When Judging What You Know Changes What You Really Know: Soliciting Metamemory Judgments Reactively Enhances Children's Learning

Zhao, W., Li, B., Shanks, D. R., Zhao, W., Zheng, J., Hu, X., Su, N., Fan, T., Yin, Y., Luo, L., & Yang, C.

The data contained in this project and the pre-registration of Experiment 2 are publicly available at Open Science Framework (OSF) and the OSF links are provided on the cover page.

Recent studies established that making concurrent judgments of learning (JOLs) can significantly alter (typically enhance) memory itself – a *reactivity* effect. The current study recruited 190 Chinese children (M_{age} = 8.68 years; 101 female) in 2020 and 2021 to explore the reactivity effect on children's learning, its developmental trajectory and associated metacognitive awareness. The results showed that making JOLs significantly enhanced retention for students in Grades 1, 3, and 5, with Cohen's *d*s ranging from 0.40 to 1.33. Grade 5 students exhibited a larger reactivity effect than Grade 1 and 3 students. Children's metacognitive appreciation of the effect was weak. Firsthand experience of the reactivity effect, induced by taking a memory test, enhanced their awareness and calibrated their judgment accuracy.

Judgements of learning (JOLs; i.e., metacognitive estimates about the likelihood of remembering a studied item on a later memory test) play a fundamental role in human learning and memory, as people typically regulate their study strategies (e.g., when, what, and how to study) based on their JOLs (e.g., selecting items which are subjectively perceived as less-well studied to restudy; Finn, 2008; Yang, Potts, & Shanks, 2017b; Thiede, Anderson, & Therriault, 2003; Metcalfe, 2009). Hence, much research over the last half-century has been conducted to determine to what extent JOLs accurately reflect actual levels of learning and memory and what factors constrain JOL accuracy (for reviews, see Rhodes & Tauber, 2011; Dunlosky & Tauber, 2016; Yang, Yu, et al., 2021).

In previous studies, JOLs were typically made immediately following the studying of each item, and JOL accuracy was frequently quantified as the signed difference between JOLs and test performance (i.e., absolute accuracy) or intra-individual correlations between JOLs and test performance (i.e., relative accuracy). However, an emerging body of recent studies has highlighted a significant problem in this research by showing that making item-by-item JOLs can significantly change memory itself – a phenomenon termed the *reactivity* effect (Yang, Huang, et al., 2021; Double, Birney, & Walker, 2018). Below, we briefly describe previous research findings about the effect and outline the rationale of the current study.

Reactivity effects

About 30 years ago, Spellman and Bjork (1992) conjectured that metamemory judgments might fail to measure what they intend to assess because such judgments may reactively change the very thing being judged. This assumption has been tested in several recent studies, which showed that making item-by-item JOLs can retrospectively change memory itself (for reviews, see Yang, Huang, et al., 2021; Double et al., 2018). For instance, Soderstrom, Clark, Halamish, and Bjork (2015)

instructed two groups of participants to study strongly-related (e.g., *teacher–student*) and weakly-related (e.g., *pond –frog*) word pairs. In a JOL group, participants first studied a pair for 4s, and then studied the same pair for another 4s, during which they were instructed to make a JOL. By contrast, in a no-JOL group, participants studied each pair for 8s in total and did not make JOLs. Even though the total exposure duration of the word pairs was identical, the JOL group recalled significantly more strongly-related pairs and numerically more weakly-related pairs than the no-JOL group, reflecting a *positive reactivity* effect (that is, making JOLs facilitates learning) (for related findings, see Witherby & Tauber, 2017).

Some recent studies have documented that, in some situations, the reactivity effect can be *negative* (that is, making JOLs impairs retention). For instance, Mitchum, Kelley, and Fox (2016) had two groups (a JOL group *vs*. a no-JOL group) of participants study related (e.g., *computer – keyboard*) and unrelated (e.g., *apple – road*) word pairs. Their results showed that the JOL group remembered significantly fewer unrelated pairs than the no-JOL group (for related findings, see Tauber, Dunlosky, & Rawson, 2015), although the JOL group recalled numerically (but not significantly) more related pairs. A recent meta-analysis, which integrated results across 17 experiments, demonstrated that the reactivity effect is moderated by material type (Double et al., 2018). Specifically, Double et al. observed positive reactivity effects on learning of related word pairs (Hedges' $g = 0.323$) and word lists ($g = 0.384$), but there is minimal influence of making JOLs on learning of unrelated or a mixed list of related and unrelated word pairs.

Although the reactivity effect has been repeatedly investigated in recent studies, little research has been conducted to explore whether this effect generalizes to different populations. Indeed, most previous studies constrained their participants to college students (e.g., Mitchum et al., 2016; Witherby

& Tauber, 2017), but with one exception (Tauber & Witherby, 2019). In five experiments, Tauber and Witherby (2019) had college students and older adults study related word pairs, either making or not making JOLs. These five experiments consistently showed that positive reactivity failed to manifest in older adults, despite being robust in college students. Thus far, no research has been conducted to determine whether positive reactivity occurs in young children.

Metacognitive awareness of the reactivity effect

There are many strategies which can effectively boost learning efficiency, yet learners tend to underappreciate their benefits (e.g., Yang et al., 2017b; Potts & Shanks, 2014; Kornell & Bjork, 2008; Kirk-Johnson, Galla, & Fraundorf, 2019). For instance, testing (i.e., retrieval practice) can more effectively consolidate long-term retention than passive restudying – the *testing* effect – but people tend to believe that restudying is more beneficial than practice retrieval (Karpicke & Roediger, 2008; Yang, Luo, Vadillo, Yu, & Shanks, 2021; Rivers, 2020; but see Hartwig & Dunlosky, 2012; Kornell & Bjork, 2007; Morehead, Rhodes, & DeLozier, 2016). Even though spaced learning promotes inductive learning more effectively than massed learning – the *spacing* effect – people often erroneously believe that massed learning is more beneficial than spaced learning (Kornell & Bjork, 2008; Kornell, 2009).

Only one study has asked whether people are metacognitively aware of the reactivity effect, but the results were inconclusive (Yang, Huang, et al., 2021). Specifically, even though C. Yang, Huang, et al. observed that participants' metacognitive judgments about reactivity varied in the same direction as the effect they actually experienced, none of C. Yang, Huang, et al.'s experiments observed a significant difference in these judgments between JOL and no-JOL conditions. Hence, it remains unknown whether people can metacognitively appreciate the reactive influence of making JOLs on their learning. This research question is important because if people lack metacognitive

awareness of the positive reactivity effect, they might be reluctant to actively generate metacognitive judgments to aid encoding during self-regulated learning (Bjork, Dunlosky, & Kornell, 2013).

Rationale of the current study

As discussed above, there are many important questions about the reactivity effect that remain unknown. The first unexplored question is whether it generalizes to young children, and the second is what the developmental trajectory of the effect is. Exploring generalizability of the effect to young children's learning and investigating its developmental trajectory bear practical significance as it is important to determine whether making JOLs enhances or impairs young children's learning.

Another less-well studied aspect of the reactivity effect is the degree of metacognitive awareness associated with it. This research question is important because if generating JOLs reactively enhances children's learning but they lack awareness of this enhancing effect, they may be unlikely to actively employ this strategy during self-regulated learning (Tauber, Dunlosky, Rawson, Wahlheim, & Jacoby, 2013; Metcalfe, 2009). Hence, another aim of the current study is to determine whether children possess metacognitive awareness about the reactivity effect of making JOLs.

If a high proportion of children lack metacognitive appreciation of the effect (which is indeed what we find), how can we enhance their awareness? Previous studies have demonstrated that testing experience can act as a practical intervention to calibrate metacognition, because test performance provides diagnostic feedback to calibrate metacognitive judgments and to update metacognitive beliefs (Dunlosky & Hertzog, 2000; Dunlosky, Rawson, & McDonald, 2002; Kelemen, Winningham, & Weaver, 2007; Szpunar, Jing, & Schacter, 2014). Hence, the current study also investigates whether testing experience can improve children's awareness of the reactivity effect and enhance their

judgment accuracy (i.e., the alignment between a given individual's judgement about the reactivity effect and the actual effect of making JOLs on her learning; see below for details).

In total, the current study took both an explanatory and confirmatory (Experiment 2 was pre-registered) approach to investigate four important questions about the reactivity effect: (1) Does the effect generalize to young children? (2) What is the developmental trajectory of the effect? (3) Do elementary children possess metacognitive awareness of the effect? (4) If not, can test experience enhance their awareness of and the accuracy of their judgments about the effect?

Overview of the current study

To investigate these questions, Experiment 1 recruited 30 elementary students from each of Grades 1, 3, and 5. They studied 100 words, 50 with concurrent JOLs and 50 without, and their memory was assessed by a forced-choice recognition test. Experiment 1 chose Grade 1, 3, and 5 students because previous studies demonstrated that even 4-year-old children possess basic metamemory abilities (Wellman & Johnson, 1979).

The reactivity effect was quantified as the difference in test performance between JOL and no-JOL words, and the developmental trajectory of the effect was quantified by the interaction between age (grade) and reactivity. To measure metacognitive awareness of the effect, after studying all words but before testing, participants were prompted to report which kind of words (words with concurrent JOLs versus words without) they thought they would remember better. To evaluate the effect of test experience on metacognitive awareness, we re-asked participants to report which kind of words they thought they remembered better after they completed the recognition test. To investigate the effect of testing experience on the accuracy of metacognitive beliefs (i.e., the alignment between their beliefs and the actual reactivity effect they experienced), we compared the accuracy of their judgments about

the reactivity effect made before and after testing. Finally, we conducted a pre-registered experiment (Experiment 2) to conceptually replicate the main findings of Experiment 1.

Experiment 1

Method

Participants

Based on a pilot study (Cohen's d = 0.54 with 14 Grade 1 children as participants), we estimated that approximately 29 children in each grade group were required to observe a significant (2-tailed, α = .05) reactivity effect at 0.80 power. In the pilot study, college students were recruited to perform the same learning task. Recognition accuracies for both JOL and no-JOL words were over 90% in the forced-choice recognition test. Because of this ceiling effect (i.e., this learning task was too easy for adults), adult participants were not included in the current study.

In total, 30 Grade 1 students (*M* age = 6.27 years, *SD* = 0.45; 16 female), 30 Grade 3 students (*M* age = 8.53 years, *SD* = 0.57; 17 female), and 30 Grade 5 students (*M* age = 10.23 years, *SD* = 0.50; 18 female) were recruited from local elementary schools in Dandong City, China. All were native Chinese speakers, had normal or corrected-to-normal vision, and did not suffer from any neurological or psychiatric diseases (as reported by their caregivers). Participants received a box of modeling clay as compensation, and their parents provided informed consent. This experiment was conducted in May, 2020.

The study was approved by the Ethics Committee of the Faculty of Psychology, Beijing Normal University.

Materials

Two hundred and twenty Chinese words were selected from first-grade textbooks, with word frequency *M* = 75.26 (*SD* = 204.82) per million (Cai & Brysbaert, 2010) and number of strokes *M* = 14.90 (*SD* = 4.45). Before initiating the study, a first-grade Chinese teacher was invited to screen these words to ensure that they were suitable for Grade 1 students. Unfamiliar words identified by this teacher were replaced. In addition, to ensure that participants were familiar with each word, we played a recording of the word concurrently with its visual presentation (see below for details). For these 220 words, 20 of them were used for practice and the other 200 were used in the main experiment. In the experiment, 100 words were studied during the learning phase, and served as "old" items in the forced-choice recognition test, with the remaining 100 words as "new" items.

To avoid any item-selection effects, for each participant, the computer randomly divided the 100 words in the learning task into four lists, with two lists randomly assigned to the JOL condition and the other two to the no-JOL condition. In addition, for each participant, the presentation sequence of the words in each list and the list sequence were randomly decided by the computer. All stimuli were presented via the Matlab *Psychtoolbox* (Kleiner, Brainard, & Pelli, 2007) on a Microsoft Surface computer, which allowed touch-screen responses.

Experimental design and procedure

The experiment involved a 2 (Study method: JOL versus no-JOL) × 3 (Grade: 1 versus 3 versus 5) mixed design, with Study method as a within-subjects variable and Grade (age) as a between-subjects variable. Participants were informed that they would study four lists of words, with 25 words in each list, in preparation for a later memory test. For two lists, they would be asked to predict the likelihood of remembering each word in a later memory test, while they would not need to make such predictions for the other two lists of words. Importantly, they were also informed that they

should remember all words equally well regardless of whether they had to make memory predictions or not, because all of them would eventually be tested.

The task procedure was adapted from Soderstrom et al. (2015). Before the main experiment, participants completed a practice task to familiarize themselves with the procedure. Following practice, participants were asked if they understood the task requirements. If not, the experimenter re-explained the task and participants re-completed the practice phase. This cycle repeated until a given participant fully understood the task requirements. Then, the main experiment began.

Participants studied four lists of words, with 25 words studied in each list. Before each list, the computer informed participants whether or not they would need to make memory predictions for the subsequent list of words. In a no-JOL list, 25 words were presented one-by-one in a random order. Before the presentation of each word, a cross sign appeared at the center of the screen for 0.5s to mark the inter-stimulus interval. Then a word appeared on screen and its spoken version was played simultaneously. The spoken word lasted about 1s and the word appeared on screen for 5s in total. Then, the next trial started. This cycle repeated until the end of the list, with a new word studied in each cycle.

The procedure for the JOL lists was similar to that for the no-JOL lists, but when a word appeared on screen, a scale was simultaneously presented below it (see Figure 1). The scale was composed of five emoji faces, including a crying face (*Sure I will not remember it*), a slightly sad face (*Possibly I will not remember it*), a neutral face (*I have no idea*), a slightly happy face (*Possibly I will remember it*), and a very happy face (*Sure I will remember it*). Participants were asked to touch one of the emojis to make a memory prediction during the 5s time-window. If they did not successfully make a JOL during the required time-window, a message box appeared to remind them to make memory

predictions for the following words during the required time-window. Participants touched the message box to remove it and to trigger the next trial. If they successfully made a JOL before 5s expired, the word remained on screen for the remaining duration of the 5s to ensure that the total exposure for each word was 5s. In addition, the selected emoji was highlighted by a red frame to inform participants that they had made a prediction.

[Figure 1 goes here]

After participants studied all four lists, they were prompted to report which kind of words they thought they would remember better by touching one of three options, which were presented in the same order for all participants: (1) *I will remember the words for which I made memory predictions better than those for which I did not make predictions*; (2) *I will remember the words for which I made memory predictions equally well as those for which I did not make predictions*; (3) *I will remember the words for which I did not make memory predictions better than those for which I made predictions*. This question measures participants' metacognitive awareness (beliefs) about the reactivity effect.

After answering this question, participants played a jigsaw puzzle game for 10 minutes, which served as a distractor task. Then all participants completed a forced-choice recognition test. The 100 studied and 100 new words were randomly combined to form 100 pairs, with each pair consisting of an "old" and a "new" word. The pairs were presented one-by-one in a random order. Before presenting each pair, a cross sign was presented for 0.5s. Next, the computer randomly selected a word from the pair to present on screen, during which its spoken version was played. Then, the word disappeared and the other word was presented with its spoken version played. Then the two words were shown simultaneously on the screen, randomly allocated to the left and right positions. Participants were

instructed to touch one of the two words to indicate which was "old". When a recognition choice was made, the next test trial started automatically.

There was no time pressure and no feedback in the forced-choice recognition test. At the end of the test, participants were re-questioned about which kind of words they thought they remembered better, and the question and response choices were the same as those presented at the end of the learning task.

Results and discussion

Below we report test performance results (i.e., the reactivity effect) and then present the judgments (i.e., metacognitive awareness). Item-by-item JOLs (i.e., JOLs made for each word in the learning task) were not of substantive research interest, and hence are reported in the Supplemental Information (SI; available at https://osf.io/azje8/).

Test performance (the reactivity effect)

Recognition performance for both JOL and no-JOL words is depicted in Figure 2A. A mixed analysis of variance (ANOVA), with Study method as a within-subjects variable and Grade as a between-subjects variable, found a main effect of Study method, $F(1, 87) = 39.28$, $p < .001$, $\eta_p^2 = .311$. As shown in Figure 2A, JOL words were recognized more accurately than no-JOL words, indicating an overall positive reactivity effect on children's learning. Pre-planned paired *t*-tests showed that positive reactivity occurred in all three grades, Grade 1: difference in accuracy between JOL and no-JOL words = .05, 95% confidence interval = [.004, .096], $t(29) = 2.21$, $p = .04$, Cohen's $d = 0.40$; Grade 3: difference = .05 [.011, .078], $t(29) = 2.74$, $p = .01$, $d = 0.50$; Grade 5: difference = .12 [.079, .160], $t(29) = 6.08$, $p < .001$, $d = 1.11$.

[Figure 2 goes here]

The main effect of Grade was not statistically significant, $F(2, 87) = 1.06$, $p = .35$, $\eta_p^2 = .02$. Of critical importance, there was a significant interaction between Study method and Grade, $F(2, 87) = 4.47$, $p = .01$, $\eta_p^2 = .09$. As shown in Figure 2A, the main driver of this interaction is that the reactivity effect (represented by the difference in accuracy between JOL and no-JOL words) was larger in the Grade 5 group ($M = .12$, $SD = .11$) than in the Grade 1 group ($M = .05$, $SD = .12$), difference = .07 [.009, .129], $t(58) = 2.31$, $p = .02$, $d = 0.60$, and larger than in the Grade 3 group, ($M = .05$, $SD = .09$), difference = .08 [.024, .126], $t(58) = 2.92$, $p = .005$, $d = 0.76$. There was no significant difference in the magnitude of the reactivity effect between the Grade 1 and Grade 3 groups, difference = .005 [-.051, .061], $t(58) = 0.19$, $p = .85$, $d = 0.05$.

In addition to the group means, we also examined the proportions of children showing positive and negative reactivity. As illustrated in Figure 2B, across all three grades, a majority of children benefited from making JOLs. Specifically, for each grade, over half the students showed a positive reactivity effect, with a smaller proportion showing a negative effect. The proportion exhibiting a positive effect increased across Grades 1, 3, and 5, and the proportion showing a negative effect steadily declined. Although an overall Chi-square test found that these proportions did not significantly vary as a function of grade, $\chi^2(4) = 6.46$, $p = .17$, the proportion showing positive reactivity (i.e., the proportion of participants who correctly recognized a larger number of JOL words over no-JOL ones) did increase significantly, $\chi^2(2) = 6.30$, $p = .04$, while the proportion showing negative reactivity numerically decreased across grades, $\chi^2(2) = 4.32$, $p = .12$.

Overall, these results demonstrate positive reactivity in children across Grades 1, 3, and 5. Importantly, the effect tends to be small at Grades 1 and 3 but appreciably larger at Grade 5. In

addition, the proportions of participants who exhibited positive reactivity linearly increased across grades.

Metacognitive awareness prior to testing

This section explores whether elementary children metacognitively appreciated the beneficial effect of making JOLs before they completed the criterion test. According to participants' responses to the judgment question about which kind of words they thought they would remember better, presented prior to testing, they were classified into three categories: (1) JOL > no-JOL (i.e., participants who believed that JOL words would be remembered better than no-JOL ones), (2) JOL = no-JOL, and (3) JOL < no-JOL.

Figure 2C shows the proportions of participants in each category at each grade. Note that there were only 30 participants in each group, which were then subdivided amongst three categories, and such a small sample size did not permit Chi-square tests to analyze the trends. We therefore collapsed judgments across the three grades to increase statistical power. In addition, as clearly shown in Figure 2C, students in Grades 1, 3, and 5 showed very similar judgment patterns, and a Chi-square test found that Grade did not significantly affect participants' judgments, $\chi^2(4) = 0.73$, $p = .95$.

Across all three grades, 54.4% (49 out of 90) of participants believed JOL > no-JOL, which was significantly greater than the proportion believing JOL = no-JOL (13.3%), $\chi^2(1) = 32.14$, $p < .001$, and also greater than the proportion believing JOL < no-JOL (32.2%), $\chi^2(1) = 8.17$, $p = .004$. The proportion believing JOL < no-JOL was greater than the proportion believing JOL = no-JOL, $\chi^2(1) = 8.09$, $p = .004$.

In summary, these findings reveal that, prior to taking the recognition test, participants were somewhat aware of the positive reactivity effect, as reflected by the finding that a slight majority

(54.5%) of them believed JOL > no-JOL. Nonetheless, one third of the children believed that no-JOL words were more memorable than JOL ones.

Metacognitive awareness after testing

We now turn to metacognitive judgments made after testing. Again, participants were classified into three categories according to their responses. The proportion of participants in each category for each grade is depicted in Figure 2C. Once more, there was no systematic difference in judgments across grades, $\chi^2(4) = 2.00$, $p = .74$, and judgments were therefore collapsed across groups.

Across grades, 66.7% (60 out of 90) of participants believed JOL > no-JOL, which was substantially greater than the proportion believing JOL = no-JOL (16.7%), $\chi^2(1) = 44.25$, $p < .001$, and greater than the proportion believing JOL < no-JOL (16.7%), $\chi^2(1) = 44.25$, $p < .001$. There was no difference between the proportion believing JOL = no-JOL and the proportion believing JOL < no-JOL. Overall, these findings demonstrate that, after testing, participants were robustly aware of the positive reactivity effect.

Effect of test experience on metacognitive awareness

Even though participants were aware of the benefit of making JOLs both before and after testing, we can ask whether test experience enhanced their metacognitive awareness. Figure 2C shows the proportions believing JOL > no-JOL, JOL = no-JOL, and JOL < no-JOL as a function of judgment timing (before testing versus after testing). It is clear from the figure that across all grades, test experience caused the proportion believing JOL > no-JOL to increase and the proportion believing JOL < no-JOL to decrease.

With judgment results collapsed across groups, a Chi-square test showed that test experience affected participants' judgments, $\chi^2(2) = 5.90$, $p = .05$. Further analyses showed that test experience

numerically increased the proportion of participants who believed JOL > no-JOL (before testing: 54.4% versus after testing: 66.7%), $\chi^2(1) = 2.33$, $p = .13$, and it significantly decreased the proportion believing JOL < no-JOL (before testing: 32.2% versus after testing: 16.7%), $\chi^2(1) = 5.08$, $p = .02$. Test experience did not significantly affect the proportion believing JOL = no-JOL (before testing: 13.3% versus after testing: 16.7%), $\chi^2(1) = 0.16$, $p = .69$.

It is noteworthy that there were 21 children who not only believed JOL < no-JOL prior to testing but simultaneously showed positive reactivity in the recognition test, demonstrating an illusion of metacognition. For these poorly calibrated participants, testing caused 66.7% of them to appreciate the positive reactivity effect, leaving only 23.8% continuing erroneously to believe that the effect is negative.

In summary, the above findings show that taking a memory test tends to enhance metacognitive awareness of the positive reactivity effect. In particular, there was a halving of the number of children believing that no-JOL words were more memorable than JOL ones.

Accuracy of metacognitive beliefs about reactivity

Accuracy of judgments about the reactivity effect was quantified by the alignment between judgments about the effect and the actual effect participants experienced. Suppose that a given participant reported JOL > no-JOL in response to the pre- or post-test judgment question. If this participant actually recognized JOL words better than no-JOL ones in the recognition test (i.e., if the judgment and the actual reactivity effect aligned), her judgment would be labeled as correct (and scored 1). If she actually recognized fewer JOL words than no-JOL words or if JOL and no-JOL words were recognized equally well (i.e., if the judgment and the actual reactivity were misaligned),

her judgment would be labeled as incorrect (and scored 0). Scores were computed correspondingly for children who reported JOL = no-JOL and JOL < no-JOL.

As shown in Figure 2D, before testing, Grade had no significant influence on judgment accuracy, $\chi^2(2) = 0.09$, $p = .96$. Hence, to increase statistical power, the data were collapsed across grades. Judgment accuracy (.42; 38 out of 90) was not significantly different from chance (.33; recall that there were three response options), $\chi^2(1) = 1.16$, $p = .28$, implying that participants' judgment accuracy of the reactivity effect was quite poor.

Next, we measured the accuracy of post-test judgments. Again, as shown in Figure 2D, there was no significant difference in accuracy across grades, $\chi^2(2) = 0.28$, $p = .87$, and so the accuracy scores were collapsed. The data showed that, after testing, judgment accuracy (.60; 54 out of 90) was significantly better than chance, $\chi^2(1) = 11.81$, $p < .001$, and better than pre-test judgment accuracy (.42), $\chi^2(1) = 5.00$, $p = .03$. Hence, these findings confirm that test experience is an effective intervention to improve the accuracy of metacognitive judgments about the reactivity effect.

Another intriguing finding was that, for the 29 participants who reported JOL < no-JOL before taking the recognition test, their pre-test judgment accuracy (.17; 5 correct out of 29) was strikingly low. This poor accuracy mainly resulted from the fact that a majority (72.4%) of them actually experienced positive reactivity and only a small proportion (17.3%) experienced negative reactivity in the recognition test. Critically, after the recognition test, their judgment accuracy significantly improved to .69, $\chi^2(1) = 13.78$, $p < .001$, reconfirming that test experience improves judgment accuracy.

Experiment 2

Experiment 1 provided evidence of positive reactivity in children in Grades 1, 3 and 5, accompanied by relatively poor metacognitive awareness that was partially improved by taking a memory test. Experiment 2 was pre-registered to conceptually replicate the main findings from Experiment 1. The pre-registration is available at https://osf.io/7gcwp/registrations.

We made one noteworthy procedural change. It is possible that the emoji face scale employed in Experiment 1 affected participants' learning motivation and influenced their responses to the metacognitive judgment questions. Hence, another aim of Experiment 2 is to determine whether the findings documented in Experiment 1 survive when the emoji face scale is replaced by a conventional digit rating scale. This experiment was conducted in May, 2021.

Method

Participants

Given that, in Experiment 1, the results from children in Grades 1 and 3 did not systematically differ, the Grade 3 group was excluded in Experiment 2. Instead, we increased the sample size in the Grade 1 and 5 groups to increase statistical power. According to the effect size (η_p^2 = .084) of the interaction between reactivity and Grade (1 versus 5) observed in Experiment 1, 45 children in each group were required to detect a significant (2-tailed, α = .05) interaction at 0.80 power. To be more conservative, we pre-registered the goal to recruit 50 children in each group. Such a sample size permitted us to obtain a significant reactivity effect at 0.98 power in Grade 1 and 0.99 power in Grade 5.

Accordingly, 50 Grade 1 (*M* age = 6.53 years, *SD* = 0.69; 28 female) and 50 Grade 5 students (*M* age = 10.37 years, *SD* = 0.67; 22 female) were recruited from local elementary schools in Dandong City, China. All were native Chinese speakers, had normal or corrected-to-normal vision, and did not

suffer from any neurological or psychiatric diseases. They received a box of modeling clay as compensation, and their caregivers provided informed consent.

Materials, design and procedure

The materials, design, and procedure were identical to those in Experiment 1 but with three differences. The first was that Experiment 2 involved a 2 (Study method: JOL versus no-JOL) × 2 (Grade: 1 versus 5) mixed design. The second was that the emoji face scale was replaced by a digit scale because the emoji faces might have increased children's learning motivation in the JOL condition. Accordingly, in Experiment 2 participants were instructed to touch one of 5 digits (1 = *Sure I will not remember it*; 5 = *Sure I will remember it*) to report their JOLs. The third change was that, to avoid the effects of presentation order on responses to the metacognitive judgment questions, for each participant the computer randomly presented the three options (i.e., JOL > no-JOL, JOL = no-JOL, and JOL < no-JOL) in one of two orders: the same as in Experiment 1 or the reverse.

Results and discussion

Test performance and metacognitive awareness judgments are reported below. Item-by-item JOLs are reported in the SI.

Test performance (the reactivity effect)

A mixed ANOVA found a main effect of Study method, $F(1, 98) = 111.14$, $p < .001$, $\eta_p^2 = .531$ (see Figure 3A). Pre-planned paired *t*-tests showed that positive reactivity occurred in both Grade 1, difference = .05 [.034, .075], $t(49) = 5.30$, $p < .001$, $d = 0.75$, and Grade 5 children, difference = .11 [.084, .129], $t(49) = 9.44$, $p < .001$, $d = 1.33$.

[Figure 3 goes here]

The main effect of Grade was marginally significant, $F(1, 98) = 3.85$, $p = .053$, $\eta_p^2 = .04$. A mini random-effects meta-analysis, which integrated the Grade 1 and Grade 5 test performance across Experiments 1 and 2, found that Grade 5 significantly outperformed Grade 1, Hedges' $g = 0.37$ [0.061, 0.686], $p = .02$, and there was minimal heterogeneity across experiments, $Q(1) = 0.006$, $p = .94$.

Critically, the interaction between Study method and Grade was significant, $F(1, 98) = 11.62$, $p < .001$, $\eta_p^2 = .11$. The reactivity effect was larger in the Grade 5 group ($M = .11$, $SD = .08$) than in the Grade 1 group ($M = .05$, $SD = .07$), difference = .05 [.022, .082], $t(98) = 3.41$, $p < .001$, $d = 0.68$, confirming the developmental trajectory of reactivity.

Figure 3B depicts the proportions of children showing positive and negative reactivity. Although an overall Chi-square test found that these proportions did not significantly vary across grades, $\chi^2(2) = 4.61$, $p = .10$, the proportion showing positive reactivity marginally increased from Grade 1 to Grade 5, $\chi^2(1) = 3.06$, $p = .08$, while the proportion showing negative reactivity numerically decreased across grades, $\chi^2(1) = 0.75$, $p = .39$.

Overall, the above results conceptually replicate the main findings from Experiment 1 by showing that both Grade 1 and Grade 5 children benefited from making JOLs, and the enhancing effect was appreciably larger at Grade 5.

Metacognitive awareness prior to testing

According to participants' responses to the judgment question presented prior to testing, they were classified into three categories (see Figure 3C). Given that a Chi-square test showed that Grade did not significantly affect participants' judgments, $\chi^2(2) = 1.87$, $p = .39$, we collapsed judgments across grades to increase power. Overall, 50.0% (50 out of 100) of children believed JOL > no-JOL, significantly greater than the proportion believing JOL = no-JOL (26.0%), $\chi^2(1) = 11.27$, $p < .001$, and

the proportion believing JOL < no-JOL (24.0%), $\chi^2(1) = 13.41$, $p < .004$. The proportion believing JOL < no-JOL was not significantly different from the proportion believing JOL = no-JOL, $\chi^2(1) = 0.03$, $p = .870$.

In summary, these findings confirmed that, prior to taking the recognition test, participants were somewhat aware of the positive reactivity effect. Nonetheless, half of them did not realize the enhancing effect of making JOLs.

Metacognitive awareness after testing

According to responses to metacognitive judgments made after testing, participants were re-classified into three categories (see Figure 3C). There was no systematic difference in metacognitive awareness between grades, $\chi^2(2) = 1.09$, $p = .58$, and hence judgments were collapsed across grades. In total, 67.0% (67 out of 100) of children believed JOL > no-JOL, greater than the proportion believing JOL = no-JOL (15.0%), $\chi^2(1) = 53.76$, $p < .001$, and the proportion believing JOL < no-JOL (18.0%), $\chi^2(1) = 47.14$, $p < .001$. There was no significant difference between the proportion believing JOL = no-JOL and the proportion believing JOL < no-JOL, $\chi^2(1) = 0.15$, $p = .70$. Overall, these results replicate the finding that, after testing, participants were robustly aware of the positive reactivity effect.

Effect of test experience on metacognitive awareness

With judgment results collapsed across groups, a Chi-square test showed that test experience significantly affected participants' judgments, $\chi^2(2) = 6.28$, $p = .04$ (see Figure 3C). Further analyses showed that test experience significantly increased the proportion believing JOL > no-JOL (before testing: 50.0% versus after testing: 67.0%), $\chi^2(1) = 5.27$, $p = .02$. In addition, testing experience numerically decreased both the proportion believing JOL = no-JOL (before testing: 26.0% versus after

testing: 15.0%), $\chi^2(1) = 3.07$, $p = .08$, and the proportion believing JOL = no-JOL (before testing: 24.0% versus after testing: 18.0%), $\chi^2(1) = 0.75$, $p = .39$.

There were 20 children who believed JOL < no-JOL prior to testing but experienced positive (JOL > no-JOL) reactivity in the recognition test, exhibiting a metacognitive illusion. For these 20 children, testing caused 55% of them to appreciate the positive reactivity effect, leaving only 35% continuing erroneously to believe that reactivity is negative. In summary, these results replicate the finding that test experience boosted metacognitive awareness of positive reactivity.

Accuracy of metacognitive beliefs about reactivity

Following Experiment 1, judgment accuracy was coded according to the alignment between judgments about the reactivity effect and the actual effect participants experienced. Before testing, Grade had no significant influence on judgment accuracy, $\chi^2(1) = 0.16$, $p = .69$ (see Figure 3D). Therefore, accuracy data were collapsed across grades. Judgment accuracy (.44; 44 out of 100) was not significantly different from chance (.33), $\chi^2(1) = 1.16$, $p = .28$, implying poor judgment accuracy before testing.

Next, we assessed the accuracy of post-test judgments. As shown in Figure 3D, there was no significant difference in accuracy across grades, $\chi^2(2) = 0.17$, $p = .92$, and so the accuracy scores were collapsed. After testing, judgment accuracy (.64; 64 out of 100) was significantly greater than chance, $\chi^2(1) = 17.62$, $p < .001$. More importantly, post-test accuracy was also greater than pre-test accuracy (.44), $\chi^2(1) = 7.27$, $p = .007$, confirming that test experience effectively enhances judgment accuracy.

In summary, this pre-registered experiment confirmed the presence of positive reactivity in children in Grades 1 and 5, with the magnitude of reactivity increasing with age, accompanied by relatively poor metacognitive awareness that was partially alleviated by taking a memory test.

General Discussion

The current study is the first to investigate the reactive effect of making JOLs on children's learning and its developmental trend. The principal finding of Experiments 1 and 2 was that making JOLs significantly enhanced children's learning and the positive reactivity effect generalized to both young and older children. All previous reactivity studies presented stimuli in a unimodal form (e.g., word pairs presented visually; Mitchum et al., 2016; Yang, Huang, et al., 2021), which is somewhat unrepresentative of materials in real educational settings. For instance, lecture contents are orally delivered by teachers, accompanied with visual presentation (e.g., PowerPoint slides). The current study revealed positive reactivity when items were presented in a combined (visual + auditory) mode, thus further extending its generalizability to multi-modal learning.

Although the current experiments found that positive reactivity generalizes to multi-modal learning, it would be premature to recommend that students make concurrent metacognitive judgments in educational settings. Further research is required to clarify its boundary conditions and moderators. Indeed, several recent studies have sought to achieve this goal. For instance, research has observed positive reactivity effects on learning of word lists (Double et al., 2018), related word pairs (Soderstrom et al., 2015), and general knowledge facts (Yang, Huang, et al., 2021), whereas no reactivity is found on learning of texts (Ariel, Karpicke, Witherby, & Tauber, 2021), unrelated word pairs (Soderstrom et al., 2015; Dougherty, Robey, & Buttaccio, 2018), or mixed lists of related and unrelated word pairs (Double et al., 2018). Even though making JOLs facilitates children's (the current study) and young adults' learning (Witherby & Tauber, 2017), older adults benefit minimally from making JOLs (Tauber & Witherby, 2019). Myers, Rhodes, and Hausman (2020) also observed that reactivity is moderated by test format, being more positive in recognition than in free recall tests.

Overall, positive reactivity is limited to certain populations, study materials, and test formats – restrictions that would have to be taken into account in any translation into the classroom. It is also worth noting that, in the current study, all participants were recruited from Chinese elementary schools, and all stimuli were Chinese words. It is important for future research to test the generalizability of the documented findings to different languages and children in different countries.

The positive reactivity effect on children's learning may be accounted for by the *positive-reactivity* theory (Mitchum et al., 2016; Rivers, 2018; Dougherty, Scheck, Nelson, & Narens, 2005), which proposes that positive reactivity is primarily caused by the fact that making item-by-item JOLs encourages learners to adopt more effective encoding strategies and exert greater study effort. For instance, frequently asking children to make predictions about subsequent test performance might remind them that they will be tested later, and this test expectancy may in turn stimulate them to adopt more effective strategies (Sahakyan, Delaney, & Kelley, 2004) and expend greater study effort (Yang, Potts, & Shanks, 2017a). In addition, in the JOL condition, children have to sustain their engagement in the learning task in order to make an appropriate JOL for each study item. Enhanced learning engagement and refined study strategies in turn are likely produce a positive reactivity effect on children's learning.

Although the positive-reactivity theory accounts for the overall positive reactivity effect we observed, it has some difficulty in explaining its developmental trajectory. This theory assumes that the magnitude of the effect should be (at least partially) determined by how poor the study strategies are or how limited the learning effort is when providing JOLs is not required. If a learner devotes a large amount of effort in the no-JOL control condition, little room is left for JOL-making to enhance study effort in the experimental condition. Indeed, Tekin and Roediger (2020) recently found that the

enhancing effect of making JOLs on items processed shallowly (e.g., perceptual judgments) was larger than the effect on items processed deeply (e.g., semantic judgments).

Because it is well-known that young children are less able to sustain their attention and maintain their effort in a learning or cognitive task than older children (Wetzel, 2014; Klenberg, Korkman, & Lahti-Nuuttila, 2001; Ruff, Lawson, Parrinello, & Weissberg, 1990), there should be more opportunity for the requirement to make concurrent JOLs to enhance young children's study effort and to boost their memory performance. Equally, because young children are less able to employ effective (complex) encoding strategies when they are not explicitly or implicitly prompted to do so (Horn, Bayen, & Michalkiewicz, 2021; Schleepen & Jonkman, 2012; Schneider, 1986), there should be more opportunity for concurrent JOLs to refine their learning strategies. Therefore, the positive-reactivity theory predicts a negative interaction between age and reactivity, with younger children benefiting more from making concurrent JOLs than their older counterparts. However, the current study found the exact reverse pattern, namely a positive relation between age and reactivity.

The positive interaction between age and reactivity might result from larger dual-task costs, induced by frequent task-switching between encoding and monitoring in the JOL condition, for young children (Janes, Rivers, & Dunlosky, 2018). Janes et al. assumed that, when making item-by-item JOLs, individuals need to frequently switch their processing mode between encoding and monitoring, and frequent task-switching in turn leads to a detrimental effect of making JOLs on learning, which is especially true when the learning task is demanding.

In the current study, the learning and monitoring tasks were more challenging for young than for older children, as revealed by the findings that (1) test performance was poorer in the Grade 1 than in the Grade 5 group (see the meta-analytic results), and (2) Grade 1 children failed to make item-by-

item JOLs to a greater proportion of words than Grade 5 children (see the SI). Hence, dual-task costs should have been larger for young than for older children, leading to a smaller enhancing effect of making JOLs for young than for older children. Obviously, the dual-task costs explanation cannot explain why the reactivity effect was positive overall for young and older children.

The positive-reactivity theory and the dual-task costs explanation are not mutually exclusive, and these two theories can jointly account for the findings documented in the current study. For instance, according to the positive-reactivity theory, because children are insufficiently able to maintain their learning engagement in the no-JOL condition, instructing them to provide JOLs while studying might have enhanced their engagement in the JOL condition, leading to a positive reactivity effect (Dougherty et al., 2005). According to the dual-task costs explanation, because the learning task was more challenging for young than for older children, generating concurrent JOLs therefore produced larger dual-task costs. The mechanisms proposed by these two theories might jointly contribute to the overall positive reactivity effect on children's learning and a positive relation between age and reactivity.

It has to be acknowledged that all above discussions are based on theoretical speculations. Future research is needed to directly investigate the (meta)cognitive underpinnings of the reactivity effect on children's learning and its development trend.

Unlike many other strategies which significantly boost learning and for which learners underappreciate their value (e.g., testing and spaced learning; Yang, Luo, et al., 2021; Kornell & Bjork, 2008; Karpicke & Roediger, 2008; Rivers, 2020), children appear to metacognitively appreciate the benefit of making JOLs, as reflected by the finding that a greater proportion of participants believed JOL > no-JOL than the proportion who believed JOL = no-JOL or JOL < no-JOL. However,

their awareness was far from impressive, with only a small majority believing JOL > no-JOL and a sizable minority believing JOL < no-JOL. By contrast, their awareness was improved by taking a memory test, as reflected by the increased proportion of participants believing JOL > no-JOL and reduced proportion believing JOL < no-JOL. These findings establish that firsthand experience of the reactivity effect, induced by taking a memory test, can be used as a practical technique to enhance children's metacognitive awareness of the benefit of making JOLs (Dunlosky & Hertzog, 2000; Dunlosky et al., 2002; Kelemen et al., 2007; Szpunar et al., 2014). In addition, children showed similar patterns of awareness improvement across grades.

A striking finding is that, prior to taking the recognition test, the accuracy of judgments about the reactivity effect was moderate at best. This might be due to the likelihood that children made their judgments simply based on their *a priori* beliefs about the effect and did not carefully consider their actual mastery levels of JOL and no-JOL words when making their judgments (Koriat, 1993; Dunlosky, Rawson, & Middleton, 2005). This hypothesis is supported by the finding that recognition test performance, which provided implicit diagnostic feedback, successfully improved their judgment accuracy (Dunlosky et al., 2002).

In sum, the takeaway messages from the current study are as follows: (1) The positive reactivity effect generalizes to children in Grades 1, 3, and 5, and making JOLs seems to be beneficial for elementary children's multi-modal learning; (2) The enhancing effect of making JOLs is larger for Grade 5 than for Grade 1 and 3 students; (3) Children tend to appreciate the benefit of making JOLs, but their metacognitive awareness leaves considerable room for improvement; (4) Firsthand experience of the reactivity effect, induced by taking a memory test, can enhance their appreciation of the effect and improve their judgmental accuracy.

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

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology, 64*, 417-444. doi:10.1146/annurev-psych-113011-143823

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLoS One, 5*, 1-8. doi:http://dx.doi.org/10.1371/journal.pone.0010729

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

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

Dougherty, M. R., Scheck, P., Nelson, T. O., & Narens, L. (2005). Using the past to predict the future. *Memory & cognition, 33*, 1096-1115. doi:10.3758/bf03193216

Dunlosky, J., & Hertzog, C. (2000). Updating knowledge about encoding strategies: A componential analysis of learning about strategy effectiveness from task experience. *Psychology & Aging, 15*, 462–474. doi:10.1037//0882-7974.15.3.462

Dunlosky, J., Rawson, K. A., & McDonald, S. L. (2002). Influence of practice tests on the accuracy of predicting memory performance for paired associates, sentences, and text material. In T. J. Perfect & B. L. Schwartz (Eds.), *Applied Metacognition* (pp. 68-92, Chapter xi, 297 Pages): Cambridge University Press, New York, NY.

Dunlosky, J., Rawson, K. A., & Middleton, E. L. (2005). What constrains the accuracy of metacomprehension judgments? Testing the transfer-appropriate-monitoring and accessibility hypotheses. *Journal of Memory and Language, 53*, 551-565. doi:http://dx.doi.org/10.1016/j.jml.2005.01.011

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

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

Hartwig, M. K., & Dunlosky, J. (2012). Study strategies of college students: Are self-testing and scheduling related to achievement? *Psychonomic Bulletin & Review*, *19*, 126-134. doi:10.3758/s13423-011-0181-y

Horn, S. S., Bayen, U. J., & Michalkiewicz, M. (2021). The development of clustering in episodic memory: A cognitive-modeling approach. *Child Development, 92*, 239-257. doi:10.1111/cdev.13407

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, 28*, 2356–2364. doi:10.3758/s13423-018-1463-4

Karpicke, J. D., & Roediger, H. L., 3rd. (2008). The critical importance of retrieval for learning. *Science, 319*, 966-968. doi:10.1126/science.1152408

Kelemen, W., Winningham, R. G., & Weaver, C. A. (2007). Repeated testing sessions and scholastic aptitude in college students' metacognitive accuracy. *European Journal of Cognitive Psychology, 19*, 689-717. doi:10.1080/09541440701326170

Kirk-Johnson, A., Galla, B. M., & Fraundorf, S. H. (2019). Perceiving effort as poor learning: The misinterpreted-effort hypothesis of how experienced effort and perceived learning relate to study strategy choice. *Cognitive Psychology, 115*, 101237. doi:https://doi.org/10.1016/j.cogpsych.2019.101237

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

Klenberg, L., Korkman, M., & Lahti-Nuuttila, P. (2001). Differential development of attention and executive functions in 3- to 12-year-old Finnish children. *Developmental Neuropsychology, 20*, 407-428. doi:10.1207/S15326942DN2001_6

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review, 100*, 609-639. doi:http://dx.doi.org/10.1037/0033-295X.100.4.609

Kornell, N. (2009). Optimising learning using flashcards: Spacing is more effective than cramming. *Applied Cognitive Psychology, 23*, 1297-1317. doi:https://doi.org/10.1002/acp.1537

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review*, *14*, 219 -224. doi:https://doi.org/10.3758/BF03194055

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction"? *Psychological Science, 19*, 585-592. doi:https://doi.org/10.1111/j.1467-9280.2008.02127.x

Metcalfe, J. (2009). Metacognitive judgments and control of study. *Current Directions in Psychological Science, 18*, 159-163. doi:10.1111/j.1467-8721.2009.01628.x

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:http://dx.doi.org/10.1037/a0039923

Morehead, K., Rhodes, M. G., & DeLozier, S. (2016). Instructor and student knowledge of study strategies. *Memory*, *24*, 257-271. doi:10.1080/09658211.2014.1001992

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

Potts, R., & Shanks, D. R. (2014). The benefit of generating errors during learning. *Journal of Experimental Psychology: General, 143*, 644-667. doi:10.1037/a0033194

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:http://dx.doi.org/10.1037/a0021705

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

Rivers, M. L. (2020). Metacognition about practice testing: A review of learners' beliefs, monitoring, and control of test-enhanced learning. *Educational Psychology Review*, *33*, 832-862. doi:10.1007/s10648-020-09578-2

Ruff, H. A., Lawson, K. R., Parrinello, R., & Weissberg, R. (1990). Long-term stability of individual differences in sustained attention in the early years. *Child Development, 61*, 60-75. doi:10.1111/j.1467-8624.1990.tb02760.x

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

Schleepen, T. M., & Jonkman, L. M. (2012). Children's use of semantic organizational strategies is mediated by working memory capacity. *Cognitive Development, 27*, 255-269. doi:https://doi.org/10.1016/j.cogdev.2012.03.003

Schneider, W. (1986). The role of conceptual knowledge and metamemory in the development of organizational processes in memory. *Journal of Experimental Child Psychology, 42*, 218-236. doi:https://doi.org/10.1016/0022-0965(86)90024-X

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*, 315-317. doi:10.1111/j.1467-9280.1992.tb00680.x

Szpunar, K. K., Jing, H. G., & Schacter, D. L. (2014). Overcoming overconfidence in learning from video-recorded lectures: Implications of interpolated testing for online education. *Journal of Applied Research in Memory and Cognition, 3*, 161-164. doi:10.1016/j.jarmac.2014.02.001

Tauber, S. K., Dunlosky, J., & Rawson, K. A. (2015). The influence of retrieval practice versus delayed judgments of learning on memory: Resolving a memory-metamemory paradox. *Experimental Psychology, 62*, 254-263. doi:10.1027/1618-3169/a000296

Tauber, S. K., Dunlosky, J., Rawson, K. A., Wahlheim, C. N., & Jacoby, L. L. (2013). Self-regulated learning of a natural category: Do people interleave or block exemplars during study? *Psychonomic Bulletin & Review, 20*, 356-363. doi:10.3758/s13423-012-0319-6

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

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

Wellman, H. M., & Johnson, C. N. (1979). Understanding of mental processes: A developmental study of "remember" and "forget". *Child Development, 50*, 79-88. doi:10.2307/1129044

Wetzel, N. (2014). Development of control of attention from different perspectives. *Frontiers in Psychology, 5*. doi:10.3389/fpsyg.2014.01000

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

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*, 399-435. doi:10.1037/bul0000309

Yang, C., Potts, R., & Shanks, D. R. (2017a). The forward testing effect on self-regulated study time allocation and metamemory monitoring. *Journal of Experimental Psychology: Applied, 23*, 263-277. doi:10.1037/xap0000122

Yang, C., Potts, R., & Shanks, D. R. (2017b). 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

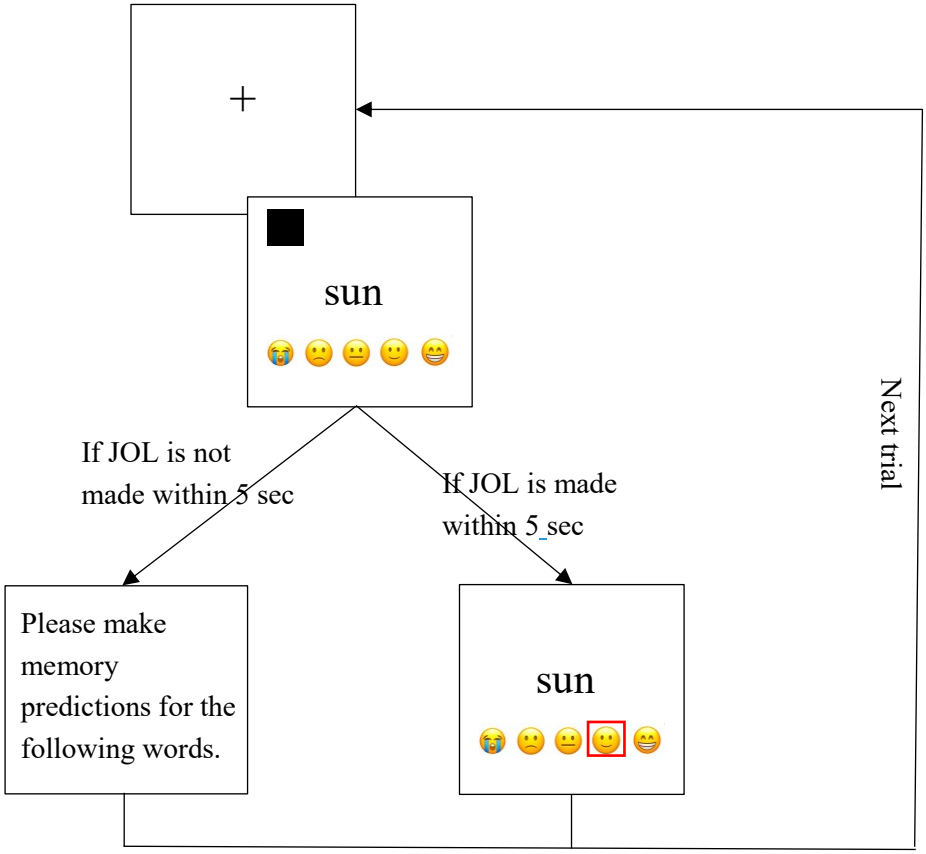

Figure 1. Schematic illustration of the task procedure in the JOL condition in Experiment 1. Note that the word and sentence presented in this figure are English translations of the original (Chinese) one.

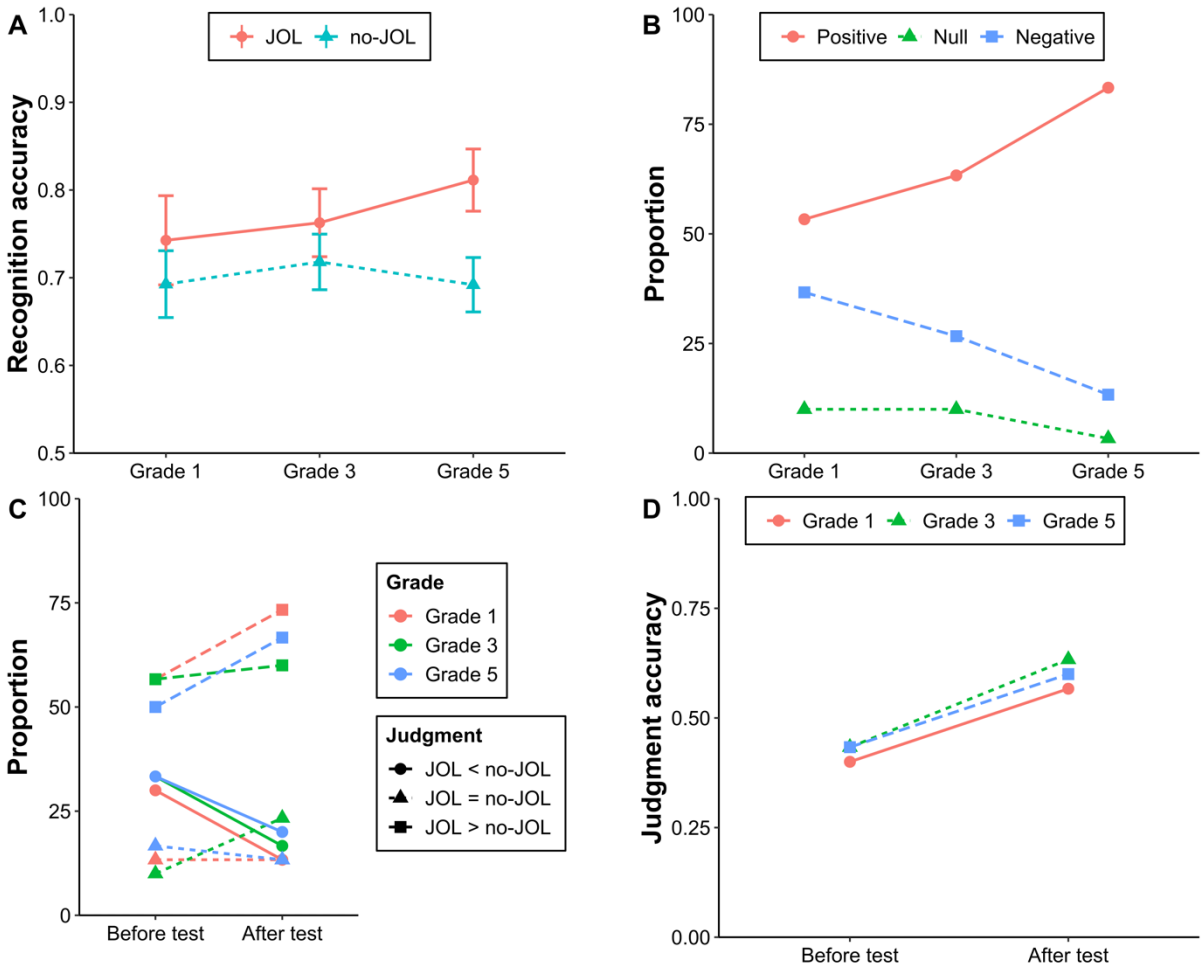

Figure 2. Results of Experiment 1. A: Recognition accuracy as a function of Study method and Grade. B: Proportions of participants who experienced positive (JOL > no-JOL), no (JOL = no-JOL), or negative (JOL < no-JOL) reactivity effects in the recognition test. C: Proportions of participants who reported JOL > no-JOL, JOL = no-JOL, or JOL < no-JOL as a function of judgment timing (before versus after testing) at each grade. D: Accuracy of judgments made before and after testing at each grade. Error bars represent 95% CI.

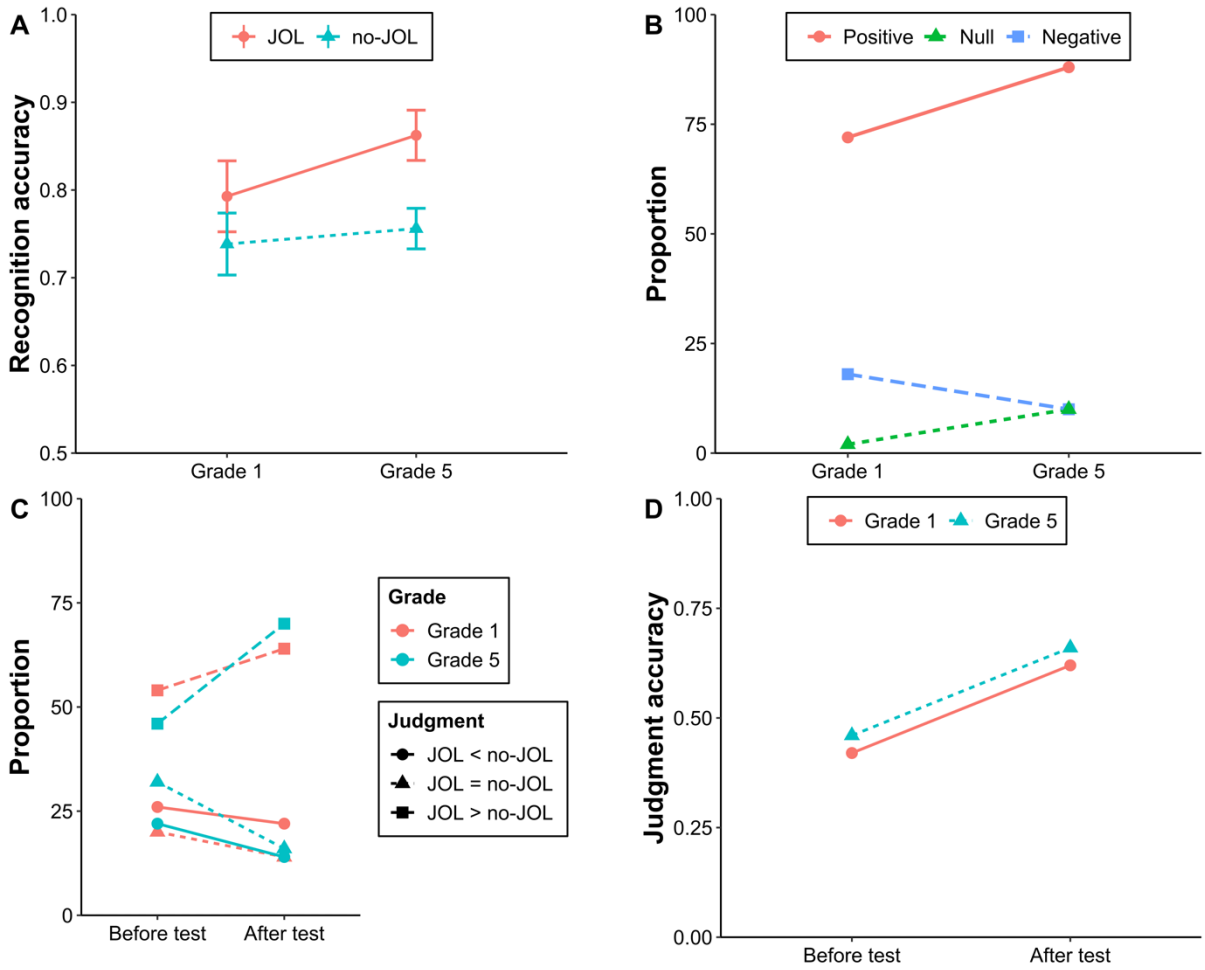

Figure 3. Results of Experiment 2. A: Recognition accuracy as a function of Study method and Grade. B: Proportions of participants who experienced positive (JOL > no-JOL), no (JOL = no-JOL), or negative (JOL < no-JOL) reactivity effects in the recognition test. C: Proportions of participants who reported JOL > no-JOL, JOL = no-JOL, or JOL < no-JOL as a function of judgment timing (before versus after testing) at each grade. D: Accuracy of judgments made before and after testing at each grade. Error bars represent 95% CI.