Memory for Inter-item Relations is Reactively Disrupted by Metamemory Judgments

Wenbo Zhao[1], Yue Yin[1], Xiao Hu[2,3], David R. Shanks[4], Chunliang Yang[3,5], Liang Luo[1,2,5]

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

[2] Faculty of Psychology, Beijing Normal University, Beijing, China.

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

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

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

Author Note

The data contained in this project is publicly available at Open Science Framework (https://osf.io/9nm3w/). 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 Research Program Funds of the Collaborative Innovation Center of Assessment toward Basic Education Quality at Beijing Normal University (2021-01-132-BZK01), and the United Kingdom Economic and Social Research Council (ES/S014616/1).

Conflict of Interest

The authors declare no conflict of interest.

Compliance with Ethical Standards

The Ethics Committee at the Collaborative Innovation Center of Assessment for Basic Education Quality, BNU, approved the present study. All participants provided informed consent.

Abstract

Item memory (e.g., recall or recognition of specific items) can reactively change when metacognitively monitored via judgments of learning (JOLs). The current research explores whether memory for inter-item relations (e.g., semantic relations among list items) is reactively influenced by JOLs. Participants in Experiment 1 studied rhyming word pairs for which the target words were semantically-related category exemplars. The word pairs were presented in a category-blocked order during the study phase to increase the saliency of inter-target relations. The results showed that making JOLs had little influence on free recall of the target words, but disrupted semantic clustering, reflecting negative reactivity in memory for inter-item relations. The pre-registered Experiment 2 successfully replicated the main findings of Experiment 1. Experiment 3 found that, when the word pairs were presented in a random order (i.e., not blocked by category, leading to minimal semantic clustering at recall) during the study phase, free recall of target words was enhanced by JOLs, reflecting positive reactivity in item memory. The dissociable reactivity effects on item and inter-item relational memory found in these experiments support an item-relational memory account which hypothesizes that when a given strategy enhances item-specific processing, it concurrently diverts encoding resources away from inter-item relational processing.

Keywords: Judgments of learning; Reactivity; Item-relational account; Inter-item relational memory; Item memory

Over the last four decades, hundreds of studies have been conducted to determine how accurately people can metacognitively monitor their learning and memory status (for reviews, see Dunlosky & Tauber, 2016; Yang, Yu, et al., 2021). In these studies, participants are typically asked to make a judgment of learning (JOL) for each study item to predict the likelihood that they will remember it in a later memory assessment. The accuracy of JOLs is commonly quantified as the signed difference (i.e., absolute accuracy) or intra-individual correlations (i.e., relative accuracy) between JOLs and actual memory performance (Huff & Nietfeld, 2009; Lipowski et al., 2013; Rhodes, 2016; Schraw, 2009). However, an emerging body of research has demonstrated that the act of making JOLs can alter the very entity being judged (i.e., memory), demonstrating that memory performance can be *reactive* to observation and monitoring (Li et al., 2021; Mitchum et al., 2016; Myers et al., 2020; Soderstrom et al., 2015; Spellman & Bjork, 1992).

As an illustration, Soderstrom et al. (2015) instructed two (JOL vs. no-JOL) groups of participants to study a list of word pairs, composed of strongly (e.g., *doctor-nurse*) and weakly (e.g., *pond-frog*) related word pairs. Participants in the JOL group made a JOL while studying each word pair, whereas those in the no-JOL group did not. In a later cued recall test, the JOL group recalled significantly more strongly related pairs and numerically more weakly related ones than the no-JOL group, reflecting a positive reactivity effect on memory.

It should be noted that such reactivity effects tend to be moderated by material type, test format, and participant sample. Regarding material type, a recent meta-analysis conducted by Double et al. (2018) showed that making JOLs enhances retention of word lists and related word pairs, but fails to benefit recall of unrelated word pairs. In addition, Ariel et al. (2020)

observed no reactivity effect on learning of text passages. Regarding test format, Myers et al. (2020) found that reactivity is evident in recognition but not in free recall tests. Lastly, evidence suggests that the effect generalizes across children in elementary school (Zhao et al., 2022) and young adults (Witherby & Tauber, 2017), but not to older adults (Tauber & Witherby, 2019).

Previous studies have focused on the reactivity effect on item memory (i.e., recognition or recall of specific items; Myers et al., 2020; Senkova & Otani, 2021), but no research has been conducted to explore if memory for inter-item relations (e.g., memory of temporal or semantic relations among list items) is similarly reactive. The current study aims to fill this gap.

Exploring reactivity in inter-item relational memory is important for at least two reasons. First, it is well-known that both its contents (i.e., item memory) and organization (i.e., memory for inter-item relations) contribute to episodic memory (Tulving, 1972). Numerous laboratory and field studies have established that both aspects are important for successful retrieval (Diamond & Levine, 2020; Senkova & Otani, 2021; Yang, Zhao, et al., 2022). In addition, it has been well-documented that, in educational settings, both fact retention and knowledge integration play critical roles in successful learning (Rohrer et al., 2020). In the classroom, students not only need to remember specific knowledge units, but also to master the coherent connections among units, which are beneficial for both knowledge organization and transfer (Kubsch et al., 2020; Roelle et al., 2017). It is hence not only important to examine reactivity in item memory, but also in inter-item relational memory.

Second, exploring JOL-induced reactivity in inter-item relational memory may facilitate our understanding of why test format (free recall *vs*. recognition) moderates the reactivity effect. Previous studies observed a large positive reactivity effect on recognition for word lists (e.g., Cohen's d = 1.33 in Zhao et al., 2022; Cohen's d = 1.23 in Li et al., 2022), but the reactivity effect on free recall of word lists is relatively weak (e.g., Senkova & Otani, 2021) or even negligible (e.g., Tauber & Rhodes, 2012). More direct evidence for the moderating effect of test format comes from Myers et al. (2020). In this study, two (JOL and no-JOL) groups of participants were asked to study a list of word pairs, and then took recognition and free recall tests for the target words. The results showed positive reactivity in the recognition test but no detectable reactivity in the free recall test. Although previous research demonstrates consistently that reactivity varies across test formats, explanations about why this happens are lacking. In the General Discussion, we elaborate on how the reactivity effect on inter-item relational memory can help explain the moderating effect of test format.

Considering the importance of this question, the motivation of the current research is to explore the potential reactivity of inter-item relational memory under metamemory monitoring. The difference between intra- and inter-item relations should be highlighted. Intra-item relations refer to relations *within* a study item, such as the cue-target relation for a given word pair or a face-name pair (Mulligan & Peterson, 2015; Peterson & Mulligan, 2013). By contrast, inter-item relations refer to relations *among* list items, such as semantic or temporal relations among list words (Hunt & McDaniel, 1993; Jonker & MacLeod, 2015).

Intra- and inter-item relational memory can be functionally dissociated. Peterson and Mulligan (2013) provided a clear demonstration of this. They asked two (restudy *vs.* test)

groups of participants to study 36 rhyming word pairs (e.g., *tape-grape*, *wear-pear*). Critically, in addition to the intra-item rhyming relations between the cues and targets, there were also inter-item semantic relations among the targets. That is, the 36 targets were exemplars from six taxonomic categories (e.g., *fruits, occupations*). After initially studying all pairs, the restudy group restudied them in a category-by-category (blocked) order to increase the saliency of inter-target relations. By contrast, the test group took a cued recall test (e.g., *tape-* ___) on these pairs, also presented in a category blocked order. Finally, both groups freely recalled the targets (Experiment 1) or took a cued recall test on the studied word pairs (Experiment 2).

The results showed that the test group outperformed the restudy group in the cued recall test, reflecting a positive testing effect on intra-item relational memory. In contrast, the test group recalled fewer targets than the restudy group in the free recall test (a negative testing effect), and more importantly, targets were less semantically clustered in the test than in the restudy group, reflecting a negative testing effect on inter-item relational memory. Similar dissociations have also been observed in the generation effect (Nairne et al., 1991; Steffens & Erdfelder, 1998). Although active generation (e.g., *force-hor*__) enhances cued recall of rhyming word pairs relative to passive reading (e.g., *cheer-deer*), reflecting a positive generation effect on intra-item relational memory, it hinders free recall of the categorized targets, reflecting a negative generation effect on inter-item relational memory.

A possible explanation of the positive effects of testing and generation on cued recall and the negative effects on free recall is the *item-relational* account (Hunt & McDaniel, 1993; Hunt & Seta, 1984; Mulligan & Peterson, 2015; Peterson & Mulligan, 2013), which proposes

that when a given strategy enhances item-specific processing (e.g., encoding of cue-target rhyming relations), it concurrently diverts encoding resources away from inter-item relational processing (e.g., encoding of inter-target semantic relations), leading to a detrimental effect on the latter (Hunt & McDaniel, 1993; Steffens & Erdfelder, 1998).

To further test the validity of the item-relational account, Peterson and Mulligan (2013) conducted a further experiment (Experiment 3), in which the rhyming word pairs were presented in a pseudorandom order (i.e., no two consecutive pairs contained targets from the same category) to reduce the saliency of inter-target categorical relations. Peterson and Mulligan assumed that random presentation would eliminate (or at least reduce) the disadvantages of testing on inter-target relational processing (that is, both the test and restudy groups would now largely be unaware of the semantic relations among targets) and hence the enhancing effect of testing on item-specific processing would emerge in the free recall test. Their results confirmed this expectation by showing that free recall of targets was better in the test than in the restudy group when the rhyming pairs were presented in a random order during the review (restudy and practice test) phase.

Many recent studies employed word pairs (e.g., *pledge – promise*) as study stimuli to explore reactivity, in which reactivity was quantified as the difference in cued recall performance between JOL and no-JOL conditions (e.g., Mitchum et al., 2016; Soderstrom et al., 2015). These studies demonstrated that making JOLs enhances cued recall of related pairs, but has minimal or even a negative effect on cued recall of unrelated pairs. However, it is unknown whether memory for inter-item relations is reactive to the requirement of making JOLs. As discussed above, intra- and inter-item relational memory can be functionally

dissociated. Hence, the current study is primarily motivated to explore the reactivity of inter-item relational memory. To achieve this aim, the experiments reported here employed rhyming word pairs with targets being exemplars from a set of taxonomic categories (e.g., Mulligan & Peterson, 2015; Peterson & Mulligan, 2013).

According to the item-relational account (Hunt & McDaniel, 1993), we predicted that soliciting JOLs, similar to testing and generation, will enhance item-specific processing and concurrently disrupt inter-item relational processing. For instance, participants have to focus on encoding and analyzing the study item itself in order to search for diagnostic cues to make an appropriate JOL for it, leading to a positive reactivity effect on item memory (Senkova & Otani, 2021; Zhao et al., 2022). This enhanced item-specific processing is predicted to borrow limited encoding resources from inter-item relational processing, leading to a negative reactivity effect on that aspect of memory.

Before moving forward, it is worth noting that another influential theory of JOL reactivity – the *cue-strengthening theory* – can also readily explain positive reactivity in item memory. This theory was originally proposed by Soderstrom et al. (2015), who claimed that individuals need to search for relevant cues to make JOLs, and that this increases processing of study items, in turn leading to positive reactivity in item memory. Additionally, this theory hypothesizes that, when the cues used as the basis of JOLs (e.g., cue-target relations) are consistent with the cues used in the memory test (e.g., cued recall), the requirement of making JOLs will strengthen subsequent recall performance. We further elaborate on the cue-strengthening theory in the General Discussion section.

In the current research, reactivity in inter-item relational memory was quantified as the difference in semantic clustering between JOL and no-JOL conditions in a free recall test. Semantic clustering was measured via the adjusted ratio of clustering (ARC) method developed by Roenker et al. (1971), which estimates the likelihood that categorical items follow each other during free recall. An ARC score of zero indicates chance level clustering, with a positive score representing above-chance clustering, and a negative score indicating below-chance clustering (Chan et al., 2018; Senkova & Otani, 2012).

Roenker et al. (1971) demonstrated that other measures of semantic clustering, such as the modified ratio of repetition (MRR) and the clustering (C) index, are susceptible to irrelevant factors, such as number of categories recalled, total number of correctly-recalled items, and distribution of correct recall across different categories. By contrast, ARC scores are immune to these irrelevant factors. In addition, ARC scores are easy to interpret. Hence, following previous studies (Chan et al., 2018; Senkova & Otani, 2012), we employed the ARC measure of semantic clustering.

Experiment 1

Experiment 1 used rhyming word pairs with inter-target relations to assess the effects of making JOLs on inter-item relational memory. According to the item-relational account, we expected to observe negative reactivity in inter-item relational memory, represented as lower ARC scores in the JOL than in the no-JOL condition.

The item-relational account does not generate a clear prediction about reactivity in free recall performance. It is well-known that free recall relies on both item and inter-item relational memory (Hunt & McDaniel, 1993; Mulligan & Peterson, 2015). Negative reactivity

in inter-item relational memory and positive reactivity in item memory may cancel out each other, leading to minimal difference in free recall between the JOL and no-JOL conditions. Weakly positive or negative reactivity in free recall might be detected if reactivity in item memory is stronger or weaker than that in inter-item relational memory. Additionally, it is difficult to make a prediction based on previous findings, because they are largely inconsistent, with some showing positive reactivity in free recall (e.g., Zechmeister & Shaughnessy, 1980) and others showing no reactivity (e.g., Myers et al., 2020; Tauber & Rhodes, 2012). Hence, we had no *a priori* expectation about reactivity in free recall performance.

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in this and the subsequent experiments.

Method

Participants

Based on the magnitude (d = 0.709) of the negative testing effect on ARC scores reported by Peterson and Mulligan (2013), a power analysis was performed via G*Power (Faul et al., 2007), which indicated that 33 participants in each group were required to detect significant (2-tailed, α = .050) reactivity on ARC scores at 0.80 power. In the final sample, 70 participants (*M* age = 22.300, *SD* = 2.380; 49 female) were recruited from XXX (masked institution information), with 35 randomly allocated to each group. They were tested individually in a sound-proofed cubicle and received monetary compensation. The Ethics Committee at the XXX (masked institution information) approved all experiments.

Materials

The stimuli were 36 rhyming Chinese word pairs, which are listed in Appendix 1. The target for each pair was an exemplar from one of six taxonomic categories (i.e., four-footed animals, body parts, fruits, vegetables, occupations, and natural earth formations), selected from the category norms developed by Van Overschelde et al. (2004). A rhyming cue was selected from a Chinese word database (Cai & Byrsbert, 2010) to accompany each target, which itself was not a member of any of the six target categories.

Design and procedure

The experiment adopted a between-subjects design (JOL *vs*. no-JOL) and was composed of four phases: initial study, restudy, distraction and free recall test. Noteworthy is that, in a pilot study, free recall performance was close to floor when participants only studied each pair once. Such poor performance made it impossible to compute ARC scores for a majority of participants (for details of ARC score calculation, see Roenker et al., 1971). To avoid this floor effect, participants restudied the word pairs before the free recall test.

The instructions asked participants to study 36 word pairs twice in preparation for a later memory test. Participants in the JOL group were also informed that they would need to make a memory prediction for each pair while studying, and they did this on both cycles. By contrast, participants in the no-JOL group did not make memory predictions. Both groups were also informed that, for each pair, the word on the left was the cue and the one on the right was the target.

The procedure was adapted from previous JOL reactivity studies (e.g., Shi et al., 2022; Zhao et al., 2022). During the initial study phase, the 36 word pairs were presented one-by-one, blocked by taxonomic category (that is, the six pairs containing targets from the same

category were presented consecutively). For each participant, the presentation sequence of word pairs in each category and the category sequence were randomized.

For the no-JOL and JOL groups, each pair was presented on screen for 8 s. The only difference was that, for the JOL group, a 0 – 100 slider was simultaneously presented below each pair, and participants were instructed to drag and click the slider during the 8 s time-window to predict the likelihood that they would remember the on-screen pair in a later memory test (0 = *Sure I will not remember it*; 100 = *Sure I will remember it*).

After the initial study phase, both groups studied all word pairs a second time. The procedure during the restudy phase was the same as that during the initial study phase, including JOLs for the JOL group, except that the order of word pairs in each category and the category order were re-randomized.

Next, both groups completed a distractor task (solving arithmetic problems) for 5 min. Finally, they took a free recall test, in which they recalled as many target words as they could, in any order they preferred and typed their answers into a blank textbox. There was no time pressure and no feedback in the free recall test.

Overall, the only difference between the JOL and no-JOL groups was that participants in the JOL group, but not the no-JOL group, made concurrent JOLs for each pair during the initial study and restudy phases.

Results

The primary research question is the effect of making JOLs on free recall and ARC scores. The accuracy of item-by-item JOLs themselves (how well participants in the JOL group predicted their later item-by-item recall) is a subsidiary aspect, and hence those results

are reported in Appendix 2. In brief, the results showed that participants' JOLs in both the study and restudy phase were significantly correlated with subsequent test performance, suggesting that they were able to discriminate well learned items from less well studied ones.

We employed Bayes factors (BF_{10}) to measure whether a finding supports the alternative hypothesis relative to the null hypothesis. $BF_{10} > 3$ indicates that the results are in favor of the alternative over the null hypothesis, with $1 < BF_{10} < 3$ providing weak evidence supporting the alternative. By contrast, $BF_{10} < 0.33$ indicates that the observed results support the null over the alternative hypothesis, with $0.33 < BF_{10} < 1$ providing weak evidence supporting the null (Li et al., 2021). All Bayes factors were calculated via JASP 0.16 (https://jasp-stats.org), with all parameters set at their defaults.

There was no statistically detectable difference in free recall between the JOL (M = .390, SD = .149) and no-JOL groups (M = .409, SD = .178), difference = -.018 [-.097, .060], $t(68)$ = -0.465, p = .643, d = -0.111 (see Figure 1A), and the Bayesian evidence supports the absence of reactivity in free recall of targets, BF_{10} = 0.270. The number of inappropriately recalled cue words and incorrect recalls of unstudied words were low and did not statistically differ between groups (ps > .100; see Table 1).

For one participant in the no-JOL group, the ARC score was not computable because the participant recalled no more than one word from each of the six categories. Hence, this participant's data were excluded from the ARC analyses.

Of critical interest, there was strong evidence that ARC scores were greater in the no-JOL (M = .495, SD = .400) than in the JOL group (M = .177, SD = .471), difference = .318 [.107, .528], $t(67)$ = 3.016, p = .004, d = 0.726, BF_{10} = 10.510 (see Figure 1B), indicating that

the JOL group clustered target words less on the basis of semantic relations than the no-JOL group, and reflecting negative reactivity in inter-item relational memory. ARC scores in the JOL ($t(34) = 2.225$, $p = .033$, $d = 0.376$, $BF_{10} = 1.583$) and no-JOL ($t(33) = 7.219$, $p < .001$, $d = 1.238$, $BF_{10} > 100$) groups were greater than chance (i.e., 0), although the Bayesian evidence in the JOL group was relatively weak. These results reflect that both groups tended to recall targets according to their categorical relations.

Previous research established that semantic (categorical) relations serve to guide successful retrieval in free recall tests (Gruenewald & Lockhead, 1980; Hunt & Seta, 1984; Josephs et al., 2016). The same pattern was found here: There was strong evidence of a positive correlation between ARC scores and free recall performance in the JOL group, $r = .497$ [.196, .712], $p = .002$, $BF_{10} = 17.460$. Although the evidence was somewhat weaker in the no-JOL group, the pattern was similar, $r = .362$ [.028, .624], $p = .035$, $BF_{10} = 1.793$.

Discussion

Experiment 1 observed negative reactivity in inter-item relational memory (as reflected by poorer ARC scores in the JOL than in the no-JOL group), consistent with the main assumption of the item-relational account.

Another finding from Experiment 1 was that making JOLs had minimal effect on target free recall. A possible explanation is that a positive reactivity effect on item-specific processing (i.e., enhanced memory for the specific targets) and a negative effect on inter-item memory canceled each other out, leading to minimal reactive influence on free recall of targets. Indeed, Experiment 1 observed lower ARC scores in the JOL group, and ARC scores

were positively related to free recall performance. We further test this explanation in Experiment 3.

Experiment 2

Experiment 2 was pre-registered to replicate the main findings of Experiment 1, namely the negative effect of making metamemory judgments on inter-item relational memory. The pre-registration can be found at the Open Science Framework (https://osf.io/ajsrg). According to the results of Experiment 1 and the item-relational account, we expected to observe negative reactivity in ARC scores and no reactivity in free recall performance.

Method

Participants

According to the effect size of the reactivity effect on ARC scores observed in Experiment 1 (d = 0.726), a power analysis revealed that 31 participants in each group were required to detect a significant (2-tailed, α = .050) reactivity effect on ARC scores at 0.80 power. Given that ARC scores might not be computable for some participants, we decided to conservatively increase the sample size to 35 participants in each group to compensate for such attrition. In the final sample, 71 participants (*M* age = 20.620 years, *SD* = 2.486; 55 females) were recruited from XXX (masked institution information), with 35 randomly allocated to the JOL group and 36 to the no-JOL group. They were tested individually in a sound-proofed cubicle and received monetary compensation.

Materials, design and procedure

The materials, experimental design, and procedure were identical to those in Experiment 1.

Results

There were no deviations from the pre-registration.[1] Replicating Experiment 1, there was minimal difference in free recall between the JOL (M = .387, SD = .194) and no-JOL groups (M = .443, SD = .227), difference = -.056 [-.156, .045], $t(69) = -1.107$, $p = .272$, $d = -0.263$, $BF_{10} = 0.413$ (see Figure 1C). Although the Bayesian result only weakly supports the null hypothesis, the result pattern was similar to that observed in Experiment 1. The number of unstudied words incorrectly recalled did not differ significantly between groups ($p = .122$; see Table 1), whereas the number of cue word intrusions was significantly greater in the JOL than in the no-JOL group, $t(46.180) = 2.666$, $p = .011$, $d = 0.636$, $BF_{10} = 5.026$ (see Table 1).

ARC scores were not computable for 4 participants in the JOL group, and hence their data were excluded from the following analyses. Replicating Experiment 1, ARC scores were greater in the no-JOL (M = .381, SD = .454) than in the JOL group (M = .111, SD = .353), difference = .270 [.069, .471], $t(65) = 2.685$, $p = .009$, $d = 0.658$, $BF_{10} = 4.951$ (see Figure 1D), and the Bayesian analysis supports the existence of negative reactivity in ARC scores.

There was strong evidence that ARC scores were greater than chance in the no-JOL group, $t(35) = 5.031$, $p < .001$, $d = 0.839$, $BF_{10} > 100$. By contrast, there was insufficient evidence to conclude whether the equivalent scores were different from chance in the JOL group, $t(30) = 1.744$, $p = .091$, $d = 0.313$, $BF_{10} = 0.738$, although the pattern was consistent with that observed in the JOL group of Experiment 1. There was strong evidence that ARC scores positively correlated with free recall performance in both the JOL, r = .645

[1] In the pre-registration, we chose an option provided by OSF ("*there is no analysis difference between original experiment and replication*") to describe the data-analysis plan. Therefore, the pre-registration did not explain the data-analysis methods in detail (such as using a Bayesian independent sample *t* test to compare ARC scores between groups) but these were pre-planned to be identical to those in Experiment 1.

[.377, .813], $p < .001$, $BF_{10} > 100$, and no-JOL, $r = .544$ [.263, .741], $p < .001$, $BF_{10} = 59.609$, groups, reconfirming that semantic clustering is beneficial for free recall.

Discussion

The pre-registered Experiment 2 successfully replicated the negative reactivity effect in inter-item relational memory found in Experiment 1. In addition, semantic clustering scores (i.e., ARC scores) strongly predicted free recall performance.

Experiment 3

Experiments 1 and 2 corroborated the hypothesis of the item-relational account by showing negative reactivity in memory for inter-target semantic relations, as measured by clustering scores. Experiment 3 aimed to test another assumption of the item-relational account: that is, making JOLs reactively enhances item-specific processing. Following Peterson and Mulligan (2013, Experiment 3), Experiment 3 presented the rhyming word pairs in a pseudorandom order (that is, no two consecutive pairs contained targets from the same taxonomic category), rather than blocked by taxonomic category as in Experiments 1 and 2.

Randomized presentation was expected to eliminate (or at least reduce) the relational processing disadvantages of making JOLs because both the JOL and no-JOL groups would be less likely to attend to the categorical relations among targets when they are presented in a random order (Peterson & Mulligan, 2013, Experiment 3). Consequently, after eliminating (or reducing) this disadvantage, the benefit of metamemory monitoring on item-specific processing should be revealed. Therefore, according to the item-relational account, we expected to observe positive reactivity in free recall of target words in Experiment 3. This expectation was also partially drawn from the well-established phenomenon that both item-

specific and inter-item relational processing contribute importantly to free recall (McDaniel & Bugg, 2008).

Method

Participants

Following Experiment 2, the pre-planned sample size was 35 participants in each group. According to the effect size observed in Peterson and Mulligan's (2013) Experiment 3 (d = 0.765), such a sample size had power of 0.88 to observe significant (2-tailed, α = .050) positive reactivity in target free recall. Due to over-recruitment, the final sample included 80 participants (*M* age = 20.113, *SD* = 1.243; 58 females), recruited from XXX (masked institution information), with 40 randomly allocated to each group.

Materials, design and procedure

The materials, experimental design, and procedure were identical to those in Experiments 1 and 2, with one exception. During the initial study and restudy phases, the word pairs were presented in a pseudorandom order: no two consecutive trials included targets from the same taxonomic category.

Results

In contrast to the findings with taxonomically-organized words, the JOL group (*M* = .400, *SD* = .163) correctly recalled more targets than the no-JOL group (*M* = .281, *SD* = .136), difference = .119 [.052, .185], $t(78) = 3.542$, $p < .001$, $d = 0.792$, $BF_{10} = 42.504$ (see Figure 1E). These results provide strong evidence supporting the existence of positive reactivity in free recall of targets when the word pairs are shown in pseudorandom order.

Incorrect recall of cue words and unstudied words did not differ between groups (*p*s > .200; see Table 1).

For 5 participants in the no-JOL group, ARC scores were not computable, and their data were excluded. In neither the JOL, M = .022, SD = .347, $t(39)$ = .404, p = .688, d = 0.064, BF_{10} = 0.184, nor the no-JOL group, M = .007, SD = .387, $t(34)$ = .102, p = .919, d = 0.017, BF_{10} = 0.182, were ARC scores statistically different from chance, and the Bayesian evidence clearly supports the null hypothesis (for related findings, see Peterson & Mulligan, 2013). In addition, there was no statistically detectable difference in ARC scores between groups, difference = .015 [-.153, .184], $t(73)$ = 0.183, p = .856, d = 0.042, BF_{10} = 0.243 (see Figure 1F). There was minimal correlation between ARC scores and free recall performance in either the JOL, r = .085 [-.232, .387], p = .600, BF_{10} = 0.225, or no-JOL, r = .156 [-.187, .465], p = .371, BF_{10} = 0.309, groups. These results reveal that the randomized presentation manipulation was successful: we infer that neither the JOL nor the no-JOL group attended to the categorical structure of the targets and that targets were not recalled on the basis of inter-target categorical relations.

Discussion

Consistent with the findings from Peterson and Mulligan (2013, Experiment 3), Experiment 3 observed that, when the negative effect of making JOLs on inter-item relational processing was eliminated, the benefit of making JOLs on item-specific processing was revealed, as reflected by positive reactivity in free recall of targets. These findings are consistent with the item-relational account's proposal that making JOLs facilitates item-specific processing.

General Discussion

The present study examined JOL-induced reactivity in inter-item relational memory. Following Peterson and Mulligan (2013), the principal stimuli employed here were rhyming word pairs with categorical (specifically, taxonomic) relations among targets. Experiment 1 found that making JOLs disrupted semantic clustering during free recall of targets, reflecting negative reactivity in inter-item relational memory. The pre-registered Experiment 2 successfully replicated this negative effect on ARC scores. These findings are consistent with the main proposal of the item-relational account that making concurrent JOLs disrupts inter-item relational processing (Hunt & McDaniel, 1993; Mulligan & Peterson, 2015; Peterson & Mulligan, 2013).

Both Experiments 1 and 2 observed minimal effect of making JOLs on target free recall. A tempting inference is that these findings are inconsistent with the item-relational account's proposal that soliciting JOLs enhances item-specific processing, because making JOLs did not enhance recall of the specific items (i.e., the targets). However, such an inference would be mistaken. Indeed, we claim that the minimal reactivity effect on free recall is wholly consistent with the item-relational account. If item-specific and inter-item relational processing are in competition for limited processing resources, then enhancement of one would be at the cost of the other, and vice versa. Hence, making JOLs might enhance memory for the specific targets (Senkova & Otani, 2021) while concurrently interfering with the encoding of inter-item categorical relations. The positive reactivity effect on item memory, and the negative effect on inter-item relational memory may therefore have counterbalanced

each other, leading to little overall reactive influence on target free recall, as observed in Experiments 1 and 2.

Direct evidence supporting this explanation came from Experiment 3, in which the rhyming pairs were presented in a pseudorandom order (rather than blocked by category) to reduce the saliency of inter-target relations (Mulligan & Peterson, 2015). By eliminating the costs of making JOLs on inter-item relational processing (as reflected by near-zero ARC scores and minimal difference in semantic clustering between groups), positive reactivity in item memory was now detectable, as reflected by superior free recall of targets in the JOL than in the no-JOL group. Overall, the findings are consistent with the item-relational account and provide evidence of dissociable monitoring-induced reactivity in item and inter-item relational memory.

The impairment effect of making JOLs on inter-item relational memory can be utilized to explain why previous studies observed that reactivity is weaker in free recall than in recognition tests (see the Introduction). It is well-known that free recall performance relies on both item and inter-item relational memory. For instance, inter-item relations (e.g., temporal or semantic relations) can be used to guide output order in free recall tests, and superior inter-item relational memory is typically associated with superior free recall performance, as observed in Experiments 1 and 2 (for related findings, see Hunt & McDaniel, 1993; McDaniel & Bugg, 2008; Peterson & Mulligan, 2013). In contrast to free recall tests, in recognition tests, test items are generally presented in random order (which means that participants cannot control output order in recognition tests), and hence recognition performance relies less on inter-item relational memory. The suppressive effect of making JOLs on inter-item relational

memory may partially cancel out the enhancing effect of making JOLs on item memory, leading to weaker reactivity in free recall than in recognition tests. The current study supports this explanation by showing that positive reactivity in free recall performance emerges when the suppressive effect of making JOLs on inter-item relational memory is rendered irrelevant.

It is worth noting that, in the current research, there was a mismatch between the demands of the monitoring requirement and the final test. That is, participants made JOLs to predict the likelihood of remembering the rhyming word pairs in a later test, whereas in the final test they were instructed to freely recall the target words (rather than to recall the target words when prompted with the cue words). According to the cue-strengthening theory (Soderstrom et al., 2015), such a mismatch between the JOL and test demands should reduce (or even eliminate) the enhancing effect of making JOLs on recall performance. Hence, the cue-strengthening theory can readily explain the null reactivity effect on free recall of targets observed in Experiments 1 and 2.

The cue-strengthening theory can also explain the positive reactivity effect on free recall of targets observed in Experiment 3. For instance, making JOLs might have strengthened the rhyming relations between the cues and targets (Soderstrom et al., 2015), and during the free recall test, cue words might act as self-generated cues to facilitate recall of targets. Consistent with this explanation, Experiments 1-3 observed either significantly or numerically superior recall (i.e., greater intrusions) of cue words in the free recall tests (see Table 1), reflecting stronger cue-target relational memory in the JOL group.

Even though the cue-strengthening theory can explain free recall results observed in each experiment, it has difficulty explaining why the positive reactivity effect on free recall of

targets appeared in Experiment 3 but disappeared in Experiments 1 and 2 because all three experiments involved a mismatch between the demands of JOLs and the final test. Furthermore, it cannot explain why making JOLs reactively disrupted ARC scores in Experiments 1 and 2. In contrast to the cue-strengthening theory, the item-relational account provides a more comprehensive framework to account for all findings observed here.

Nevertheless, we do not conclude that the observed findings run counter to the cue-strengthening theory. Instead, the free recall results are somewhat in line with its theoretical explanation. The cognitive mechanisms proposed by the cue-strengthening theory and the item-relational account might jointly contribute to the reactivity effects documented here.

Besides the theoretical implications discussed above, the present findings also bring some practical implications. Both item and inter-item memory are critical for successful learning and retrieval (Hunt & McDaniel, 1993; McDaniel et al., 2016; McDaniel & Einstein, 1989; Smith & Hunt, 2000). In educational settings, students not only need to memorize specific knowledge concepts, but also have to construct coherent knowledge networks to structurally organize them. Furthermore, related concepts or topics in textbooks are typically structured according to their inherent similarities or relations (Rohrer et al., 2020), which is expected to facilitate knowledge organization and integration and benefit text comprehension. Experiments 1 and 2 jointly indicate that making JOLs disrupted encoding of relations among items, suggesting that asking students to monitor their learning by providing metacognitive judgments in the classroom may hinder knowledge integration.

Instructors should bear in mind these reactivity effect on inter-item relational memory when their courses require students to construct coherent knowledge networks. In such

scenarios, instructors may consider employing metamemory monitoring to facilitate students' learning of specific knowledge concepts. Meanwhile, they should also utilize other relational processing strategies (e.g., concept-mapping) to facilitate students' processing of inter-concept relations and to offset the harmful effect of JOLs on inter-item relational memory. Of course, it should be highlighted that the study materials employed here (i.e., word pairs) are not representative of real educational materials (e.g., text passages). Future research could profitably examine reactivity in memory for inter-item relations by using more realistic educational materials (such as statistical concepts with inter-concept relations).

Experiment 3 found positive reactivity in free recall of target words. However, intriguingly, Myers et al.'s (2020) Experiment 2 found no reactivity in a comparable free recall test. There are many divergences in experimental design between our Experiment 3 and Myers et al.'s (2020) Experiment 2 which might explain the divergent reactivity findings between these two experiments. For instance, in Myers et al.'s (2020) Experiment 2, participants only made JOLs during the restudy phase but not during the initial study phase, whereas in our Experiment 3 JOLs were elicited during both the initial study and restudy phases. The enhancing effect of making JOLs on item memory should accordingly be stronger in our Experiment 3 than in Myers et al.'s (2020) Experiment 2.

Another key difference concerns the materials employed. The stimuli in Myers et al.'s (2020) Experiment 2 were a mixed list of related and unrelated word pairs while those in our Experiment 3 were a pure list of rhyming word pairs with inter-target semantic relations. Previous studies showed that reactivity is moderated by material type. For instance, Mitchum et al. (2016) found that reactivity in memory for a mixed list of related and unrelated pairs

was different from that found in memory for pure lists of word pairs. This difference in material type might explain why our Experiment 3 and Myers et al.'s (2020) Experiment 2 obtained inconsistent reactivity results. In line with our Experiment 3, Senkova and Otani (2021) also observed positive reactivity in free recall of categorized words presented in random order during the study phase.

Overall, there are many differences in experimental design between our Experiment 3 and Myers et al.'s (2020) Experiment 2 and it is difficult to speculate which difference(s) causes the divergent reactivity findings. Future research on this research question is called for.

Concluding Remarks

Concurrent metamemory monitoring via JOLs enhances item memory but simultaneously disrupts memory for inter-item relations. The item-relational account provides a viable account for the dissociable reactivity of item and inter-item relational memory.

References

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

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

Chan, J. C. K., Manley, K. D., Davis, S. D., & Szpunar, K. K. (2018). Testing potentiates new learning across a retention interval and a lag: A strategy change perspective. *Journal of Memory and Language, 102*, 83–96. doi:10.1016/j.jml.2018.05.007

Daniels, K. A., Toth, J. P., & Hertzog, C. (2009). Aging and recollection in the accuracy of judgments of learning. *Psychology and Aging, 24*, 494–500. Doi:10.1037/a0015269

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.

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

Gruenewald, P. J., & Lockhead, G. R. (1980). The free recall of category examples. *Journal*

of Experimental Psychology: Human Learning and Memory, 6, 225–240. Doi:10.1037/0278-7393.6.3.225

Huff, J. D., & Nietfeld, J. L. (2009). Using strategy instruction and confidence judgments to improve metacognitive monitoring. *Metacognition and Learning, 4*, 161–176. Doi:10.1007/s11409-009-9042-8

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

Hunt, R. R., & Seta, C. E. (1984). Category size effects in recall: The roles of relational and individual item information. *Journal of Experimental Psychology: Learning, Memory and Cognition, 10*, 454–464. Doi:10.1037/0278-7393.10.3.454

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

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

Josephs, E. L., Draschkow, D., Wolfe, J. M., & Vo, M. L. H. (2016). Gist in time: Scene semantics and structure enhance recall of searched objects. *Acta Psychologica, 169*, 100–108. Doi:10.1016/j.actpsy.2016.05.013

Kubsch, M., Touitou, I., Nordine, J., Fortus, D., Neumann, K., & Krajcik, J. (2020). Transferring knowledge in a knowledge-in-use task—Investigating the role of knowledge organization. *Education Sciences, 10*, 20. Doi:10.3390/educsci10010020

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

Lipowski, S. L., Merriman, W. E., & Dunlosky, J. (2013). Preschoolers can make highly accurate judgments of learning. *Developmental Psychology, 49*, 1505–1516. Doi:10.1037/a0030614

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

McDaniel, M. A., Cahill, M. J., & Bugg, J. M. (2016). The curious case of orthographic distinctiveness: Disruption of categorical processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*, 104–113. doi:10.1037/xlm0000160

McDaniel, M. A., & Einstein, G. O. (1989). Material-appropriate processing: A contextualist approach to reading and studying strategies. *Educational Psychology Review, 1*, 113–145. Doi:10.1007/BF01326639

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., & 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

Rhodes, M. G. (2016). Judgments of learning: Methods, data, and theory. In J. Dunlosky & S. K. Tauber (Eds.), *The Oxford handbook of metamemory* (pp. 65-80). Oxford University Press.

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

Roelle, J., Hiller, S., Berthold, K., & Rumann, S. (2017). Example-based learning: The benefits of prompting organization before providing examples. *Learning and Instruction, 49*, 1–12. doi:10.1016/j.learninstruc.2016.11.012

Roenker, D. L., Thompson, C. P., & Brown, S. C. (1971). Comparison of measures for the estimation of clustering in free recall. *Psychological Bulletin, 76*, 45–48.

Doi:10.1037/h0031355

Rohrer, D., Dedrick, R. F., & Hartwig, M. K. (2020). The scarcity of interleaved practice in mathematics textbooks. *Educational Psychology Review, 32*, 873–883. Doi:10.1007/s10648-020-09516-2

Schneider, W., Visé, M., Lockl, K., & Nelson, T. O. (2000). Developmental trends in children's memory monitoring: Evidence from a judgment-of-learning task. *Cognitive Development, 15*, 115–134. Doi:10.1016/S0885-2014(00)00024-1

Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and Learning, 4*, 33–45. Doi:10.1007/s11409-008-9031-3

Senkova, O., & Otani, H. (2012). Category clustering calculator for free recall. *Advances in Cognitive Psychology, 8*, 292–295. doi:10.2478/v10053-008-0124-y

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

Shi, A., Xu, C., Zhao, W., Shanks, D. R., Hu, X., Luo, L., & Yang, C. (2022). Judgments of learning reactively facilitate visual memory by enhancing learning engagement. *Psychonomic Bulletin & Review*, *Advance online publication*. doi:10.3758/s13423-022-02174-1

Smith, R. E., & Hunt, R. R. (2000). The effects of distinctiveness require reinstatement of organization: The importance of intentional memory instructions. *Journal of Memory and Language, 43*, 431–446. Doi:10.1006/jmla.2000.2707

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

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

Steffens, M. C., & Erdfelder, E. (1998). Determinants of positive and negative generation effects in free recall. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology, 51*, 705–733. Doi:10.1080/027249898391350

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

Tulving, E. (1972). 12. Episodic and Semantic Memory. In E. Tulving & W. Donaldson (Eds.), *Organization of memory* (pp. 381–403). New York: Academic Press.

Van Overschelde, J. P., Rawson, K. A., & Dunlosky, J. (2004). Category norms: An updated and expanded version of the Battig and Montague (1969) norms. *Journal of Memory and Language, 50*, 289–335. doi:10.1016/j.jml.2003.10.003

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., 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

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. (2022). 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. *M* (*SD*) number of incorrectly recalled cue words and unstudied words in Experiments 1-3

		JOL	no-JOL
Experiment 1			
	Cue words	0.743 (1.172)	0.543 (0.780)
	Unstudied words	0.457 (0.611)	0.600 (1.143)
Experiment 2			
	Cue words	1.029 (1.361)	0.361 (0.593)
	Unstudied words	0.857 (1.438)	0.444 (0.652)
Experiment 3			
	Cue words	1.125 (1.265)	0.825 (1.299)
	Unstudied words	0.425 (0.931)	0.500 (0.751)

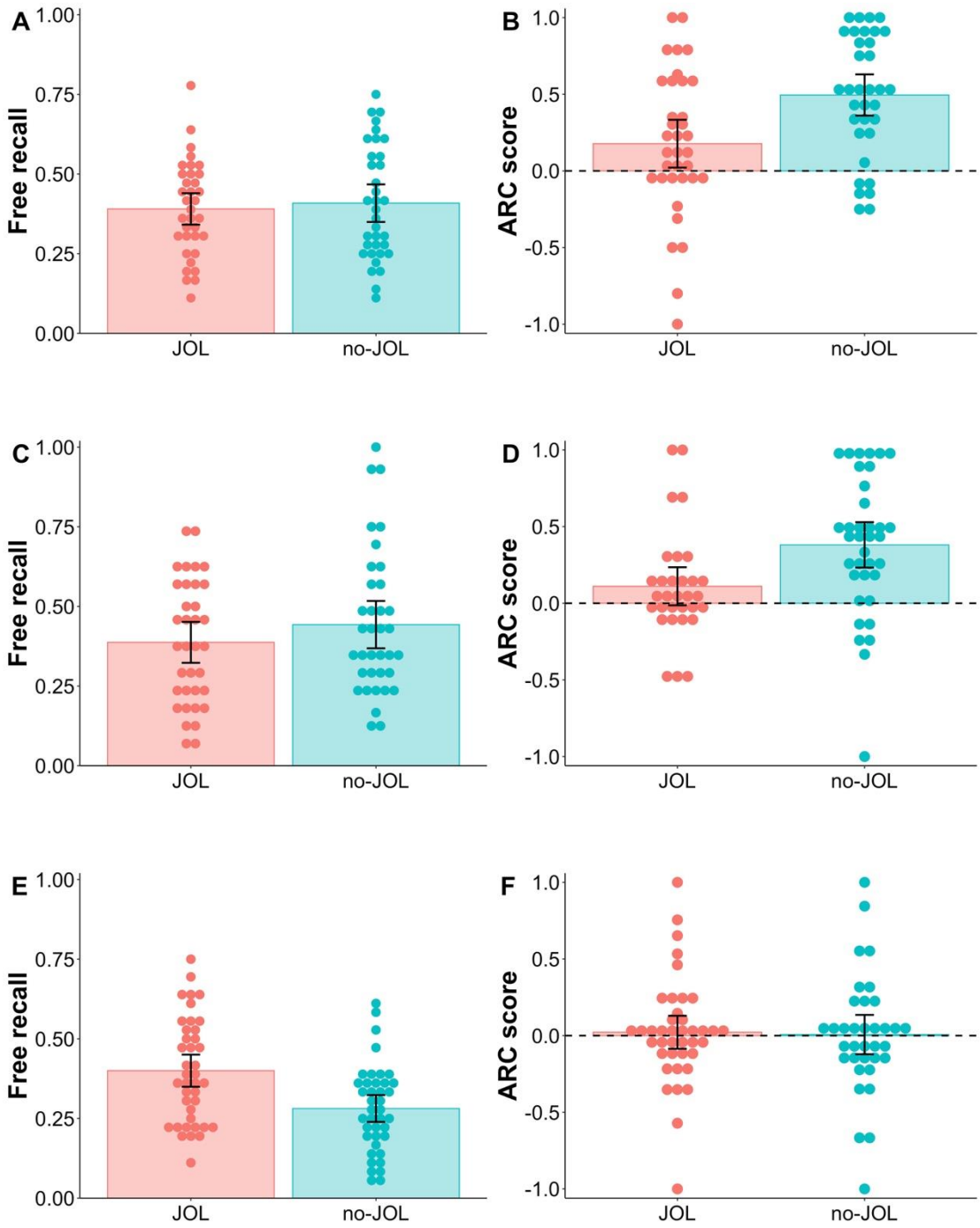

Figure 1. A, C, & E: Mean proportion of targets recalled (accuracy) as a function of study method (JOL *vs.* no-JOL) in Experiments 1-3, respectively. B, D, & F: Mean ARC score as a function of study method (JOL *vs.* no-JOL) in Experiments 1-3, respectively. Error bars represent 95% CI. Each dot represents one participant's score.

Appendix 1: Rhyming Chinese Word Pairs

Cue			Target		
Mandarin	English	International Phonetic Alphabet	Mandarin	English	International Phonetic Alphabet
			Four-footed animals		
小妞	Girl	ɕiɑu(214) niou(55)	奶牛	Cow	nai(214) niou(35)
画像	Portrait	xuA(51) ɕiɑŋ(51)	大象	Elephant	tA(51) ɕiɑŋ(51)
太阳	Sun	tʻai(51) iɑŋ(35)	山羊	Goat	ʂan(55) iɑŋ(35)
砝码	Weight	fA(214) mA(214)	斑马	Zebra	pan(55) mA(214)
公路	Highway	kuŋ(55) lu(51)	麋鹿	Elk	mi(35) lu(51)
玫瑰	Rose	mei(35) kuei(55)	海龟	Turtle	xai(214) kuei(55)
			Body parts		
小偷	Thief	ɕiɑu(214) tʻou(55)	额头	Forehead	ɤ(35) tʻou(35)
西藏	Tibet	ɕi(55) tsɑŋ(51)	心脏	Heart	ɕin(55) tsɑŋ(51)
乞丐	Beggar	tɕʻi(214) kai(51)	膝盖	Knee	ɕi(55) kai(51)
玛瑙	Agate	mA(214) nɑu(214)	大脑	Brain	tA(51) nɑu(214)
木棒	Stick	mu(51) pɑŋ(51)	肩膀	Shoulder	tɕian(55) pɑŋ(214)
甲醇	Methanol	tɕiA(214) tʂʻuən(35)	嘴唇	Lips	tsuei(214) tʂʻuən(35)
			Fruits		
包裹	Package	pɑu(55) kuo(214)	苹果	Apple	pʻiŋ(35) kuo(214)
文字	Text	uən(35) tsɿ(51)	桔子	Orange	tɕy(35) tsɿ(214)
手套	Glove	ʂou(214) tʻɑu(51)	樱桃	Cherry	iŋ(55) tʻɑu(35)
黄鹂	Oriole	xuɑŋ(35) li(35)	雪梨	Pear	ɕyɛ(214) li(35)
话梅	Prune	xuA(51) mei(35)	草莓	Strawberry	tsʻɑu(214) mei(35)
马褂	Jacket	mA(214) kuA(51)	西瓜	Watermelon	ɕi(55) kuA(55)
			Occupations		
野兽	Beast	iɛ(214) ʂou(51)	教授	Professor	tɕiɑu(51) ʂou(51)
专辑	Album	tʂuan(55) tɕi(35)	会计	Accountant	kʻuai(51) tɕi(51)
电源	Power	tian(51) yan(35)	警员	Police	tɕiŋ(214) yan(35)
杏仁	Almond	ɕiŋ(51) zən(35)	商人	Businessman	ʂɑŋ(55) zən(35)
果酱	Jam	kuo(214) tɕiɑŋ(51)	木匠	Carpenter	mu(51) tɕiɑŋ(51)
香味	Scent	ɕiɑŋ(55) uei(51)	守卫	Guard	ʂou(214) uei(51)
			Vegetables		
木材	Wood	mu(51) tsʻai(35)	菠菜	Spinach	po(55) tsʻai(51)

妻妾	Wife	tɕʻi(55) tɕʻiɛ(51)	番茄	Tomato	fan(55) tɕʻiɛ(35)
水痘	Chickenpox	ʂuei(214) tou(51)	豌豆	Pea	uan(55) tou(51)
秘密	Secret	mi(51) mi(51)	玉米	Corn	y(51) mi(214)
烟囱	Chimney	ian(55) tsʻuŋ(55)	洋葱	Onion	iɑŋ(35) tsʻuŋ(55)
广播	Broadcast	kuɑŋ(214) po(55)	萝卜	Turnip	luo(35) po(55)
			Natural earth formations		
丹麦	Denmark	tan(55) mai(51)	山脉	Mountain	ʂan(55) mai(51)
杨柳	Willow	iɑŋ(35) liou(214)	河流	River	xɤ(35) liou(35)
衬衫	Shirt	tʂʻən(51) ʂan(55)	火山	Volcano	xuo(214) ʂan(55)
琥珀	Amber	xu(214) pʻo(51)	湖泊	Lake	xu(35) pʻo(55)
年龄	Age	nian(35) liŋ(35)	丘陵	Hill	tɕʻiou(55) liŋ(35)
校园	Campus	ɕiɑu(51) yan(35)	高原	Plateau	kɑu(55) yan(35)

Appendix 2: JOL Accuracy

Experiment 1

During the initial study phase, participants made concurrent JOLs to 97.1% (*SD* = 2.97%) of word pairs in the JOL group. The average JOL was 57.908 (*SD* = 15.144). During the restudy phase, participants provided JOLs to 98.8% (*SD* = 1.94%) of word pairs and the average JOL was 63.579 (*SD* = 16.223).

A gamma (*G*) correlation was calculated to measure the relative accuracy of JOLs for each participant. Specifically, the target words were dummy coded (correctly recalled = 1; unrecalled = 0), and then, for each participant in the JOL group, we calculated *G* between JOLs and free recall performance during the initial study and restudy phases, respectively. For the JOLs made during the initial study phase, average *G* was 0.109 (*SD* = 0.248, 95% CI [.024, .194]), which was greater than chance (0), $t(34) = 2.606$, $p = .013$, Cohen's $d = 0.441$, $BF_{10} = 3.316$, indicating that participants were overall able to distinguish well-learned from less-well-learned items. For those made during the restudy phase, average *G* was 0.180 (*SD* = 0.247, 95% CI [.094, .266]), which was also greater than chance, $t(33) = 4.254$, $p < .001$, Cohen's $d = 0.730$, $BF_{10} > 100$.

Experiment 2

Participants provided concurrent JOLs to 98.5% (*SD* = 2.81%) and 99.0% (*SD* = 1.66%) of word pairs during the initial study and restudy phases, respectively. The average JOL was 48.200 (*SD* = 12.544) during the initial study phase and 53.412 (*SD* = 14.430) during the restudy phase. The average *G* for JOLs made during the initial study phase was 0.162 (*SD* = 0.231, 95% CI [.083, .242]), greater than 0, $t(34) = 4.153$, $p < .001$, Cohen's $d = 0.702$, $BF_{10} >$

100. The average *G* for JOLs made during the restudy phase was 0.179 (*SD* = 0.306, 95% CI [.074, .284]), greater than 0, $t(34) = 3.458$, $p = .001$, Cohen's $d = 0.584$, $BF_{10} = 22.404$.

Experiment 3

Participants provided concurrent JOLs to 96.0% (*SD* = 3.32%) and 98.5% (*SD* = 2.67%) of word pairs during the initial study and restudy phases, respectively. The average JOL was 53.666 (*SD* = 12.734) during the initial study phase and 56.921 (*SD* = 13.790) during the restudy phase. The average *G* for JOLs made during the initial phase was 0.159 (*SD* = 0.254, 95% CI [.078, .240]), greater than 0, $t(39) = 3.959$, $p < .001$, Cohen's $d = 0.626$, $BF_{10} = 87.456$. The average *G* for JOLs made during the restudy phase was 0.314 (*SD* = 0.202, 95% CI [.250, .379]), greater than 0, $t(39) = 9.842$, $p < .001$, Cohen's $d = 1.556$, $BF_{10} > 100$.