = .14), difference = .060 [.035, .085], $t(59) = 4.74$, $p < .001$, Cohen's d = 0.61, reflecting a positive

[9] As supplemental results, we also calculated absolute difference scores of serial positions in the order reconstruction test. These results showed the exact same pattern, with larger distance scores in the JOL (M = 3.34, SD = 0.87) than in the no-JOL condition (M = 2.87, SD = 0.97), difference = 0.469 [0.277, 0.661], $t(59) = 4.89$, $p < .001$, Cohen's d = 0.63. Furthermore, a Spearman rank correlation was calculated for each participant in each word list, which was then averaged in the JOL and no-JOL conditions, respectively. The results also showed that the correlation was weaker in the JOL (M of r_s = .21, SD = .27) than in the no-JOL (M of r_s = .35, SD = .29) condition, difference = -.136 [-.197, -.076], $t(59) = -4.50$, $p < .001$, Cohen's d = -0.58.

reactivity effect on recognition (see the right pair of bars in Figure 5A). As shown in Figure 5B, in the forced-choice recognition test, 40 participants showed a positive reactivity effect, 17 showed the converse pattern, and there were 3 ties.

Discussion

Experiment 4 successfully detected dissociable effects of making JOLs on order reconstruction (negative) and forced-choice recognition (positive) with the same participants and stimuli, supporting the item-order account to explain the reactivity effect on word list learning.

Meta-analysis

Finally, a meta-analysis was conducted to integrate results across studies to explore the moderating role of test format in the reactivity effect on word list learning. As discussed above, according to the item-order account, we expected to observe a negative reactivity effect in order reconstruction tests and a positive effect in recognition tests. We also expected to observe a larger (i.e., more positive) reactivity effect on recognition than on free recall because the former is more dependent on item memory and less on temporal order memory.

Literature identification, inclusion criteria, and calculation methods

We used the following terms to search for relevant studies in electronic databases: [“judgment* of learning” OR “judgement* of learning” OR “JOL*”] AND [“reactivity” OR “reactive influence*”]. The literature search was performed in Web of Science and ProQuest (composed of 26 databases, including PsychArticles, PsychInfo, Psychology Database, Education Database, ProQuest Dissertations & Theses Global Database, Ebook Central, Business Market Research Collection, and so on). In addition, the reference list of a recent

meta-analysis was manually screened (Double et al., 2018). Furthermore, the 70 Google Scholar citations of Mitchum et al. (2016) were checked. The literature search was finished in October 2021.

The inclusion and exclusion criteria were as follows:

1. Only studies which compared making JOLs with not making JOLs (i.e., passive study control) were included. Studies or effects which compared making JOLs with other processing strategies (e.g., forming mental images) were excluded.
2. Because the current study especially focused on the reactivity effect on word list learning and the item-order account, only studies using word lists were included. Other studies, which explored reactivity effects on learning of other types of materials such as word pairs (Janes et al., 2018) and text passages (Ariel et al., 2021), were excluded.
3. Only studies written in English were considered.

In total, we identified 8 studies eligible for the meta-analysis (Halamish, 2018; Senkova & Otani, 2021; Stevens, 2019; Tauber & Rhodes, 2012; Tekin & Roediger, 2020; H. Yang et al., 2015; Zechmeister & Shaughnessy, 1980; Zhao et al., 2021). In addition, the results from our Experiments 1-4 were also included. Recall that, in Experiment 4, each participant undertook both the order reconstruction and forced-choice recognition tests. To avoid dependency among these two effects, we randomly assigned the 60 participants to two groups, with 30 participants in each group. For one group, we calculated a Cohen's *d* to represent the reactivity effect on order reconstruction, and for the other group we computed a

Cohen's *d* to represent the reactivity on recognition. In this way, we ensured that these two effects were derived from independent participants.

Overall, from these studies, we extracted 25 effects (Cohen's *d*s) based on data from 1,638 participants. Note that, for the 14 recognition effects listed in Figure 6, their effect sizes were calculated based on hit rates, because all studies reported hit rate results and some of them did not report discrimination (*d'*) results, making it impossible to calculate effect sizes based on *d'* (e.g., Halamish, 2018). In addition, our Experiments 3 and 4 and Zhao et al. (2021) used a forced-choice recognition procedure, which is distinct from the conventional old/new recognition test. Hence, for consistency, we calculated effect sizes using hit rates for all recognition effects.

Following Double et al. (2018), all Cohen's *d*s were transformed to Hedges' *g*s using the formulae provided by Borenstein, Hedges, Higgins, and Rothstein (2009). The characteristics of the 25 effects (test formats and effect sizes) are summarized in Figure 6. A positive value of *g* indicates a positive reactivity effect and a negative value represents a negative effect.

All meta-analyses were conducted via the R *metafor* package (Viechtbauer, 2010). Given that some effects were extracted from the same studies (e.g., Halamish, 2018), all meta-analyses were conducted using multilevel random-effects models.

Results

Reactivity effects on word list learning

A multilevel random-effects meta-analysis found that providing JOLs significantly enhanced memory (combining free recall, recognition, and order reconstruction), Hedges' g = 0.47, 95% CI [0.280, 0.653], $Z = 4.90$, $p < .001$ (see Figure 6), reflecting an overall positive

reactivity effect on word list learning. Heterogeneity amongst the effects was substantial, $Q(24) = 155.79$, $p < .001$, indicating the necessity of exploring potential moderators of the included effects.

Figure 7 is a funnel plot showing the relationship between the effects and their corresponding standard errors (*SE*s). There was no obvious asymmetry of the funnel plot, suggesting little need to worry about publication bias. In addition, a multilevel random-effects meta-regression analysis (Stanley, 2008) found no statistically detectable relationship between effect sizes and *SE*s, slope coefficient = -0.37 [-5.327, 4.594], $Z = -0.15$, $p = .89$, confirming low risk of publication bias. Note however that the *SE*s are fairly similar across studies (because they included similar sample sizes), so this test for bias is likely to be underpowered.

Moderating role of test format

Now we move to the main interest of the meta-analysis, to test the moderating role of test format in reactivity effects on word list learning. A multilevel random-effects sub-group meta-analysis found that the heterogeneity amongst different test formats was substantial, $Q(2) = 44.72$, $p < .001$, indicating that test format did significantly moderate reactivity effects.

The integrated results for each test format are summarized in Figure 6. The reactivity effect on order reconstruction was significantly negative, $g = -0.61$, 95% CI [-0.988, -0.231], $Z = -3.16$, $p = .002$, as expected given that the only studies included in this group are our Experiments 1 and 4. The reactivity effect on free recall was significantly positive, but it was relatively weak, $g = 0.32$, 95% CI [0.126, 0.511], $Z = 3.25$, $p = .001$. By contrast, the effect

size of reactivity on recognition was medium-to-large, g = 0.72, 95% CI [0.573, 0.867], Z = 9.60, p < .001.

Critically, as predicted by the item-order account, the reactivity effect on recognition was significantly larger than the effect on free recall, difference in g = 0.40 [0.159, 0.644], Z = 3.25, p < .001, consistent with the main proposal of the item-order account.

Excluding results from the current study

It is intriguing that our Experiment 2 observed no reactivity in the free recall test, while the meta-analysis demonstrated positive (albeit modest) reactivity in free recall tests. These divergent findings might result from the difference in experimental procedures between our Experiment 2 and previous studies. As noted above, many of the studies included in the meta-analysis asked participants to make item-by-item JOLs immediately after studying each item (Tauber & Rhodes, 2012; H. Yang et al., 2015). The positive reactivity effects on free recall documented in previous studies might result from additional processing. Specifically, when participants made JOLs after studying, they might re-process (continue to think about) the just-studied word, leading to an enhancing effect (for detailed discussion, see Zechmeister & Shaughnessy, 1980). By contrast, our Experiment 2 instructed participants to make concurrent JOLs while they studied each word, and the total exposure duration was matched between the JOL and no-JOL conditions. Hence, it is unsurprising that Experiment 2 observed no reactivity effect on free recall. Noteworthy is that Senkova and Otani (2021) also detected no reactivity effect on free recall of unrelated words when the total exposure duration was matched between the JOL and no-JOL conditions.

Considering the divergence in experimental procedures, we re-ran the meta-analysis to assess the reactivity effects on free recall and recognition with the five effects from the current study excluded. This meta-analysis was conducted to test if the findings remain the same after excluding results from the current study.

The meta-analytic results indeed showed the same patterns. The overall reactivity effect on word list learning was positive, $g = 0.57$, 95% CI [0.417, 0.718], $p < .001$, the effect on free recall was significantly positive, $g = 0.35$, 95% CI [0.130, 0.567], $p = .002$, and the same for the effect on recognition, $g = 0.71$, 95% CI [0.535, 0.874], $p < .001$. In addition, the reactivity effect on recognition was significantly greater than the effect on free recall, difference in $g = 0.36$, 95% CI [0.080, 0.633], $p = .010$. Overall, after excluding all results from the current study, the meta-analysis showed the same patterns, establishing the reliability of the results and providing further support for the item-order account to explain the reactivity effect on word list learning.

Discussion

These meta-analyses demonstrate that making item-by-item JOLs significantly impairs order reconstruction, but enhances free recall and recognition. Moreover, the enhancing effect on recognition is significantly greater than the effect on free recall, regardless of whether results from the current study were included or excluded. These findings jointly support the item-order account's explanation for the reactivity effect on word list learning.

General Discussion

The current research is the first to (1) explore the reactive influence of making metacognitive judgments on temporal (serial) order memory – an important form of inter-item relational memory, and (2) test whether the item-order account is a valid explanation of the reactivity effect on word list learning. The major results are summarized in Table 1. Below we discuss the theoretical and practical implications of the documented findings.

Theoretical implications

The theoretical goal of the current study was to test whether the item-order account is a valid explanation of the reactivity effect on word list learning. Recall that all existing theories (i.e., cue-strengthening, changed-goal, and positive reactivity) have difficulty in explaining the reactivity effect on word list learning (see Appendix A), and the current study is the first to propose the item-order account to explain this effect.

The item-order account assumes that making JOLs enhances item-specific processing because participants have to closely encode and analyze the current item in order to make an appropriate JOL, and this enhanced item-specific processing then detracts from relational processing, leading to inferior memory for inter-item relations (McDaniel & Bugg, 2008). The findings from Experiments 1-4 and the meta-analysis jointly support both aspects of this account. Regarding inter-item relational processing, Experiment 1 observed that making JOLs decreased order reconstruction accuracy, and Experiment 2 demonstrated a negative reactivity effect on temporal clustering in free recall. Regarding item-specific processing, Experiment 3 observed that making concurrent JOLs substantially boosted forced-choice recognition accuracy. Experiment 4 simultaneously replicated the negative reactivity effect on order reconstruction and the positive effect on recognition within the same participants and stimuli.

The positive reactivity effect on recognition, as documented in Experiments 3 and 4, should also allay the concern that negative reactivity on order reconstruction (Experiments 1 and 4) and temporal clustering (Experiment 2) simply resulted from dual-task costs (i.e., costs induced by frequent task-switching between encoding and monitoring) rather than an impairment effect of making JOLs on relational processing.

The meta-analytic findings also support the item-order account. Even though the reactivity effect on word list learning was positive overall, test format significantly moderated the reactive influences. The overall reactivity effect on order reconstruction was negative, re-confirming the impairment effect of making concurrent JOLs on temporal order memory. By contrast, the effect on recognition was positive, consistent with the hypothesis that making JOLs promotes item-specific processing.

Finally, but importantly, the meta-analysis showed that the positive reactivity effect on recognition was significantly greater than the effect on free recall. Because making JOLs facilitated item memory but impaired temporal order memory, these facilitating and impairing consequences may offset each other, leading to a modest reactivity effect on free recall (Mulligan & Peterson, 2015). By contrast, recognition performance is heavily dependent on item memory and less related to temporal order memory (Mulligan & Peterson, 2015), and therefore a larger positive reactivity effect on recognition is obtained.

There is another possible explanation for the larger reactivity effect on recognition than free recall. It is possible that making JOLs facilitates both familiarity and recollection of study items. It is well-known that both aspects of memory contribute to recognition performance, whereas free recall performance is mainly related to recollection memory

(Hockley & Consoli, 1999; Yonelinas, 2002). The larger reactivity effect on recognition might result from the additional benefits of making JOLs on familiarity. Although this might account for the larger reactivity effect on recognition than free recall, it cannot explain why the reactivity effect on temporal order memory is negative (as shown in Experiments 1, 2, and 4). By contrast, the item-order account provides a reasonable explanation for all findings obtained here.

In summary, the current study found that making concurrent JOLs reactively enhances item memory but impairs temporal order memory. The findings support the item-order account to explain the reactivity effect on word list learning.

Although the item-order account is a viable explanation of the reactivity effect on word list learning, it has difficulty in explaining reactivity effects on memory for other types of materials. For instance, it cannot explain why making JOLs improves memory for related word pairs but exerts minimal reactive influence on memory for unrelated word pairs (Soderstrom et al., 2015). Similarly, although the cue-strengthening theory provides a reasonable explanation of the reactivity effect on memory for word pairs, it cannot explain the reactivity effect on word list learning (see Appendix A for further discussion). It is possible that different mechanisms may contribute to reactivity effects on memory for different types of information. More research on the cognitive underpinnings of reactivity effects is called for.

Practical implications

Besides the above-discussed theoretical implications, our findings also have important implications for future research design and interpretation. Numerous studies have been

conducted to assess JOL accuracy, in which participants made item-by-item JOLs during or after studying each item (for reviews, see Rhodes & Tauber, 2011; C. Yang, Yu, et al., 2021). Word lists are one of the most widely-used material types in previous JOL research (Rhodes & Tauber, 2011; C. Yang, Yu, et al., 2021). These studies relied on an unverified assumption that collecting item-by-item JOLs does not affect memory (Spellman & Bjork, 1992).

This assumption has been repeatedly disproved by previous research and the current study demonstrating reactivity effects on item memory (Double et al., 2018; Myers et al., 2020; Witherby & Tauber, 2017; Zhao et al., 2021). Furthermore, our findings provide the first illustration that making JOLs not only reactively changes item memory, but also alters inter-item relational (temporal order) memory. These reactivity effects highlight that future research should develop more nuanced procedures to remove or at least mitigate these reactivity effects when assessing metacognitive (JOL) accuracy. In addition, researchers should bear the reactivity effect in mind when interpreting their JOL accuracy results because JOLs might fail to measure what they intend to assess.

Limitations and future research directions

The current study suffers from several limitations. First, all of the reported experiments were conducted in the laboratory, and the stimuli were word lists. It is hence premature to infer the existence of reactive influences of making JOLs on inter-item relational memory in real life and educational settings. Future research can profitably conduct field studies and use naturalistic stimuli (e.g., grocery lists, skeleton bone names) to explore this critical question.

Second, the current study mainly explored the reactive influences of making item-by-item JOLs on inter-item relational memory. However, it is unusual that learners judge their

learning status while or after studying each item. Instead, in educational settings, learners may frequently judge their learning status while reading each text section (or a book chapter) or after attending each class. Future research is encouraged to explore whether section-by-section JOLs reactively affect inter-section relational memory.

Third, temporal order memory is just one form of inter-item relational memory. It is unknown whether making JOLs reactively disrupts other forms such as memory for semantic relations among study items. For instance, does making JOLs affect knowledge integration among related knowledge points? This research question is of considerable importance for guiding educational practice.

Finally, the current experiments quantified the reactivity effect as the signed difference in memory performance between the JOL and no-JOL conditions. In the no-JOL condition, participants did not need to attend to any additional slider-rating task. Likewise, most previous reactivity studies compared making JOLs with a passive control. Future research should employ a better-matched no-JOL control condition (e.g., asking participants to make other forms of slider ratings, unrelated to metamemory) to examine reactivity.

Concluding Remarks

Although soliciting JOLs reactively facilitates word list learning overall, it significantly impairs temporal order memory. Test format moderates the reactivity effect on word list learning, with the effect being negative on order reconstruction and temporal clustering, but positive on recognition. The reactivity effect on recognition is larger than the effect on free recall. The item-order account is a valid explanation of the enhancing effect of making JOLs on item memory and the disrupting effect on inter-item relational (temporal order) memory.

References

Aka, A., Phan, T. D., & Kahana, M. J. (2021). Predicting recall of words and lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 47*, 765–784. doi:10.1037/xlm0000964

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. doi:10.1007/s10648-020-09556-8

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & Cognition, 35*, 201–210. doi:10.3758/BF03193441

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

Cai, Q., & Byrsbert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *Plos One, 5*, e10729. doi:10.1371/journal.pone.0010729

Dalezman, J. J. (1976). Effects of output order on immediate, delayed, and final recall performance. *Journal of Experimental Psychology: Human Learning and Memory, 2*, 597–608. doi:10.1037/0278-7393.2.5.597

Diamond, N. B., & Levine, B. (2020). Linking detail to temporal structure in naturalistic-event recall. *Psychological Science, 31*, 1557–1572. doi:10.1177/0956797620958651

Double, K. S., Birney, D. P., & Walker, S. A. (2018). A meta-analysis and systematic review of reactivity to judgements of learning. *Memory, 26*, 741–750.

doi:10.1080/09658211.2017.1404111

Dunlosky, J., & Tauber, S. K. (2016). *The Oxford handbook of metamemory*: Oxford University Press.

Engelkamp, J., Biegelmann, U., & McDaniel, M. A. (1998). Relational and item-specific information: Trade-off and redundancy. *Memory, 6*, 307–333. doi:10.1080/741942360

Farrell, S. (2012). Temporal clustering and sequencing in short-term memory and episodic memory. *Psychological Review, 119*, 223–271. doi:10.1037/a0027371

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. doi:10.3758/ BF03193146

Fernandes, M. A., Wammes, J. D., & Meade, M. E. (2018). The surprisingly powerful influence of drawing on memory. *Current Directions in Psychological Science, 27*, 302–308. doi:10.1177/0963721418755385

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

Forrin, N. D., & MacLeod, C. M. (2016). Order information is used to guide recall of long lists: Further evidence for the item-order account. *Canadian Journal of Experimental Psychology, 70*, 125–138. doi:10.1037/cep0000088

Guynn, M. J., McDaniel, M. A., Strosser, G. L., Ramirez, J. M., Castleberry, E. H., & Arnett, K. H. (2014). Relational and item-specific influences on generate–recognize processes in recall. *Memory & Cognition, 42*, 198–211. doi:10.3758/s13421-013-0341-6

Halamish, V. (2018). Can very small font size enhance memory? *Memory & Cognition, 46*, 979–993. doi:10.3758/s13421-018-0816-6

Hockley, W. E., & Consoli, A. (1999). Familiarity and recollection in item and associative recognition. *Memory & Cognition, 27*, 657–664. doi:10.3758/BF03211559

Hunt, R. R., & Einstein, G. O. (1981). Relational and item-specific information in memory. *Journal of verbal learning and verbal behavior, 20*, 497–514. doi:10.1016/S0022-5371(81)90138-9

Hunt, R. R., & McDaniel, M. A. (1993). The enigma of organization and distinctiveness. *Journal of Memory and Language, 32*, 421–445. doi:10.1006/jmla.1993.1023

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. doi:10.3758/s13423-018-1463-4

Jeunehomme, O., Folville, A., Stawarczyk, D., Van der Linden, M., & D'Argembeau, A. (2018). Temporal compression in episodic memory for real-life events. *Memory, 26*, 759–770. doi:10.1080/09658211.2017.1406120

Jonker, T. R., Levene, M., & MacLeod, C. M. (2014). Testing the item-order account of design effects using the production effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*, 441–448. doi:10.1037/a0034977

Jonker, T. R., & MacLeod, C. M. (2015). Disruption of relational processing underlies poor memory for order. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*, 831–840. doi:10.1037/xlm0000069

Jonker, T. R., Wammes, J. D., & MacLeod, C. M. (2019). Drawing enhances item information

but undermines sequence information in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 45*, 689–699. doi:10.1037/xlm0000610

Kahana, M. J. (1996). Associative retrieval processes in free recall. *Memory & Cognition, 24*, 103–109. doi:10.3758/BF03197276

Karpicke, J. D., & Zaromb, F. M. (2010). Retrieval mode distinguishes the testing effect from the generation effect. *Journal of Memory and Language, 62*, 227–239. doi:10.1016/j.jml.2009.11.010

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

Li, B., Zhao, W., Zheng, J., Hu, X., Su, N., Fan, T., . . . Luo, L. (2021). Soliciting judgments of forgetting reactively enhances memory as well as making judgments of learning: Empirical and meta-analytic tests. *Memory & Cognition, Advance online publication.* doi:10.3758/s13421-021-01258-y

Lohnas, L. J., Polyn, S. M., & Kahana, M. J. (2011). Contextual variability in free recall. *Journal of Memory and Language, 64*, 249–255. doi:10.1016/j.jml.2010.11.003

MacLeod, C. M., & Bodner, G. E. (2017). The production effect in memory. *Current Directions in Psychological Science*, *26*, 390–395. doi:10.1177/0963721417691356

Mangels, J. A. (1997). Strategic processing and memory for temporal order in patients with frontal lobe lesions. *Neuropsychology, 11*, 207–221. doi:10.1037/0894-4105.11.2.207

McCurdy, M. P., Viechtbauer, W., Sklenar, A. M., Frankenstein, A. N., & Leshikar, E. D. (2020). Theories of the generation effect and the impact of generation constraint: A

meta-analytic review. *Psychonomic Bulletin & Review, 27*, 1139–1165. doi:10.3758/s13423-020-01762-3

McDaniel, M. A., & Bugg, J. M. (2008). Instability in memory phenomena: A common puzzle and a unifying explanation. *Psychonomic Bulletin & Review, 15*, 237–255. doi:10.3758/PBR.15.2.237

Mickes, L., Wixted, J. T., Shapiro, A., & Scarff, J. M. (2009). The effects of pregnancy on memory: Recall is worse but recognition is not. *Journal of Clinical and Experimental Neuropsychology, 31*, 754–761. doi:10.1080/13803390802488111

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. doi:10.1037/a0039923

Mulligan, N. W., & Lozito, J. P. (2007). Order information and free recall: Evaluating the item-order hypothesis. *Quarterly Journal of Experimental Psychology, 60*, 732–751. doi:10.1080/17470210600785141

Mulligan, N. W., & Peterson, D. J. (2015). Negative and positive testing effects in terms of item-specific and relational information. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*, 859–871. doi:10.1037/xlm0000056

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

Nairne, J. S., Riegler, G. L., & Serra, M. (1991). Dissociative effects of generation on item and order retention. *Journal of Experimental Psychology: Learning, Memory, and*

Cognition, 17, 702–709. doi:10.1037/0278-7393.17.4.702

Peterson, D. J., & Mulligan, N. W. (2013). The negative testing effect and multifactor account. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*, 1287–1293. doi:10.1037/a0031337

Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009). A context maintenance and retrieval model of organizational processes in free recall. *Psychological Review, 116*, 129–156. doi:10.1037/a0014420

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. doi: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. doi:10.1080/09658211.2021.1985143

Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 17*, 249–255. doi:10.1111/j.1745-6916.2006.00012.x

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychonomic Bulletin, 140*, 1432–1463. doi:10.1037/a0037559

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. doi:10.3758/BF03206472

Sederberg, P. B., Miller, J. F., Howard, M. W., & Kahana, M. J. (2010). The temporal contiguity effect predicts episodic memory performance. *Memory & Cognition, 38*, 689–699. doi:10.3758/mc.38.6.689

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

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. doi:10.1037/a0038388

Solway, A., Murdock, B. B., & Kahana, M. J. (2012). Positional and temporal clustering in serial order memory. *Memory & Cognition, 40*, 177–190. doi:10.3758/s13421-011-0142-8

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*, 315–316. doi:10.1111/j.1467-9280.1992.tb00680.x

Stanley, T. D. (2008). Meta-regression methods for detecting and estimating empirical effects in the presence of publication selection. *Oxford Bulletin of Economics and Statistics, 70*, 103–127. doi:10.1111/j.1468-0084.2007.00487.x

Stevens, A. R. (2019). *Examining the Effects of Making Judgments of Learning on True and False Memory.* (M.S.). Texas A&M University. Retrieved from https://search.proquest.com/dissertations-theses/examining-effects-making-

judgments-learning-on/docview/2239311839/se-2?accountid=8554

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

Tauber, S. K., & Witherby, A. E. (2019). Do judgments of learning modify older adults' actual learning? *Psychology and Aging, 34*, 836–847. doi: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. doi:10.1027/2151-2604/a000425

Tekin, E., & Roediger, H. L. (2021). The effect of delayed judgments of learning on retention. *Metacognition and Learning, 16*, 407–429. doi:10.1007/s11409-021-09260-0

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

Tulving, E. (1972). 12. Episodic and Semantic Memory. *Organization of memory/Eds E. Tulving, W. Donaldson, NY: Academic Press*, 381–403.

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

White, R. (2002). Memory for events after twenty years. *Applied Cognitive Psychology, 16*, 603–612. doi:10.1002/acp.819

Witherby, A. E., & Tauber, S. K. (2017). 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. doi:10.1016/j.jarmac.2017.08.004

Yang, C., Luo, L., Vadillo, M. A., Yu, R., & Shanks, D. R. (2021). Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review. *Psychological Bulletin, 147,* 394–435. doi:10.1037/bul0000309

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. doi:10.1037/xlm0000363

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. doi:10.1007/s11409-020-09254-4

Yang, C., Zhao, W., Luo, L., Sun, B., Potts, R., & Shanks, D. R. (2021). Testing potential mechanisms underlying test-potentiated new learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, Advance online publication*. doi:10.1037/xlm0001021

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

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46*, 441–517. doi:10.1006/jmla.2002.2864

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. doi:10.3758/bf03329756

Zhao, W., Li, B., Shanks, D. R., Zhao, W., Zheng, J., Hu, X., . . . Yang, C. (2021). When judging what you know changes what you really know: Soliciting metamemory judgments reactively enhances children's learning. *Child Development, 93,* 405–417. doi:10.1111/cdev.13689

Table 1. Summary of main findings from Experiments 1-4 and the meta-analysis.

Experiments	Measures	Effect size	Reactivity
Experiment 1	Order reconstruction	*d* = -0.703	Negative
Experiment 2			
	Free recall	*d* = 0.094	Minimal
	Temporal clustering	*d* = -0.454	Negative
Experiment 3	Forced-choice recognition	*d* = 1.004	Positive
Experiment 4			
	Order reconstruction	*d* = -0.538	Negative
	Forced-choice recognition	*d* = 0.612	Positive
Meta-analysis			
	Order reconstruction	*g* = -0.601	Negative
	Free recall	*g* = 0.319	Positive
	Recognition	*g* = 0.720	Positive

Note: Cohen's *d*s and Hedges' *g*s represent standardized differences in test performance between the JOL and no-JOL conditions, with positive values representing positive reactivity and negative values representing negative reactivity.

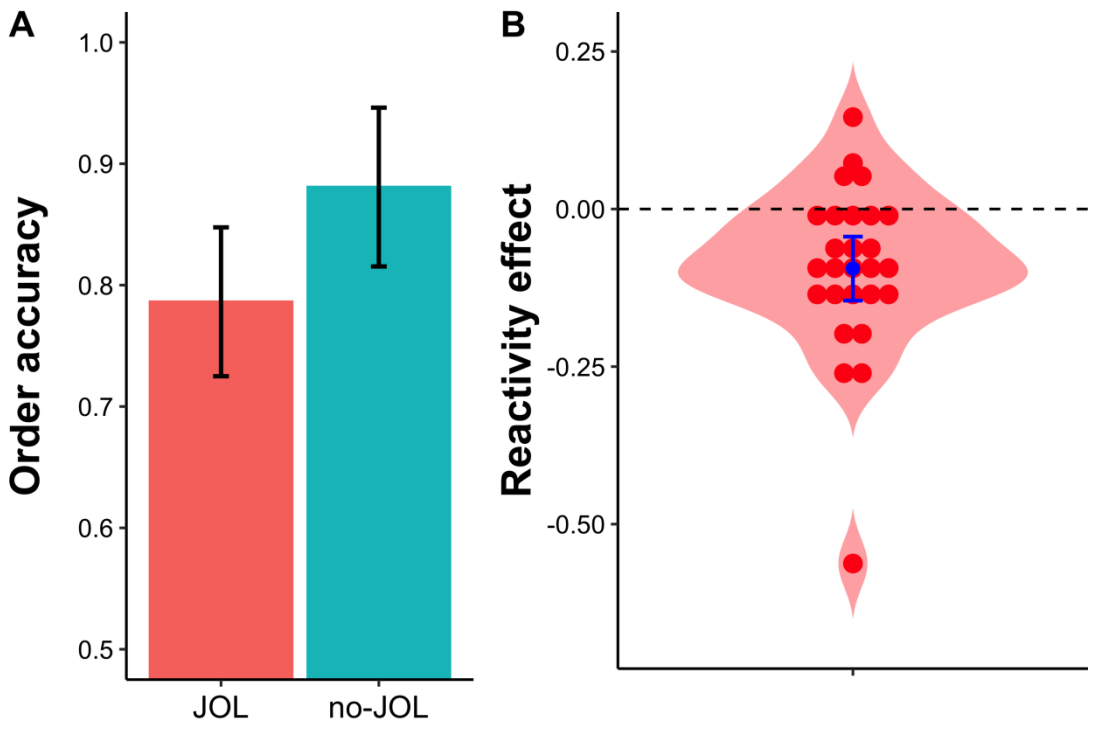

Figure 1. A: Order reconstruction performance (accuracy) as a function of study method (JOL *vs*. no-JOL) in Experiment 1. B: Violin plot showing the reactivity effect (i.e., the difference in test performance between JOL and no-JOL conditions). Each red dot represents one participant's reactivity effect score. Error bars represent 95% CI.

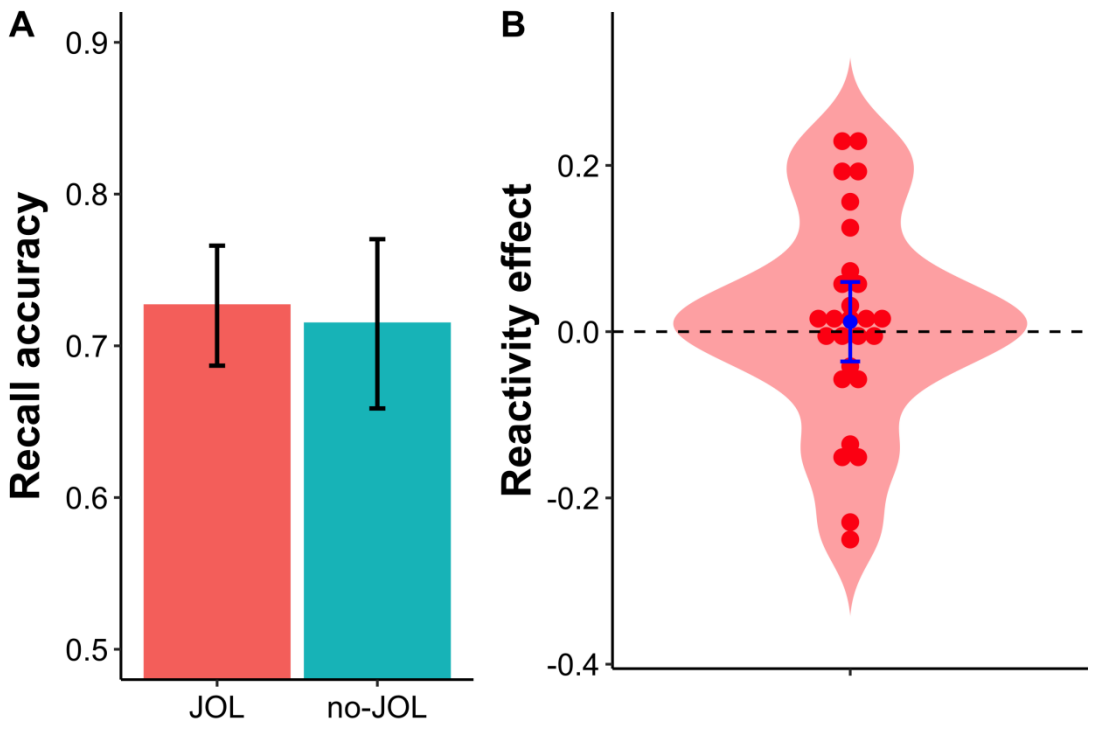

Figure 2. A: Recall performance (accuracy) as a function of study method (JOL *vs*. no-JOL) in Experiment 2. B: Violin plot showing the reactivity effect (i.e., the difference in test performance between JOL and no-JOL conditions). Each red dot represents one participant's reactivity effect score. Error bars represent 95% CI.

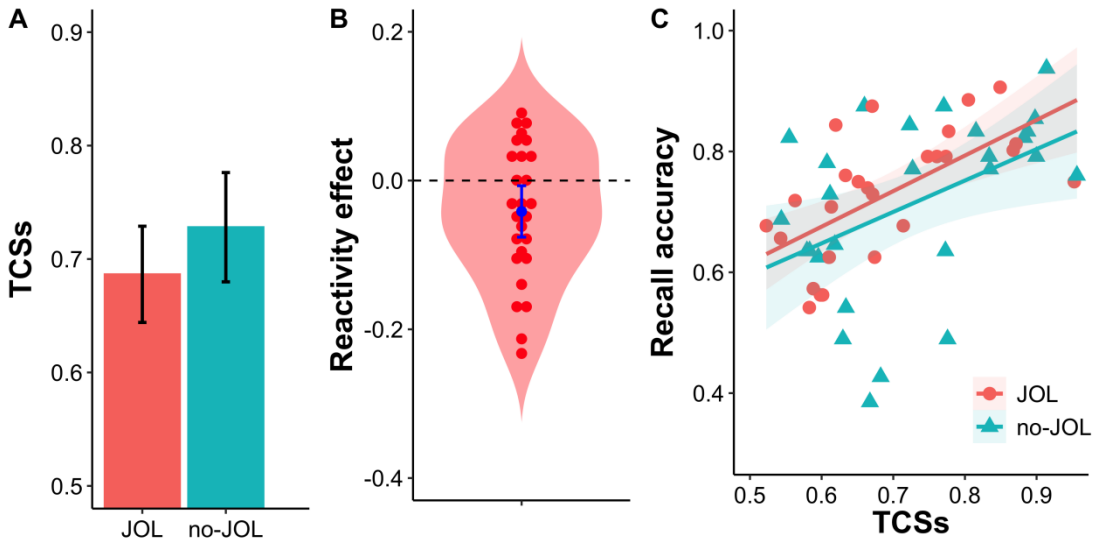

Figure 3. A: TCSs as a function of study method (JOL *vs*. no-JOL) in Experiment 2. B: Violin plot showing the reactivity effect (i.e., the difference in TCSs between JOL and no-JOL blocks). Each red dot represents one participant's difference in TCSs. Error bars represent 95% CI. C: Relationship between free recall accuracy and TCSs for each condition.

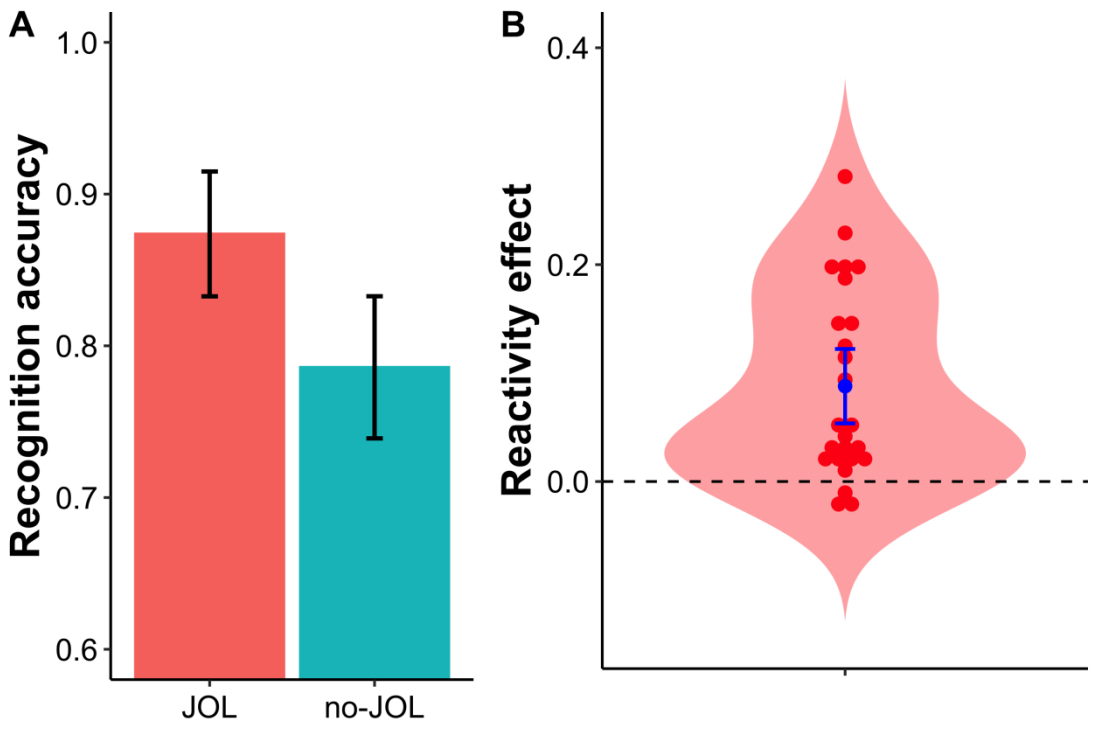

Figure 4. A: Recognition performance (accuracy) as a function of study method (JOL *vs*. no-JOL) in Experiment 3. B: Violin plot showing the reactivity effect (i.e., the difference in test performance between JOL and no-JOL conditions). Each red dot represents one participant's reactivity effect score. Error bars represent 95% CI.

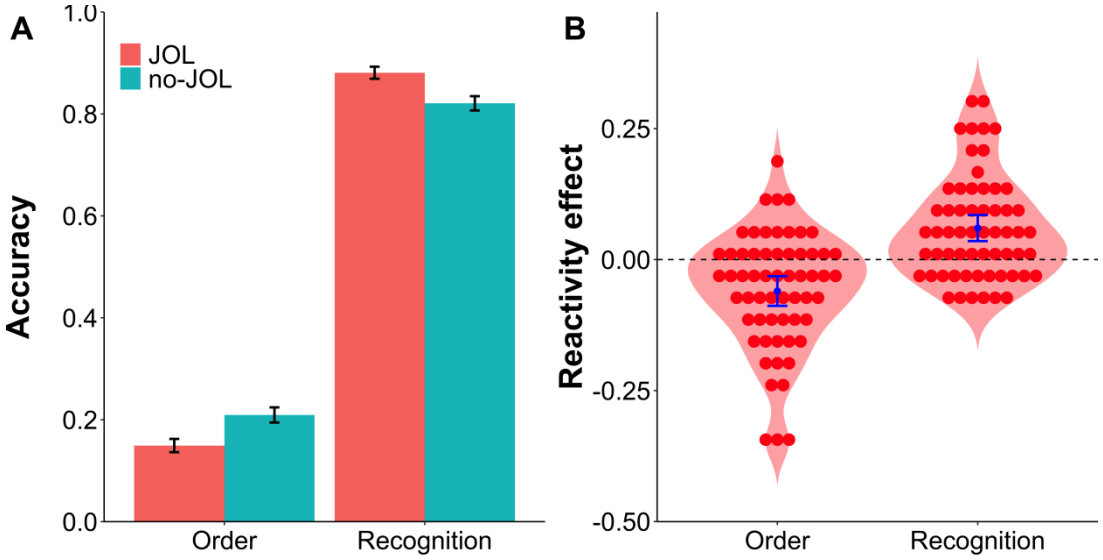

Figure 5. A: Accuracy (order reconstruction or recognition performance) as a function of study method (JOL *vs*. no-JOL) in Experiment 4. B: Violin plot showing the reactivity effect (i.e., the difference in test performance between JOL and no-JOL conditions) on order reconstruction and recognition respectively. Each red dot represents one participant's reactivity effect score. Error bars represent 95% CI.

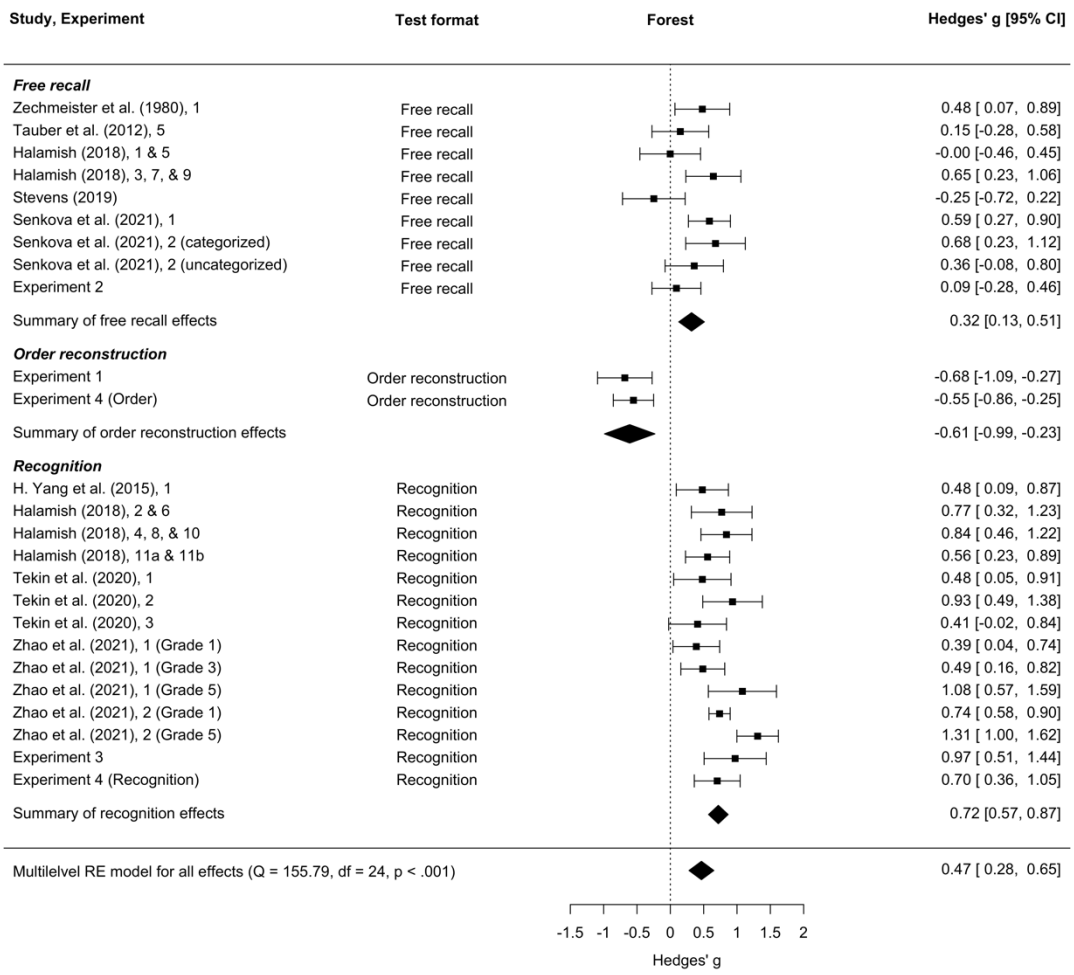

Figure 6. Forest plot summarizing the 25 effect sizes (Hedges' *g*s), their experimental characteristics (test format and effect sizes), and the multilevel random-effects (RE) meta-analysis results. Error bars represent 95% CI.

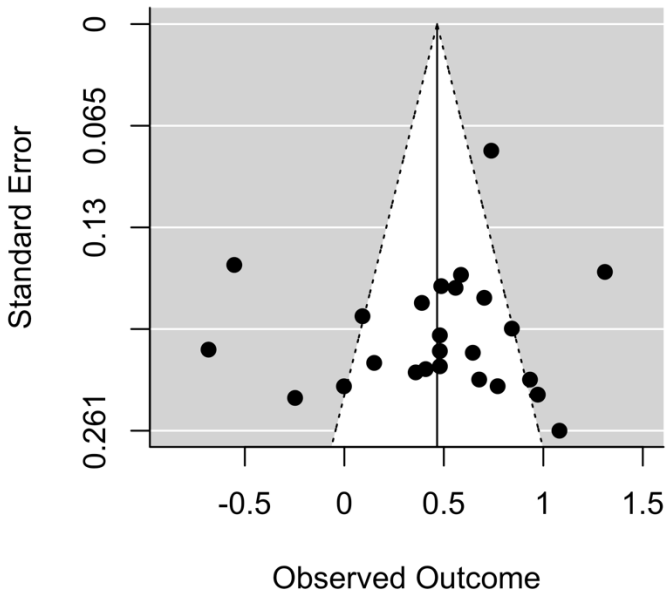

Figure 7. Funnel plot of the relationship between effect sizes (Hedges' *g*s) and their corresponding *SE*s.

Appendix A: Other Theories of Reactivity

Soderstrom et al. (2015) proposed a *cue-strengthening theory* to account for the reactivity effect on learning of word pairs. Soderstrom and colleagues observed that making JOLs significantly enhances recall of related word pairs, but has minimal reactive influence on recall of unrelated word pairs. They assumed that participants have to search for relevant cues to make a reasonable JOL for a given word pair, and the activated cues in turn strengthen the association between the cue and target for related word pairs, leading to positive reactivity. Because there is no pre-existing relatedness between the cue and target for unrelated pairs, making JOLs therefore fails to boost recall of unrelated pairs (for related discussion, see Janes et al., 2018).

Although the cue-strengthening theory provides a good explanation for reactivity effects on memory of word pairs, it cannot explain why making JOLs enhances recognition of unrelated list words (H. Yang et al., 2015; Zhao et al., 2021). In previous word list learning studies (H. Yang et al., 2015; Zhao et al., 2021), there were no semantic relations among the words, but making JOLs significantly improved recognition. Obviously, the cue-strengthening theory has difficulty in accounting for this finding.

Another account is the *changed-goal theory*, proposed by Mitchum et al. (2016) to explain the positive and negative reactivity effects on learning of a mixed list of related and unrelated word pairs. Mitchum et al. observed that making JOLs, compared with not making JOLs, numerically improved cued recall of related word pairs and significantly impaired recall of unrelated word pairs when related and unrelated pairs were studied in a mixed list.

Mitchum and colleagues assumed that making JOLs should enhance awareness of the difference in learning difficulty between related (easy) and unrelated (difficult) word pairs in a mixed list, and participants then adjust their learning goals from remembering all word pairs to especially remembering easy ones with a sacrifice of difficult ones, leading to positive reactivity on memory of related pairs and negative reactivity on memory of unrelated pairs. Although the changed-goal theory can explain the positive and negative reactivity effects on learning of a mixed list of easy and difficult materials, it cannot explain the overall positive reactivity effect on recognition of pure word lists.

The third account is the *positive reactivity theory* (Mitchum et al., 2016), which assumes that making JOLs drives participants to adopt more effective study strategies, leading to superior learning outcomes. Supporting evidence comes from Sahakyan, Delaney, and Kelley (2004), who observed that asking participants to make a JOL (i.e., predicting the number of words they would remember in a later memory test) following the study of a list of words caused them to shift from poor learning strategies (e.g., rote rehearsal) to more effective ones during subsequent study of a new list. It is worth noting that Mitchum et al. (2016) observed no difference in reported study strategies between their JOL and no-JOL conditions, which is inconsistent with the positive reactivity theory (for related findings, see Rivers, Janes, & Dunlosky, 2021). According to the positive reactivity theory, reactivity effects on word list learning should always be positive. However, previous findings run counter to this expectation. For instance, Stevens (2019) observed no reactivity effect on free recall of word lists (for related findings, see Tauber & Rhodes, 2012).

Overall, all the above-discussed theories (cue-strengthening, changed-goal, and positive reactivity) fail to provide compelling explanations for reactivity effects on word list learning. The results obtained in our Experiments 1-4 provide further challenges for all three of these theories.

Appendix B: JOL Results

Experiment 1

Participants provided concurrent JOLs to 98.2% (*SD* = 2.4%) of words in the JOL blocks. The average JOL was 69.10 (*SD* = 20.37). A gamma (*G*) correlation was calculated to measure the relative accuracy of JOLs for each participant. Specifically, order reconstruction performance was dummy coded (correct = 1; incorrect = 0), and then we calculated *G* between JOLs and order reconstruction performance across JOL words for each participant. Average *G* across participants was 0.27 (*SD* = 0.27, 95% CI [.153, .377]), which was significantly greater than 0, $t(24) = 4.87$, $p < .001$, Cohen's $d = 0.97$.

Experiment 2

Participants provided item-by-item JOLs to 97.9% (*SD* = 2.0%) of words in the JOL blocks. The average JOL was 60.35 (*SD* = 13.07) and average *G* across participants was 0.09 (*SD* = 0.23, 95% CI [-.001, .180]), which is marginally greater than 0, $t(26) = 2.03$, $p = .05$, Cohen's $d = 0.39$.

Experiment 3

Participants provided item-by-item JOLs to 98.9% (*SD* = 1.2%) of words in the JOL blocks. The average JOL was 62.30 (*SD* = 17.10) and the average *G* across participants was 0.04 (*SD* = 0.35, 95% CI [-.109, .182]), which is not significantly different from 0, $t(24) = 0.51$, $p = .61$, Cohen's $d = 0.10$.

Experiment 4

Participants provided item-by-item JOLs to 98.1% (*SD* = 2.3%) of words in the JOL blocks. The average JOL was 53.65 (*SD* = 13.49). The relative accuracy of JOLs for words tested in the order reconstruction test was -0.02 (*SD* = 0.37, 95% CI [-.116, .084]), which is not significantly different from 0, $t(59) = -0.32$, $p = .75$, Cohen's $d = -0.04$. In the same way, the relative accuracy of JOLs for words tested in the forced choice recognition test (*M* = 0.04, *SD* = 0.39, 95% CI [-.059, .145]) was not significantly different from 0, $t(59) = -0.32$, $p = .75$, Cohen's $d = -0.04$.