Metacognition and Learning (2021) 16:407–429
https://doi.org/10.1007/s11409-021-09260-0

The effect of delayed judgments of learning on retention

Eylul Tekin[1] · Henry L. Roediger III[1]

Received: 25 March 2020 / Accepted: 7 February 2021 / Published online: 27 February 2021

Abstract
Evidence is mixed concerning whether delayed judgments of learning (JOLs) enhance learning and if so, whether their benefit is similar to retrieval practice. One potential explanation for the mixed findings is the truncated search hypothesis, which states that not all delayed JOLs lead to a full-blown covert retrieval attempt. In three paired-associate learning experiments, we examined the effect of delayed JOLs on later recall by comparing them to conditions of restudy, overt retrieval, and various other delayed JOL conditions. In Experiment 1, after an initial study phase, subjects either restudied word pairs, practiced overt retrieval, or made cue-only or cue-target delayed JOLs. In Experiments 2a and 2b, where conditions were manipulated within-subjects, subjects either restudied word pairs, practiced overt retrieval, made cue-only delayed JOLs, made cue-only delayed JOLs followed by a *yes/no* retrieval question or, in another condition, by an overt retrieval prompt. The final cued recall tests were delayed by two days. In Experiment 1, recall after cue-only delayed JOLs did not reliably differ from recall after overt retrieval or restudy. In Experiments 2a and 2b, delayed JOLs consistently produced poorer recall relative to overt retrieval. Furthermore, reaction times for delayed JOLs were shorter relative to delayed JOLs paired with overt retrieval prompts. We conclude that only some delayed JOLs elicit covert retrieval attempts, a pattern supporting the truncated search hypothesis.

Keywords Delayed judgments of learning · Covert retrieval · Truncated search · Reactivity

✉ Eylul Tekin
elifeylultekin@wustl.edu

Henry L. Roediger, III
roediger@wustl.edu

[1] Department of Psychological and Brain Sciences, Washington University in St. Louis, St Louis, MO 63130-4899, USA

Judgments of learning (JOLs) are predictive assessments made during study about how well information will be retrieved on a future test and they are likely to determine future study behavior (Nelson and Narens 1990). A robust finding in the metacognition literature is that JOLs are highly accurate in predicting future retention when people provide JOLs after a delay rather than immediately, referred to as *the delayed-JOL effect* (Nelson and Dunlosky 1991; Dunlosky and Nelson 1992, 1994). That is, in paired-associate learning studies (e.g., *artist - baker*), cue-only JOLs (*artist -?*) given after a delay showed higher discrimination between recalled and non-recalled word pairs on a cued recall test (i.e., higher relative metacognitive accuracy or resolution), relative to JOLs (*artist -?*) given immediately after study and cue-target delayed JOLs (*artist - baker).* Although the delayed-JOL effect has been replicated across many studies, whether delayed JOLs influence retention is still debated. Therefore, the current study examined the effect of delayed JOLs on retention and its underlying mechanism.

The idea that making delayed JOLs might affect retention has been around since the discovery of the delayed-JOL effect. For instance, Spellman and Bjork (1992) argued that making delayed-JOLs changed retention of at least some items through spaced retrieval practice, thus yielding reactive effects for delayed JOLs (for a review, see Rhodes and Tauber 2011). The empirical studies investigating the effects of delayed JOLs on retention, however, have provided mixed results. In a meta-analysis, Rhodes and Tauber (2011) compared recall for immediate and delayed JOLs across 98 effect sizes and reported only a modest benefit for delayed JOLs ($g = 0.08$). In addition, the effect was only present for cue-target delayed JOLs ($g = 0.70$), but not for cue-only delayed JOLs ($g = 0.03$). In all these studies, however, delayed JOLs were compared to immediate JOLs. Critically, recent studies have revealed that immediate JOLs can be also reactive measures in paired-associate learning paradigms (Mitchum et al. 2016; Soderstrom et al. 2015). That is, making immediate JOLs influences learning outcomes relative to not making JOLs. As a result, immediate JOLs are not ideal control conditions to investigate the effect of delayed JOLs on retention. Given that the main purpose of prior studies was to examine the theoretical accounts of the delayed JOL effect, these studies rarely compared delayed JOLs to a neutral control condition (i.e., without reactive effects).

Kelemen and Weaver's (1997) research was one exception that included a control condition without JOLs. In their paired-associate learning experiments, subjects studied word pairs while making immediate JOLs after some pairs, cue-only delayed JOLs for other pairs, and without making JOLs for the remaining pairs. They found that delayed JOLs improved recall compared to not making any JOLs, suggesting that the act of making delayed JOLs might improve retention compared to a neutral control condition. Of course, the positive effect of delayed JOLs could be due to spaced retrieval practice, at least for the items that were retrieved during cue-only JOLs. The primary goal of the current study was to further examine whether making delayed JOLs affects later retention by comparing them to various learning conditions.

In a typical delayed JOL paradigm, JOLs are either made after a few study trials or after the initial study phase. As indicated by Spellman and Bjork (1992), this procedure allows subjects to have spaced retrieval practice for at least some of the items that were retrieved while making delayed JOLs, thereby selectively boosting retention for those items. When retention after study without delayed JOLs is compared to retention after delayed JOLs as in Kelemen and Weaver (1997), items in the control condition have an inherent disadvantage: They are only presented once, and thus do not benefit from the spaced practice. Given that pairs followed by delayed JOLs potentially receive spaced retrieval practice, whereas words studied without delayed JOLs do not, the comparison of these conditions cannot separate the potential effects

of spacing on retention from the effects of making delayed JOLs. Therefore, an alternative way to examine whether making delayed JOLs influences retention is to use a neutral control condition that also benefits from spacing. This control condition would ensure that the two conditions only differ in the act of making delayed JOLs. Therefore, in the present experiments, we examined how delayed JOLs affect learning in comparison to restudy without JOLs.

As a second goal, we explored the mechanisms that account for delayed JOLs' effect on retention. First, we investigated covert retrieval (i.e., retrieving without responding aloud or in writing) as a potential mechanism to explain how delayed JOLs influence retention. The two theoretical accounts of the delayed-JOL effect, the monitoring-dual-memories and self-fulfilling prophecy hypotheses, both state that people make delayed JOLs by attempting covert retrieval of the target. Nelson and Dunlosky (1991, 1992) proposed that people make cue-target and cue-only immediate JOLs based on target information accessible dominantly in short-term memory and to a lesser degree in long-term memory, whereas they make cue-only delayed JOLs based on target information that they covertly retrieve only from long-term memory (i.e., the monitoring-dual-memories or MDM hypothesis). The delayed-JOL effect occurs because recall also depends on retrieval from long-term memory. In contrast, Spellman and Bjork (1992) proposed the self-fulfilling prophecy or SFP hypothesis, which maintains that the delayed JOL effect is a memory-based phenomenon rather than a monitoring one. That is, people provide high delayed JOLs if they covertly retrieve the target word while making delayed JOLs, and thus high JOL pairs benefit from spaced retrieval practice, whereas low JOL pairs do not (Cepeda et al. 2006). Accordingly, the delayed-JOL effect occurs because the selective spaced retrieval practice of the high JOL pairs increases their retention on the final test relative to low JOL pairs (that are not retrieved), thereby increasing the monitoring accuracy of delayed JOLs. Despite proposing different explanations for the delayed JOL effect, the MDM and SFP hypotheses assume that cue-only delayed JOLs elicit covert retrieval attempts (i.e., the covert retrieval hypothesis) and solely involve a covert retrieval process (for a review, see Rhodes 2016; Rhodes and Tauber 2011).

Engaging in retrieval for a set of material one has previously studied enhances retention compared to restudying the material when the test is delayed, a robust finding known as *the testing effect*, (Roediger and Karpicke 2006; for a review, see McDermott 2021). Covert retrieval in paired-associate learning paradigms often produces similar testing benefits relative to restudy on delayed recall (Putnam and Roediger 2013; Smith et al. 2013). For instance, Putnam and Roediger (2013, Experiment 3) contrasted restudy, covert retrieval, and overt retrieval and found that covert and overt retrieval yielded similar levels of recall and enhanced recall relative to restudy. If delayed JOLs consistently evoke covert retrieval attempts, then making delayed JOLs should also enhance recall relative to restudy and produce benefits comparable to those of overt retrieval practice based on the findings in the testing effect literature (Putnam and Roediger 2013; Smith et al. 2013). Therefore, in the current study, we examined covert retrieval as one potential mechanism for memory effects of delayed JOLs by comparing delayed JOLs to overt retrieval and restudy conditions.

Two prior studies have directly compared delayed JOLs to overt retrieval. In Jönsson et al. (2012), subjects studied Swahili-Swedish word pairs in four study-practice trials and took a one-week delayed cued recall test. In the practice phases, they were either tested on all target words (overt retrieval), tested on target words with a dropout method (i.e., correctly recalled items were dropped out from future tests), or asked to make delayed JOLs. They found that the dropout method produced poorer recall compared to the delayed JOLs and overt retrieval conditions, whereas there was no reliable recall difference between delayed JOL and overt

retrieval conditions. These results suggest that delayed JOLs produce comparable benefits to that of overt retrieval practice, potentially through covert retrieval. Nevertheless, multiple study-test practice trials are not the standard procedure in the delayed JOL literature and might have contributed to these results.

Tauber et al. (2015) also compared delayed JOLs to overt retrieval and immediate JOL conditions, but they employed a single-trial paired-associate learning paradigm. They found that retrieval practice produced greater recall than delayed JOLs on a two-day delayed cued recall test; however, this recall difference did not reach conventional statistical significance in a one-tailed t-test ($p = .08$). By conducting a cross-experimental analysis to increase power, they reported a small but reliable benefit of retrieval practice over delayed JOLs. In addition, no recall difference between immediate JOLs and delayed JOLs was obtained. Tauber et al. concluded that delayed JOLs did not enhance learning because they promoted truncated retrieval attempts that short-circuited recall during delayed JOLs relative to retrieval practice. Tauber et al. argued that these results supported a truncated search hypothesis, which states that people stop their search for the target words prematurely (i.e., they truncate retrieval or make shorter retrieval attempts) during delayed JOLs instead of attempting to fully retrieve the target words as in retrieval practice (see also Son and Metcalfe 2005). When people truncate their search, they can still covertly retrieve some of the target words, but critically, some target words that would have been retrieved through retrieval practice are not retrieved during delayed JOLs. The reaction time results further supported the truncated search hypothesis: Subjects spent less time on making delayed JOLs than on practicing retrieval. Thus, relative to retrieval practice, subjects truncated their search for the target word earlier while making JOLs than in practicing retrieval.

Truncated search is yet another mechanism through which JOLs might influence retention. However, in Tauber et al. (2015), only the cross-experimental analysis revealed a reliable effect that supported the truncated search hypothesis, and these results were inconsistent with the findings of Jönsson et al. (2012). Therefore, it is important to further examine the truncated search hypothesis as a potential mechanism that may account for the effect of delayed JOLs on retention. Furthermore, neither of these studies employed a restudy condition, which is one of the standard control conditions in the testing effect literature. Delayed JOLs might still benefit recall relative to restudy, even when their effect on memory is not identical to that of retrieval practice (Putnam and Roediger 2013). Although Tauber et al. (2015) concluded that delayed JOLs did not enhance learning, they compared delayed JOLs to immediate JOLs instead of restudy. However, as stated before, because immediate JOLs have reactive effects, they are not ideal control conditions to examine how delayed JOLs affect later retention. The current experiments address both these issues.

Both of the aforementioned accounts of delayed JOL effects state that these JOLs elicit some form of retrieval. However, delayed JOLs might affect recall through mechanisms other than retrieval (i.e., non-retrieval reactivity). For instance, making delayed JOLs might prompt people to elaborate on the meaning of the word pairs more than they would otherwise, which might in turn influence their retention (Tekin & Roediger, 2020). In line with this possibility, recent studies have reported that making immediate JOLs in paired-associate learning enhances later recall, relative to not making JOLs. In Soderstrom et al. (2015), subjects who made immediate JOLs while studying related and unrelated pairs recalled more related pairs than subjects who studied the word pairs without JOLs (i.e., positive JOL reactivity for related pairs). Witherby and Tauber (2017) replicated positive JOL reactivity for related pairs and showed that similar results were obtained on a two-day delayed cued recall test as well (also see Janes et al. 2018). Indeed, in a meta-analysis, Double et al. (2018) reported that immediate

JOLs were moderately reactive measures for related pairs (also see Rhodes 2016). Although the mechanisms behind immediate JOL reactivity are still debated, two prominent accounts suggested that immediate JOLs benefit retention either through strengthening the information used to make JOLs when the memory test is sensitive to such information (Myers et al. 2020; Soderstrom et al. 2015) or through making people aware of different levels of item difficulty, and thus altering their learning goals (for more details, see Janes et al. 2018; Mitchum et al. 2016).

Although these proposed mechanisms have not been investigated for delayed JOLs, they might still apply to them. That is, rather than covert or truncated retrieval, delayed JOLs might influence retention through the mechanisms documented for immediate JOLs. Therefore, we lastly examined whether non-retrieval mechanisms can explain the possible effect of delayed JOLs on recall (i.e., JOL reactivity). To test this possibility, we included a comparison condition of cue-target delayed JOLs (e.g., *artist - baker*). Cue-target delayed JOLs should not evoke any retrieval attempts because the target is presented at the time of the JOL. Thus, any recall differences we might observe between cue-target delayed JOLs and restudy would be attributable to a non-retrieval mechanism that is similar to the reactivity of immediate JOLs. To test non-retrieval reactivity, we employed related and unrelated word pairs in Experiment 1, because prior studies have frequently reported that immediate JOLs improved recall of related pairs (i.e., reactive effects) in paired-associate learning paradigms but not for unrelated pairs (Double et al. 2018; Janes et al. 2018).

Although delayed JOLs have previously been contrasted to immediate JOLs and to retrieval practice in examining their effects on retention, there has been no systematic comparison of delayed JOLs to various learning conditions including restudy. We examined how delayed JOLs affect paired-associate learning by comparing cue-only delayed JOLs to overt retrieval, restudy, and various other types of delayed JOLs. If delayed JOLs solely elicit covert retrieval attempts, they should enhance later recall similarly to overt retrieval practice and more so than restudy (Putnam and Roediger 2013; Smith et al. 2013). If delayed JOLs elicit truncated retrieval attempts, as hypothesized by Tauber et al. (2015), overt retrieval should benefit learning more than delayed JOLs, and delayed JOLs may or may not boost recall relative to restudy. We also analyzed reaction times of delayed JOLs and retrieval conditions to test the truncated search hypothesis. Lastly, if cue-only delayed JOLs evoke non-retrieval reactivity, cue-target delayed JOLs should also enhance recall relative to restudy, and this benefit should be similar to that of cue-only delayed JOLs.

Experiment 1

In Experiment 1, subjects studied related and unrelated cue-target word pairs and then practiced the word pairs in one of the following between-subjects conditions before taking a two-day delayed cued recall test: 1) restudy, 2) cue-target delayed JOLs, 3) cue-only delayed JOLs, and 4) overt retrieval practice. A pilot experiment using the same word pairs reported in the supplementary online material (SOM, osf.io/yaguz/) revealed no differences among conditions and ceiling effects for related word pairs on an immediate cued recall test. Therefore, in Experiment 1, the final test was delayed two-days. We asked: 1) Do delayed JOLs produce similar proportions of recall and reaction times as overt retrieval? 2) Do delayed JOLs enhance recall relative to restudy? and 3) If we obtain a benefit in recall for delayed JOLs, what is the underlying mechanism for the effect?

Method

Subjects One hundred eighty-four undergraduates were recruited for course credit or payment ($10). After excluding 19 subjects who did not return for Session 2 and one for experimental error, 164 subjects remained. There were 43 subjects in the cue-only JOL condition, 41 in the cue-target JOL condition, 40 in the restudy condition, and 40 in the overt retrieval condition.

Materials and design Sixty word pairs were selected from Nelson et al. (2004). The words had concreteness levels ranging from 3.5 to 7 on a scale of 7 (Nelson et al. 2004), and their HAL frequencies range from 6.92 to 13.02 (Balota et al. 2007). Half of the word pairs were related, meaning that the target was one of the three strongest associations of the cue (e.g., *table – chair*), while the unrelated word pairs were not associatively related (e.g., *wind – circle*) and were picked randomly from Nelson et al. (2004).[1] A president recognition test was used as the filler task between the initial study phase and the practice phase of the experiment; the task consisted of 123 president and non-president names with subjects instructed to identify the president (Roediger and DeSoto 2016).

A 4 (practice) × 2 (relatedness) mixed factorial design was used for this experiment. After studying the pairs, the practice phase was manipulated between-subjects and consisted of four conditions: cue-only JOL, cue-target JOL, restudy, and overt retrieval. The relatedness of word pairs was manipulated within-subjects.

Procedure The experiment consisted of two sessions. In Session 1, the initial study phase was identical for all subjects. Subjects studied 60 word pairs, 30 related and 30 unrelated, in a randomized order for each subject. Each word pair was presented for 4-s. Following the study phase, subjects completed the filler task (i.e., the president recognition task) that lasted approximately 10 min.

During the practice phase, subjects were randomly assigned to one of four conditions (cue-only JOL, cue-target JOL, restudy, and overt retrieval) and received relevant instructions. For the JOLs, subjects were informed about the time and format of the final test. Figure 1a provides an example trial from the practice phase for each condition. In the cue-only JOL condition, subjects only saw cues of cue-target pairs and rated their likelihood of remembering the target on a final test given the cue, on a 100-point scale, with 100 indicating *I would definitely recall the word*. Subjects in the cue-target JOL condition saw both cues and targets, and they rated their likelihood of remembering the target given the cue. Subjects were allowed to use the numeric keypad on the keyboard to enter their ratings. Subjects in the overt retrieval condition were presented with cues and asked to recall target words. For JOLs and overt retrieval trials, subjects had 8-s to make a response and then they received feedback (the intact word pair) for 2-s. For JOL and overt retrieval trials, there were no submission keys; the experiment automatically moved onto feedback after 8-s. Subjects were allowed to correct their responses and use the 'backspace' key within the 8-s window. In the restudy condition, subjects were presented with cue-target word pairs for 10-s each; they were not asked to give any kind of response. Session 1 ended after the practice phase. Subjects came back for Session 2 after two days and they received the final cued recall test in which they had 8-s per trial to recall the target for each cue.

[1] Although targets did not appear equally among related and unrelated pairs, the characteristics of targets were similar across related and unrelated pairs. For related targets, the average word length was 4.8 letters, the average logarithmic HAL frequency was 10.28 and the average concreteness was 4.53 out of 5. For unrelated targets, the average word length was 5.1 letters, the average logarithmic HAL frequency was 9.99 and the average concreteness was 4.39 out of 5.

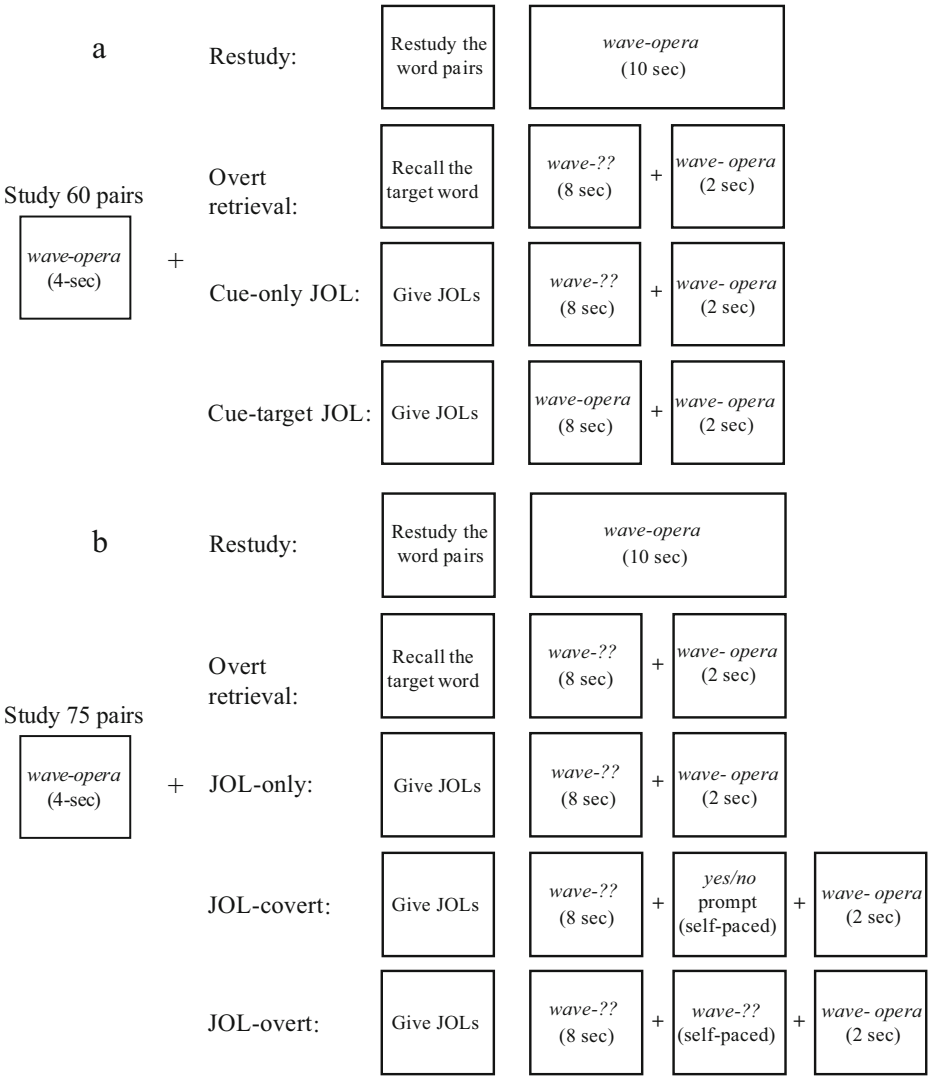

Fig. 1 An example trial from the practice phase for each condition in Experiment 1 (a) and in Experiments 2a and 2b (b)

Results

Our primary interest was the proportion of correct recall on the final test, so we first report analyses of the final cued recall test. We then report reaction time analyses and the initial proportion recalled during overt retrieval practice. For cued recall tests, we report the results based on lenient scoring because it allowed for simple typos (e.g., typing "hammar" instead of "hammer"). The results did not change when we used a strict scoring criterion. Reaction times were measured as the time it took subjects to press the first key in each trial during practice conditions because first keypress latencies were not affected by typing differences across items (e.g., word length) and conditions (e.g., typing a numeric rating is much shorter than typing a

word). If subjects did not respond within 8-s on any trial, their reaction time for that trial was not included in data analysis (12% out of 2400 trials in the overt retrieval condition). The mean reaction time for each condition was computed from mean reaction times per subject. If the sphericity assumption is violated, the Greenhouse-Geisser correction was used. For any reported pairwise comparisons, Tukey's HSD was conducted unless otherwise reported. The results for the JOLs are reported in the Appendix.

Final cued recall test Figure 2 shows that higher proportion of related word pairs were recalled than unrelated word pairs, and subjects in the overt retrieval condition recalled slightly more than those in the restudy condition. A 2 (relatedness) × 4 (practice) mixed factorial ANOVA confirmed a main effect of relatedness, $F(1,160) = 2114.38$, $p < .001$, $\eta^2_p = .93$, and a main effect of practice condition, $F(3,160) = 2.80$, $p = .042$, $\eta^2_p = .05$. The interaction was not reliable, $F(3,160) = 1.73$, $p = .163$, $\eta^2_p = .03$.

On average, recall proportions were .44, .45, .50, and .52 for the restudy, cue-target JOL, cue-only JOL and overt retrieval conditions, respectively. Although there was a $.09^2$ recall difference between the overt retrieval and restudy conditions, it did not reach conventional levels of statistical significance, $p = .056$, $d = .57$. None of the other pairwise comparisons were significant [restudy vs. cue-target JOL, $p = .967$, $d = .11$; restudy vs. cue-only JOL, $p = .264$, $d = .38$; overt retrieval vs. cue-target JOL, $p = .157$, $d = .54$; overt retrieval vs. cue-only JOL, $p = .869$, $d = .16$; cue-target JOL vs. cue-only JOL, $p = .521$, $d = .31$].

Reaction times The middle section of Table 1 shows reaction times in seconds measured by first key press of subjects for the cue-only JOL, cue-target JOL and overt retrieval conditions. A 2 (relatedness) × 3 (practice) mixed factorial ANOVA showed that subjects were faster to respond to related word pairs than unrelated word pairs, $F(1,121) = 184.39$, $p < .001$, $\eta^2_p = .64$. The main effect of practice condition on reaction time was also reliable, $F(2,118) = 8.54$, $p < .001$, $\eta^2_p = .12$, and it was moderated by a reliable interaction, $F(2,118) = 20.78$, $p < .001$, $\eta^2_p = .26$. Pairwise comparisons showed that for related pairs, subjects took less time to start typing target words than typing JOL ratings, [overt retrieval vs. cue-only JOL $p < .001$, $d = 1.29$; overt retrieval vs. cue-target JOL, $p < .001$, $d = 1.52$; cue-target JOL vs. cue-only JOL $p = .985$, $d = .07$], whereas for unrelated pairs, subjects' reaction times did not reliably differ, a finding in opposition with the truncated search hypothesis [overt retrieval vs. cue-only JOL $p = .195$, $d = .38$; overt retrieval vs. cue-target JOL, $p = .971$, $d = .10$; cue-target JOL vs. cue-only JOL $p = .391$, $d = .30$]. However, subjects gave ratings to all items in the two JOL conditions, whereas they only responded to items they could recall in the overt retrieval condition. Thus, item selection inherent in this procedural difference probably also contributed to the observed results.

Practice cued recall test For each subject in the overt retrieval condition, proportion correct was calculated for the practice cued recall test. The middle section of Table 2 shows mean recall for related and unrelated pairs on the practice cued recall test, and the final recall results appear in Fig. 2b. A 2 (relatedness) × 2 (test) repeated measures ANOVA revealed a main effect of relatedness, $F(1,39) = 772.86$, $p < .001$, $\eta^2_p = .95$, and a main effect of test, $F(1,39) = 13.12$, $p = .001$, $\eta^2_p = .25$. On average, subjects' recall showed a small but reliable increase from the practice test ($M = .48$, $SE = .02$) to the final test ($M = .52$, $SE = .02$), even

[2] The actual recall difference between overt retrieval (.524) and restudy (.437) is .087, due to the rounding, recall differences of .08 (i.e., .52–.44) and .09 do not match.

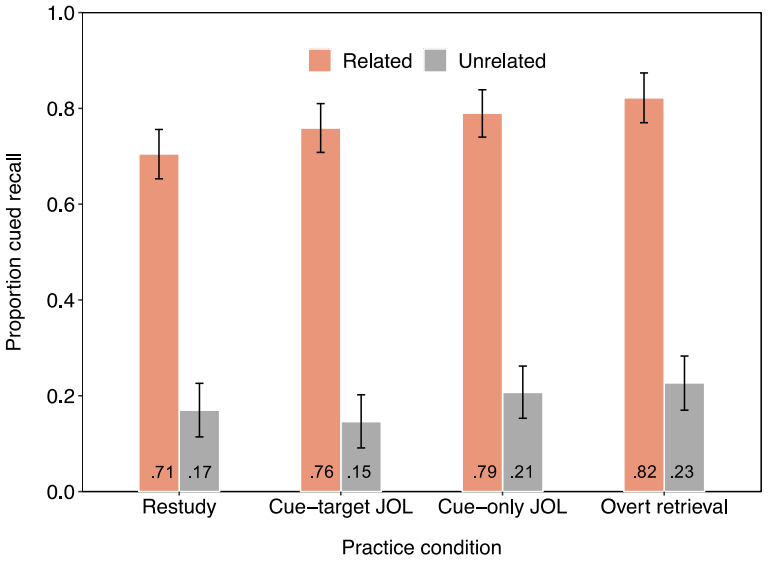

Fig. 2 Final recall performance in Experiment 1. Error bars indicate 95% confidence intervals

after a two-day delay possibly due to feedback. The interaction was also reliable, $F(1,39) = 25.97$, $p < .001$, $\eta^2_p = .40$. Recall for unrelated pairs did not change across tests, $p = .80$, $d = .09$, whereas recall for related pairs improved from the practice test to the final test, $p < .001$, $d = .91$.

Discussion

In Experiment 1, we examined the effects of delayed JOLs on recall on a two-day delayed test. If cue-only delayed JOLs only elicited covert retrieval attempts, we would expect to observe 1)

Table 1 Reaction Times in Seconds Across Conditions and Experiments

Experiment 1				
Condition	Related		Unrelated	
	M	*SD*	*M*	*SD*
Cue-target JOL	3.00	.51	3.27	.59
Cue-only JOL	2.96	.62	3.49	.85
Overt retrieval	2.26	.48	3.21	.59
Experiment 2a				
Condition	JOL		Overt retrieval	
	M	*SD*	*M*	*SD*
JOL-only	3.33	.72		
JOL-covert	3.49	.66		
JOL-overt	3.65	.58	1.65	1.25
Overt retrieval			2.70	.50
Experiment 2b				
JOL-only	3.46	.67		
JOL-covert	3.58	.61		
JOL-overt	3.98	.77	1.60	1.02
Overt retrieval			2.67	.56

Table 2 Recall Performance on Practice Cued Recall Tests Across Conditions and Experiments

Experiment 1				
Condition	Related		Unrelated	
	M	*SD*	*M*	*SD*
Overt retrieval	.73	.13	.24	.16
Experiment 2a				
Condition	Weakly related			
	M	*SD*		
Overt retrieval	.47	.22		
JOL-overt	.44	.25		
Experiment 2b				
Overt retrieval	.45	.25		
JOL-overt	.50	.26		

similar recall between cue-only JOLs and overt retrieval and 2) higher recall from both conditions compared to restudy. In Experiment 1, we obtained no reliable recall differences between cue-only JOLs and overt retrieval; however, cue-only JOLs did not reliably increase recall compared to restudy either (though it did numerically). This finding contradicts the idea that delayed JOLs are strictly a form of covert retrieval because covert retrieval enhances learning relative to restudy (Putnam and Roediger 2013; Smith et al. 2013). Of course, subjects may covertly retrieve on some trials when making delayed JOLs, just not all trials (Son and Metcalfe 2005). Our descriptive results showed that cue-only delayed JOLs yielded recall (.50) in between restudy (.44) and overt retrieval (.52). This finding seems to be in line with the truncated search hypothesis (Tauber et al. 2015), which states that when people make delayed JOLs, they do not benefit as much as in the overt retrieval condition because they engage in truncated search for target words during delayed JOLs rather than full-blown covert retrieval attempts.

The reaction time results, however, did not replicate Tauber et al. (2015) who found that delayed JOLs yielded shorter reaction times than overt retrieval, suggesting that people truncated their search prematurely. They concluded that this finding supported the truncated search hypothesis. Instead, we found that compared to overt retrieval, subjects took longer to press a key while making delayed JOLs for related pairs with no reaction time occurring for unrelated pairs. These outcomes suggested that subjects did not truncate their search and spent longer time searching during delayed JOLs for related pairs. However, there were some methodological differences between the two studies. In Tauber et al. (2015), delayed JOLs and overt retrieval were self-paced (i.e., subjects had unlimited time), whereas, in our study, subjects had 8-s to make a JOL or recall the target word. Therefore, the timing differences might have contributed to the observed discrepancy.

Experiment 1 may not have had sufficient power to detect recall differences between practice conditions. For instance, Putnam and Roediger (2013) employed a within-subjects design when they found that covert retrieval produced superior recall relative to a restudy condition. Similarly, Tauber et al. (2015) reported reliable differences between overt retrieval and delayed JOLs after increasing their power through a cross-experimental analysis. To test this possibility, we conducted a post-hoc sensitivity analysis for the practice between-subjects factor (G*Power; Faul et al. 2007) to find the minimum effect size Experiment 1 could detect. The minimum detectable effect size was .23 (Cohen's f). This effect size was larger than the observed effect size in Experiment 1. This suggested that Experiment 1 was under-powered to distinguish effects below .23 from zero which might have been

responsible for the results observed in Experiment 1. Therefore, in Experiment 2, we aimed to increase our statistical power using a within-subjects design and we again investigated the covert retrieval or truncated search hypotheses as potential mechanisms for delayed JOLs' influence on later recall.

Experiment 1 also examined non-retrieval reactivity as a potential mechanism for delayed JOLs. That is, instead of retrieval, delayed JOLs could affect recall through changing people's study goals or strengthening the relationship between related pairs as reported for immediate JOLs (Mitchum et al. 2016; Soderstrom et al. 2015). For instance, employing a similar paired-associate learning paradigm and similar numbers of subjects and items, Witherby and Tauber (2017) reported that making immediate JOLs increased overall recall relative to not making JOLs on a two-day delayed cued recall test. In our experiment, however, recall for the cue-target delayed JOL condition was .45, whereas recall for the restudy condition was .44. Furthermore, although recall for related pairs was numerically higher in the cue-target delayed JOL condition than in the restudy condition, a pattern consistent with prior studies (Janes et al. 2018; Mitchum et al. 2016; Soderstrom et al. 2015), there was no reliable interaction between relatedness and practice conditions to suggest (positive or negative) reactivity for cue-target delayed JOLs. Thus, in Experiment 2, we dropped the cue-target JOL condition because the effects of delayed JOLs on retention did not appear to be mainly driven by the same mechanisms as for immediate JOLs.

Experiment 2

In Experiment 1, cue-only delayed JOLs did not reliably improve recall relative to restudy, suggesting that delayed JOLs did not always elicit covert retrieval (Putnam and Roediger 2013; Smith et al. 2013). This result contradicted a strict covert retrieval hypothesis, which states that covert retrieval is the sole mechanism for delayed JOLs (Nelson and Dunlosky 1991; Spellman and Bjork 1992). Nonetheless, it is possible that while subjects engaged in covert retrieval during some delayed JOL trials, they did not attempt to covertly retrieve the target words in all JOL trials (Son and Metcalfe 2005). Unfortunately, in Experiment 1, we did not obtain any measures on whether subjects retrieved a word while making cue-only JOLs. We addressed this issue in Experiment 2. To examine whether subjects covertly retrieved a word while making cue-only delayed JOLs, in one of the conditions, after each JOL rating, subjects answered a *yes/no* question about whether they retrieved the target word while making the JOL (Putnam and Roediger 2013). This step allowed us to examine the frequency of subjects' (correct or incorrect) covert retrievals during JOLs.

The results of Experiment 1 also showed that cue-only delayed JOLs yielded recall in between restudy and overt retrieval. This outcome suggested that varying the degree of retrieval processes between making delayed JOLs to engaging in overt retrieval might provide insights into whether delayed JOLs elicit covert retrieval or truncated search. To promote varying degrees of retrieval, two additional delayed JOL conditions were included in Experiment 2. First, we investigated whether the aforementioned covert retrieval question altered recall relative to making delayed JOLs alone. That is, asking about covert retrieval might cause subjects to retrieve covertly while making JOLs more so than they would otherwise. Thus, making delayed JOLs along with the covert retrieval question should increase recall relative to making delayed JOLs alone if delayed JOLs normally do not elicit any covert retrieval attempts. Second, we included a condition in which delayed JOLs were followed by overt

retrieval (i.e., delayed JOL + overt retrieval). We predicted that the inclusion of overt retrieval in this condition would promote subjects to engage in retrieval while making delayed JOLs. Therefore, making delayed JOLs followed by overt retrieval should increase recall relative to making delayed JOLs alone if delayed JOLs do not elicit any covert retrieval attempts.

In Experiment 2, we also tested the truncated search hypothesis by comparing reaction times across conditions. Although the observed latency results in Experiment 1 seemed in opposition with the truncated search hypothesis, the reaction time differences might be due to different activities required by these tasks rather than reflecting reaction time differences in search or retrieval. That is, making a likelihood rating about remembering a word on a future test is inherently different than trying to recall that word. Furthermore, subjects in the JOL conditions made JOL ratings within 8-s on 99% of the trials (99% and 98% for related and unrelated pairs, respectively), whereas subjects in the overt retrieval condition did not give any responses in 7% and 18% of trials for related and unrelated pairs, respectively. Because these pairs did not have any reaction time data for overt retrieval, item selection effects may have afflicted our comparisons. To address this issue, we compared reaction times of the same tasks (i.e., delayed JOLs to delayed JOLs and overt retrieval to overt retrieval). For this comparison, we used the new delayed JOL condition in which subjects were asked to recall the target word after each JOL rating (i.e., delayed JOLs + overt retrieval). We predicted that the additional overt retrieval phase would promote subjects to engage in covert retrieval during delayed JOLs. Accordingly, if subjects truncate their search of target words while making delayed JOLs alone, they should spend longer time on delayed JOLs followed by overt retrieval. Furthermore, overt retrieval following delayed JOLs should take less time than overt retrieval alone because people have already engaged in retrieval during delayed JOLs. On the other hand, if delayed JOLs always elicit covert retrieval, no RT differences should be observed between delayed JOLs alone and delayed JOLs followed by overt retrieval.

Finally, in Experiment 2, we employed weakly related pairs instead of related and unrelated pairs. We changed our material to make subjects less likely to guess targets for related pairs, and to avoid floor effects for unrelated pairs. Because there was no interaction between the relatedness of word pairs and practice conditions in previous experiments, we did not expect this change to affect our results. In Experiment 2, subjects studied weakly related cue-target word pairs and practiced the word pairs in one of the five following within-subjects conditions before taking a two-day delayed cued recall test: 1) restudy, 2) cue-only delayed JOLs (JOL-only), 3) cue-only delayed JOLs followed by the *yes/no* questions (JOL-covert), 4) cue-only delayed JOLs with overt retrieval (JOL-overt) and 5) overt retrieval practice without JOLs (overt).

Experiment 2a

Method

Subjects For Experiment 2a, we conducted a priori power analysis to determine sufficient sample size for a within-subjects main effect in a repeated-measures ANOVA using a small effect size ($f = .15$), an alpha of .05, and a power of .80. The required sample size was 55. Sixty undergraduates were recruited for course credit or payment ($10). After excluding five subjects who did not show up for Session 2, and two experimental errors, 53 subjects remained.

Materials and design Seventy-five weakly related word pairs (e.g., *paper – angle*) were taken from Nelson et al. (2004). The word pairs had a forward cue-to-target strength and a backward target-to-cue strength between 0 and .019. These word pairs ranged in concreteness from 2.0 to 5.0 on a scale out of 5 (Brysbaert et al. 2014). As with Experiment 1, the president recognition test was used as a filler task.

In Experiment 2a, a within-subjects design with five levels was used. All subjects participated in five practice conditions after studying the 75 pairs: restudy, JOL-only, JOL-covert, JOL-overt retrieval, and overt retrieval conditions. These conditions were presented in blocks (15 word pairs per block) and were counterbalanced using a Latin Square (i.e., 10 counterbalancing conditions).[3]

Procedure The experiment consisted of two sessions. In the first session, all subjects studied word pairs for 4-s each, followed by the filler task. During the practice phase, subjects completed five practice conditions in blocks. Figure 1b shows an example trial for each condition. The JOL-only condition was identical to the cue-only JOL condition in Experiment 1: Subjects only saw cues of cue-target pairs and made JOLs. In the JOL-covert condition, subjects again saw cues of cue-target pairs and made JOLs. After each JOL rating, subjects answered a *yes*/*no* covert retrieval question (*Did you remember the target word while giving the rating? Be honest!*). In the JOL-overt condition, subjects were asked to overtly recall the target word after making a JOL rating. In these conditions, subjects had 8-s to make a JOL rating, whereas covert retrieval and overt retrieval sections were self-paced. As with Experiments 1 and 2, subjects received feedback for 2-s at the end of each trial. In the overt retrieval condition, subjects were asked to recall the target word for each cue in 8-s (without making JOLs), followed by the 2-s feedback. Finally, in the restudy condition, subjects restudied cue-target word pairs for 10-s each. Subjects returned to the lab after two-days and took a final cued recall test. As with Experiment 1, for each word pair, subjects had 8-s to respond.

Results

Final cued recall test Figure 3a shows that proportion of items recalled showed a gradual increase across conditions from restudy to overt retrieval. A one-way repeated measures ANOVA with five levels confirmed a main effect of practice condition, $F(4,208) = 13.18$, $p < .001$, $\eta^2_p = .20$. Pairwise comparisons revealed that restudy produced poorer recall than the three JOL and overt retrieval conditions, $ps < .01$. When restudy was compared to overt retrieval, JOL-overt, JOL-covert and JOL-only conditions, the effect sizes for these comparisons were $d = .92$, $d = .72$, $d = .62$ and $d = .63$, respectively. Critically, in the JOL-only condition, subjects also recalled reliably fewer target words than in the overt retrieval condition, $p = .026$, $d = .49$. No other comparison was reliable [JOL-only vs. JOL-covert, $p = .79$, $d = .14$; JOL-only vs. JOL-overt, $p = .52$, $d = .24$; overt retrieval vs. JOL-covert, $p = .34$, $d = .22$; overt retrieval vs. JOL-overt, $p = .62$, $d = .20$; JOL-covert vs. JOL-overt, $p = .99$, $d = .06$].

[3] Because we employed two JOL conditions (JOL-covert and JOL-overt) that promoted retrieval in Experiments 2a and 2b, we examined whether these conditions altered how subjects made delayed JOLs in the JOL-only condition. That is, subjects might have continued to engage in covert retrieval during the JOL-only practice phase if they participated in JOL-covert and JOL-overt conditions beforehand. Therefore, we divided the 10 counterbalancing conditions into two groups based on whether the JOL-only condition preceded or followed JOL-covert and JOL-overt conditions and analyzed whether there were any order effects. We did not find any recall and reaction time differences between the two groups in either experiment.

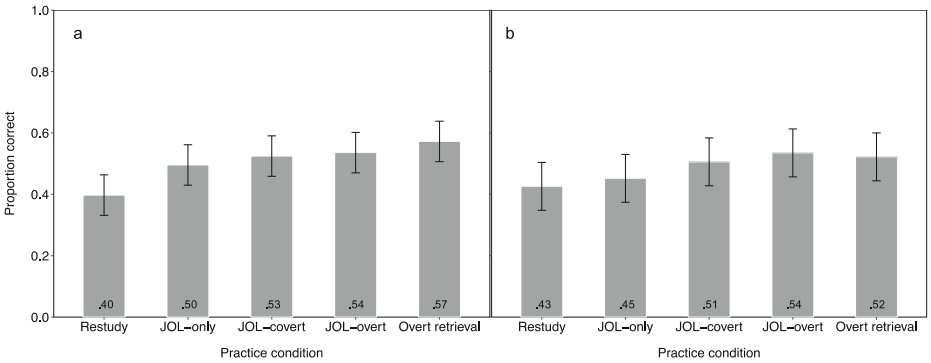

Fig. 3 Final recall performance in Experiments 2a (a) and 2b (b). Error bars indicate 95% confidence intervals

Reaction times The middle section of Table 1 shows reaction time measured in seconds by the first keypress of subjects for the JOL-only, JOL-covert, JOL-overt, and overt retrieval conditions. The left column indicates reaction times for JOLs, whereas the right column indicates reaction times for overt retrieval in the JOL-overt and overt retrieval conditions. In the overt retrieval condition, subjects did not provide a response in 26% of the trials. A repeated measures ANOVA replicated a main effect of practice on reaction time, $F(2.16,112.30) = 36.61$, $p < .001$, $\eta^2_p = .41$. As in Experiment 1, pairwise comparisons showed that subjects took less time to start typing target words than typing their JOL ratings in all JOL conditions (left column), $ps < .001$. When overt retrieval was compared to JOL-overt, JOL-covert, and JOL-only conditions, the effect sizes for these comparisons were $d = 1.36$, $d = 1.01$, and $d = .79$, respectively. Of more interest were the comparisons between the JOL-only and JOL-overt conditions. Subjects took longer to provide JOL ratings in the JOL-overt condition than the JOL-only condition (left column), $p = .004$, $d = .49$, whereas they took less time to overtly retrieve in the JOL-overt condition than the overt retrieval condition (right column), $t(52) = 5.73$, $p < .001$, $d = .79$. These results supported the truncated search hypothesis because the reaction times of JOL-overt and JOL-only conditions differed in the expected direction.

Covert retrieval prompt To examine whether JOLs elicited covert retrieval, in the JOL-covert condition, subjects were asked whether they remembered the target word while making JOLs. Subjects responded *yes* on half of the trials (.50), whereas they reported not retrieving the target word (i.e., *no* response) for the other half of the trials. This outcome suggested that at least half of the time they engage in covert retrieval while making JOLs; however, this proportion was still lower than the proportion of overt responses in the JOL-overt condition (.85) and overt retrieval condition (.74). The conditional analyses on *yes/no* responses showed that on the two-day delayed final test, subjects recalled more target words for *yes* responses ($M = .71$, $SE = .03$) than for *no* responses ($M = .34$, $SE = .03$), $t(52) = 10.51$, $p < .001$, $d = 1.45$. Not surprisingly, they also made higher JOL ratings for items to which they gave *yes* responses ($M = 75.09$, $SE = 2.07$) than *no* responses ($M = 19.46$, $SE = 1.64$), $t(52) = 23.12$, $p < .001$, $d = 3.17$. The calibration plot for *yes* and *no* responses is reported in the SOM.

Practice cued recall test For each subject in the overt retrieval and JOL-overt condition, the proportion of correct recall was calculated for the practice cued recall test. The bottom section of Table 2 shows the mean recall on the practice cued recall test for overt retrieval and JOL-overt conditions, with recall on the final test presented in Fig. 3. A 2 (practice condition) × 2

(test) repeated measures ANOVA revealed a main effect of test, $F(1,52) = 5.07$, $p = .029$, $\eta^2_p = .09$. On average, subjects' recall showed a reliable increase from the practice test ($M = .46$, $SE = .03$) to the two-day delayed final test ($M = .56$, $SE = .03$) due to feedback given on the practice test. The main effect of practice condition did not reach conventional levels of significance, $F(1,52) = 3.69$, $p = .062$, $\eta^2_p = .07$.The interaction was not reliable, $F < 1$.

Discussion

Employing a within-subjects design, we found that making delayed JOLs reliably improved recall relative to restudy. However, making only delayed JOLs did not produce similar benefits in recall relative to overt retrieval. In fact, recall after making delayed JOLs was reliably lower than recall after overt retrieval. These findings support the truncated search hypothesis and not the strictest version of the covert retrieval hypothesis. That is, people did not always engage in covert retrieval attempts during delayed JOLs; instead, our results indicate that they truncated their search of the target words when making delayed JOLs, which produced poorer final recall compared to the overt retrieval condition. Critically, truncated retrieval induced by delayed JOLs still enhanced learning relative to no retrieval (i.e., restudy) in Experiment 2a.

The reaction time results further supported the truncated search hypothesis. By employing a JOL condition that also involved overt retrieval (i.e., the JOL-overt condition), we found that subjects spent more time on delayed JOLs followed by overt retrieval than on delayed JOLs alone. This outcome suggests that students engaged in full-blown covert retrieval during delayed JOLs when delayed JOLs were followed by overt retrieval, whereas they truncated their search (or engaged in truncated retrieval) when they made delayed JOLs. Two findings support this conclusion. First, subjects reported retrieving the target only 50% of the time in the JOL-covert condition, whereas in the JOL-overt condition they were able to recall the target item 79% of the time. Second, the overt retrieval reaction times also affirmed the truncated search results: Overt retrieval following delayed JOLs (i.e., JOL-overt) took reliably less time than overt retrieval alone, indicating that subjects often retrieved target words at the time of delayed JOLs when delayed JOLs were followed by overt retrieval.

Critically, these results do not indicate that subjects never engaged in covert retrieval attempts while making delayed JOLs. When we asked subjects whether they retrieved a word while making delayed JOLs, they answered *yes* half of the time. This suggests that subjects engaged in covert retrieval at least in half of the trials. In addition, even when subjects made *no* responses, they may still have attempted to retrieve the target word while making JOLs but failed. Furthermore, asking the covert retrieval question after delayed JOLs did not produce a reliable benefit in recall relative to only making delayed JOLs alone (though it did numerically), suggesting that subjects were already engaging in covert retrieval during some JOL-only trials.

Experiment 2b

Because of the design and material differences between Experiments 1 and 2a, we wanted to ensure the replicability of our results in Experiment 2a. Therefore, in Experiment 2b, we conducted a direct replication of Experiment 2a. Our main interest was the reaction time results.

Method

Subjects Forty-nine undergraduates were recruited for course credit or payment ($10). After excluding four subjects who did not complete Session 1, and three who did not complete Session 2, 41 subjects remained.

Materials and design The material and design were identical to Experiment 2a.

Procedure In Experiment 2b, the data collection was online due to the Covid-19 pandemic. Subjects received a link for Session 1 after they signed up for the experiment and completed it in their own time at their preferred private setting. Subjects emailed a verification code at the end of the experiment to the experimenter to mark the date they completed Session 1. Subjects then received the link for Session 2 in the morning 2-days after their completion of Session 1 and were given 24 h to complete it. Due to the 24-h completion period, the time difference between Session 1 and Session 2 was not exactly 2-days as in Experiment 2a. Both sessions were identical to those in Experiment 2a.

Results

Final cued recall test Figure 3b shows that proportion of items recalled showed a gradual ncrease across conditions from restudy to overt retrieval. A one-way repeated measures ANOVA with five levels confirmed a main effect of practice condition, $F(3.13,125.19) = 5.23$, $p = .002$, $\eta^2_p = .12$. Pairwise comparisons revealed that restudy produced reliably poorer recall than the JOL-overt and overt retrieval conditions, $ps < .05$, whereas the recall difference between the restudy and JOL-covert conditions did not reach statistical significance, $p = .05$. The effect sizes for these comparisons were $d = .49$, $d = .47$, and $d = .39$, respectively. Critically, restudy yielded similar recall to the JOL-only condition in Experiment 2b, $p = .90$, $d = .11$. The JOL-only condition produced poorer recall than the JOL-overt condition, $p = .04$, $d = .48$, whereas the recall difference between the JOL-only and overt retrieval conditions did not reach conventional levels of significance although it was in the expected direction, $p = .12$, $d = .42$. Other comparisons were not reliable [JOL-only vs. JOL-covert, $p = .35$, $d = .33$; JOL-overt vs. JOL-covert, $p = .85$, $d = .20$; overt retrieval vs. JOL-covert, $p = .98$, $d = .10$ overt retrieval vs. JOL-overt, $p = .99$, $d = .09$].

Reaction times The bottom section of Table 1 shows reaction times measured in seconds by the first keypress of subjects for the JOL-only, JOL-covert, JOL-overt, and overt retrieval conditions. In the overt retrieval condition, subjects did not provide a response in 22% of the trials. A repeated measures ANOVA replicated a main effect of practice on reaction time, $F(3,117) = 61.16$, $p < .001$, $\eta^2_p = .61$. As in Experiment 2a, pairwise comparisons showed that subjects took less time to start typing target words than typing their JOL ratings in all JOL conditions (left column), $ps < .001$. When overt retrieval was compared to JOL-overt, JOL-covert, and JOL-only conditions, the effect sizes for these comparisons were $d = 1.78$, $d = 1.48$, and $d = 1.32$, respectively. Furthermore, subjects took longer to provide JOL ratings in the JOL-overt condition than the JOL-only and JOL-covert conditions (left column), $p < .001$, $d = .74$, and $p = .007$, $d = .80$, respectively. They took less time to overtly retrieve in the JOL-overt condition than the overt retrieval condition (right column), $t(39) = 7.62$, $p < .001$, $d = 1.20$. These results replicated Experiment 2a and supported the truncated search hypothesis.

Covert retrieval prompt In Experiment 2b, subjects responded *yes* on .53 of the trials, whereas they responded *no* for the remaining trials (.47). This proportion was lower than proportion of overt responses in the JOL-overt condition (.92) and overt retrieval condition (.78). The conditional analyses on *yes/no* responses showed that subjects recalled more target words for *yes* responses ($M = .76$, $SE = .04$) than for *no* responses ($M = .26$, $SE = .04$), $t(40) = 11.04$, $p < .001$, $d = 1.72$. They made higher JOL ratings for items to which they gave *yes* responses ($M = 78.20$, $SE = 2.75$) than *no* responses ($M = 20.02$, $SE = 2.19$), $t(40) = 18.41$, $p < .001$, $d = 2.88$. The calibration plot for *yes* and *no* responses is reported in the SOM.

Practice cued recall test The bottom section of Table 2 shows mean recall on the practice cued recall test for overt retrieval and JOL-overt conditions, with recall on the final test presented in Fig. 3b. A 2 (practice condition) × 2 (test) repeated measures ANOVA revealed a main effect of test, $F(1,40) = 9.16$, $p = .004$, $\eta^2_p = .19$. As with Experiment 2a, recall increased from the practice test ($M = .48$, $SE = .04$) to the final test ($M = .53$, $SE = .04$). The main effect of practice condition and the interaction were not reliable, $F(1,40) = 1.43$, $p = .24$, $\eta^2_p = .03$, and $F(1,40) = 1.60$, $p = .21$, $\eta^2_p = .04$. respectively.

Discussion

In Experiment 2b, we replicated the reaction time results of Experiment 2a in a less controlled environment. These results provided further evidence for the truncated search hypothesis of delayed JOLs: When retrieval was promoted in the JOL-overt condition, subjects took longer to make JOLs, suggesting that in JOL-only conditions, they might have truncated their search for target words prematurely. Critically, in Experiment 2b, making delayed JOLs alone did not improve retention relative to restudy (i.e., a recall difference of .02). This finding was in contrast to Experiment 2a that reported a reliable difference between restudy and delayed JOLs (i.e., a recall difference of .10). The online nature of Experiment 2b might have led subjects to engage in fewer retrieval attempts while making delayed JOLs unless provided with a prompt or retrieval attempt. Nonetheless, this finding was still in line with the truncated search hypothesis. That is, in Experiment 2b, subjects might have prematurely stopped their search for most of the target words while making delayed JOLs, which in turn might have eliminated the delayed JOL advantage observed in Experiment 2a. Combined, the results of Experiments 2a and 2b indicate that the effect of delayed JOLs on retention might be dependent on the amount of truncated search and covert retrieval attempts. That is, if most search lead to full-blown retrieval attempts, delayed JOLs might enhance retention, and if they end prematurely, the advantage of delayed JOLs over restudy might disappear.

General discussion

The main purpose of the current study was to examine how delayed JOLs affected paired-associate learning by comparing cue-only delayed JOLs to overt retrieval, restudy, and various other delayed JOL conditions (i.e., the cue-target delayed JOL condition in Experiment 1, and the JOL-covert and JOL-overt conditions in Experiments 2a and 2b). We also explored three potential mechanisms for the effects of delayed JOLs on later recall: 1) the covert retrieval hypothesis, 2) the truncated search hypothesis, and 3) non-retrieval reactivity. We first review the effects of delayed JOLs on recall and then discuss their implications for each proposed mechanism.

Delayed JOLs have been traditionally contrasted to immediate JOLs, seldom to retrieval practice, and hardly ever to restudy for memory effects (Jönsson et al. 2012; Rhodes and Tauber 2011; Tauber et al. 2015). In Experiment 1, we compared cue-only delayed JOLs to restudy, overt retrieval practice, and cue-target JOLs in between-subjects designs with a two-day cued recall test. Although we obtained a recall difference of .09 between overt retrieval and restudy conditions, it was not reliable. Of the main interest, recall for cue-only delayed JOLs was in between restudy and overt retrieval conditions but did not reliably differ from either. A sensitivity analysis confirmed that Experiment 1 suffered from relatively low power to detect small effects. Therefore, in Experiments 2a and 2b, we employed a within-subjects design and compared cue-only delayed JOLs to restudy, overt retrieval, delayed JOLs followed by a covert retrieval question, and delayed JOLs followed by overt retrieval. In Experiment 2a, all JOL conditions as well as overt retrieval practice reliably enhanced learning relative to restudy, indicating that delayed JOLs might also benefit learning; however, the advantage of delayed JOLs over restudy disappeared in Experiment 2b. The discrepancy between these findings might have stemmed from different levels of truncated search elicited in the two experiments. Nonetheless, in both experiments, recall showed a gradual increase from the delayed JOL alone condition to the overt retrieval condition (Fig. 3). Among these four conditions, the only reliable recall differences were between overt retrieval practice and delayed JOL alone in Experiments 2a and 2b and between delayed JOLs followed by overt retrieval and delayed JOL alone in Experiment 2b. These results were in line with previous findings reported by Putnam and Roediger (2013).

We first evaluate the covert retrieval hypothesis in light of our results. The strict form of the covert retrieval hypothesis proposes that all cue-only delayed JOLs evoke attempts at covert retrieval (Nelson and Dunlosky 1991; Spellman and Bjork 1992). If this is the case, according to previous studies that compared covert and overt retrieval (Putnam and Roediger 2013), delayed JOLs should enhance learning relative to restudy and as much as overt retrieval. Our results, however, contradicted this strong form of the covert retrieval hypothesis. In Experiment 1, although recall for cue-only delayed JOLs did not reliably differ from that of overt retrieval, it was also not reliably higher than restudy. Furthermore, when we employed a more powerful design in Experiment 2, delayed JOLs yielded poorer recall than overt retrieval. Therefore, we conclude that a strict covert retrieval hypothesis cannot explain the memory effects of delayed JOLs, a conclusion supported by previous studies as well (Son and Metcalfe 2005; Tauber et al. 2015).

Instead, our recall results supported the truncated search hypothesis, which states that compared to overt retrieval practice, people put less effort into retrieval during delayed JOLs, and thus truncate their retrieval attempts prematurely. Due to these subtle retrieval differences, delayed JOLs should produce poorer learning than retrieval practice but they may or may not improve learning relative to restudy. In line with the truncated search hypothesis, in Experiments 1, 2a, and 2b, recall in delayed JOLs was consistently in between restudy and overt retrieval conditions. Furthermore, reaction time results from Experiments 2a and 2b also supported the truncated search hypothesis: When delayed JOLs were followed by overt retrieval (i.e., JOL-overt condition), people took longer to make JOL ratings relative to the JOL-only and JOL-covert conditions, and they took less time to retrieve overtly compared to the overt retrieval condition. These findings indicated that asking people to engage in overt retrieval after delayed JOLs promoted full-blown retrieval attempts during delayed JOLs, and these retrieval attempts increased reaction times for delayed JOLs. If delayed JOLs alone caused full-blown effortful retrieval as in retrieval practice, we would not have observed any reliable reaction time differences.

Lastly, in Experiment 1, we investigated whether memory effects were driven by non-retrieval reactivity, as has been reported for immediate JOLs. That is, making immediate JOLs alters how people study word pairs, or strengthens information used to make JOLs, which in turn affects the learning outcome (Mitchum et al. 2016; Soderstrom et al. 2015). To test this possibility for delayed JOLs, we compared a cue-target delayed JOL condition to a restudy condition in Experiment 1, because cue-target JOLs should not elicit any retrieval attempts, yet they can show other reactive effects as immediate JOLs. Previous studies that compared a study condition with immediate JOLs to a study condition without immediate JOLs have obtained JOL reactivity for related pairs on both immediate and delayed cued recall tests (Double et al. 2018; Janes et al. 2018; Soderstrom et al. 2015; Witherby and Tauber 2017). Although we observed a similar pattern, we did not find any reliable recall differences between cue-target JOLs and restudy conditions. This was the case even when these conditions in Experiment 1 had either greater or similar numbers of subjects as prior studies. Thus, we conclude that the reactive effects of delayed JOLs cannot be solely explained through the non-retrieval mechanisms responsible for immediate JOL reactivity.

Our findings are similar to those reported by Tauber et al. (2015), who found a reliable recall difference between delayed JOLs and overt retrieval in a cross-experimental analysis. However, there are a few differences that warrant attention. First, Tauber et al. compared delayed JOLs to immediate JOLs and concluded that delayed JOLs did not enhance recall on a delayed test because "if participants sometimes truncate their search of memory for responses when making delayed JOLs, then less elaborative information would be activated during encoding and hence would not be available to support recall during the criterion test" (p. 261). Nonetheless, immediate JOLs might be a biased comparison group due to their possible reactive effects. To overcome this possibility, we compared delayed JOLs to restudy and found that delayed JOLs might benefit learning relative to passive restudy of word pairs. The benefit of delayed JOLs, however, was not consistent across Experiments 2a and 2b. This outcome suggested that the effect of delayed JOLs on retention might be dependent on what type of processes were dominantly elicited by delayed JOLs (i.e., truncated search or full-blown retrieval attempts). That is, if delayed JOLs mostly elicit truncated retrieval (or less elaborative information) as suggested by Tauber et al., they might not boost learning. Second, in Tauber et al., reaction times for the first keypress in the retrieval practice condition were longer than reaction times in the delayed JOL condition; however, in two experiments, we consistently found the opposite pattern (i.e., longer or equal reaction times for delayed JOLs). These results might have stemmed from methodological differences between Tauber et al. and our study. In Tauber et al., retrieval practice and delayed JOLs were self-paced, whereas, in the present experiments, subjects had 8-s to type the target word or the JOL rating. Therefore, subjects might have felt selectively pressured to start typing earlier during overt retrieval practice due to the time limit, because typing a word is more likely to take longer than typing a number between 0 and 100. Nonetheless, when we compared the same tasks across JOL-only and JOL-overt conditions, our reaction time results supported the truncated search hypothesis.

Although our results supported the truncated search hypothesis, they also suggested that people do not always prematurely truncate their search of target words. That is, people still engage in covert retrieval while making delayed JOLs. In fact, in Experiments 2a and 2b, subjects reported recalling the target word while making delayed JOLs around 50% of the time (i.e., a *yes* response). Critically, asking subjects whether they recalled a target word during the time of JOLs was a conservative measure of covert retrieval, because when they reported not recalling the target word (i.e., a *no* response), they might have still unsuccessfully attempted to

retrieve it. Nonetheless, the reported proportion of covert retrievals was lower than overt responses in the JOL-overt and overt retrieval conditions in both experiments. Although our results indicate that the covert retrieval hypothesis, which assumes people make delayed JOLs through a single-process retrieval mechanism, was false, we demonstrated that people engage in both covert retrieval and truncated search while making delayed JOLs.

When do delayed JOLs lead to covert retrieval and when do they lead to truncated search? Son and Metcalfe (2005) proposed a two-stage processing model for delayed JOLs according to which the first stage is a quick preliminary assessment and the second (conditional) stage is retrieval attempt (i.e., covert retrieval). To test the two-stage model, Son and Metcalfe asked subjects either to only make delayed JOLs or to attempt retrieval and then make delayed JOLs. They then compared latencies for delayed JOLs alone to retrieval latencies across JOL levels. If all delayed JOL ratings evoked retrieval (i.e., the single process), the function of delayed JOLs should have been identical to that of retrieval. When subjects attempted to retrieve the targets first and then made JOLs, low JOLs corresponded to longer retrieval latencies and high JOLs corresponded to shorter retrieval latencies, revealing a decreasing function across JOL levels. When subjects made only delayed JOLs, both low and high JOLs corresponded to shorter JOL latencies, whereas JOL ratings in the middle of the scale corresponded to longer JOL latencies, demonstrating an inversed U-shaped function across JOL levels. In other words, for high JOLs, JOL and retrieval latency functions paralleled each other, suggesting similar processes between delayed JOLs and retrieval for high JOLs. However, for low JOLs, the latency functions went in the opposite direction, suggesting different processes between delayed JOLs and retrieval for low JOLs. Son and Metcalfe (2005) concluded that people do not spontaneously engage in retrieval while making delayed JOLs, but instead they first quickly assess cue familiarity and if it is not familiar enough, they terminate (or truncate) the search for the target and they provide a fast low JOL rating. If the cue is familiar enough, then they engage in a full-blown retrieval attempt and if they retrieve a target, they provide a high JOL rating. Therefore, according to the two-stage model, truncated search occurs for low delayed JOLs whereas covert retrieval occurs for high delayed JOLs. Metcalfe and Finn (2008) further supported the two-stage model by demonstrating that when delayed JOLs were not speeded, they showed effects of both cue familiarity and target retrievability; however, when they were speeded, they only showed effects of cue familiarity, but not target retrievability. The present findings are in line with the two-stage model of delayed JOLs. Indeed, in Experiment 2, we found that when subjects in the JOL-covert condition retrieved a word (*yes* response) while making JOLs, they made higher JOL ratings, whereas when they did not (*no* response), they made lower JOL ratings.

The current study found that making cue-only delayed JOLs did not improve learning as much as overt retrieval practice. Our results demonstrate that subjects engaged in both truncated retrieval and covert retrieval while making delayed JOLs. These results further supported the truncated search hypothesis, which assumes that during delayed JOLs people terminate their search for target words prematurely. Our findings were also compatible with Son and Metcalfe's (2005) two-stage process model of delayed JOLs that proposes a quick initial cue assessment stage during JOLs before an attempt is made at covert retrieval. Future research should examine the circumstances that promote covert retrieval and truncated retrieval during delayed JOLs to advance our understanding of their mechanisms in aiding recall.

Appendix

JOL Results

Experiment 1

The top section of Table 3 shows average JOL ratings for related and unrelated word pairs across each JOL condition. A 2 (relatedness) × 2 (JOL type) repeated measures ANOVA revealed a main effect of relatedness, $F(1,82) = 397.74$, $p < .001$, $\eta^2_p = .83$ and a reliable interaction, $F(1,82) = 28.26$, $p < .001$, $\eta^2_p = .26$. The main effect of JOL type did not reach conventional levels of significance, $F(1,82) = 2.88$, $p = .094$, $\eta^2_p = .03$. Pairwise comparisons revealed that for related word pairs, cue-target JOLs did not statistically differ from cue-only JOLs, $p = .76$, $d = .23$, whereas for unrelated word pairs, cue-only JOLs were higher than cue-target JOLs, $p < .001$, $d = .84$.

Experiment 2a

The middle section of Table 3 shows average JOL ratings for weakly related word pairs across each JOL condition. A one-way repeated measures ANOVA was conducted to test any differences. The main effect of practice condition did not reach conventional levels of statistical significance, $F(1.78,92.74) = 3.20$, $p = .051$, $\eta^2_p = .06$. Because we did not have any a priori hypotheses about JOL ratings, we did not conduct pairwise comparisons. The descriptive results suggested that JOLs in the JOL-overt condition might be lower than the other two JOL conditions, possibly because in the JOL-overt condition subjects were engaging in full-blown retrieval.

Experiment 2b

The bottom section of Table 3 shows average JOL ratings for weakly related word pairs across each JOL condition in Experiment 2b. A one-way repeated measures ANOVA was conducted to test any differences. The main effect of practice condition was not significant, $F(2,80) = 1.87$, $p = .161$, $\eta^2_p = .05$.

Table 3 JOLs Across Conditions and Experiments

Experiment 1				
Condition	Related		Unrelated	
	M	*SD*	*M*	*SD*
Cue-target JOL	70.9	14.0	29.5	16.7
Cue-only JOL	67.5	16.5	43.5	18.1
Experiment 2a				
Condition	Weakly related			
	M	*SD*		
JOL-only	45.5	17.4		
JOL-covert	46.7	17.9		
JOL-overt	41.7	19.8		
Experiment 2b				
JOL-only	50.7	28.7		
JOL-covert	50.0	20.1		
JOL-overt	46.3	21.9		

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11409-021-09260-0.

Availability of data and material The datasets and material supporting the conclusions of this article are available in the Open Science Framework repository, osf.io/yaguz.

Code availability Not applicable.

Authors' contributions ET conceived the idea for the study and HR supervised throughout the project. ET programmed the research, carried it out, and analyzed the results. ET and HR were authors of the report. Both authors approve the final version.

Funding The research was funded by the Psychological and Brain Sciences Department's Departmental Research Fund.

Declarations

Ethics approval and consent to participate The study was approved by Washington University Institutional Review Board. At the beginning of the study, informed consent to participate was obtained from each subject. In addition, each subject was debriefed at the end of the study.

Consent for publication Not applicable.

Conflict of interest/competing interests The authors have no competing interests with respect to their authorship or the publication of this article.

References

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods, 39*(3), 445–459.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods, 46*(3), 904–911.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132*, 354–380.

Dunlosky, J., & Nelson, T. O. (1992). Importance of the kind of cue for judgments of learning (JOL) and the delayed-JOL effect. *Memory & Cognition, 20*(4), 374–380.

Dunlosky, J., & Nelson, T. O. (1994). Does the sensitivity of judgments of learning (JOLs) to the effects of various study activities depend on when the JOLs occur? *Journal of Memory and Language, 33*(4), 545–565.

Double, K. S., Birney, D. P., & Walker, S. A. (2018). A meta-analysis and systematic review of reactivity to judgements of learning. *Memory, 26*(6), 741–750.

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.

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*(6), 2356–2364.

Jönsson, F. U., Hedner, M., & Olsson, M. J. (2012). The testing effect as a function of explicit testing instructions and judgments of learning. *Experimental Psychology, 59*, 251–257.

Kelemen, W. L., & Weaver III, C. A. (1997). Enhanced memory at delays: Why do judgments of learning improve over time? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*(6), 1394-1409.

McDermott, K. (2021). Practicing retrieval facilitates learning. *Annual Review of Psychology., 72*, 609–633.
Metcalfe, J., & Finn, B. (2008). Familiarity and retrieval processes in delayed judgments of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(5), 1084–1097.
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*(2), 200–219.
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, 1–14.
Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect". *Psychological Science, 2*(4), 267–271.
Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers, 36*(3), 402–407.
Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. *Psychology of Learning and Motivation, 26*, 125–173.
Putnam, A. L., & Roediger, H. L. (2013). Does response mode affect amount recalled or the magnitude of the testing effect? *Memory & Cognition, 41*(1), 36–48.
Rhodes, M. G., & Tauber, S. K. (2011). The influence of delaying judgments of learning on metacognitive accuracy: A meta-analytic review. *Psychological Bulletin, 137*(1), 131–148.
Rhodes, M. G. (2016). Judgments of learning: Methods, data, and theory. *The Oxford handbook of metamemory, 1*, 65–80.
Roediger, H. L., & DeSoto, K. A. (2016). Recognizing the presidents: Was Alexander Hamilton president? *Psychological Science, 27*(5), 644–650.
Roediger III, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*(3), 249–255.
Smith, M. A., Roediger III, H. L., & Karpicke, J. D. (2013). Covert retrieval practice benefits retention as much as overt retrieval practice. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(6), 1712–1725. https://doi.org/10.1037/a0033569.
Son, L. K., & Metcalfe, J. (2005). Judgments of learning: Evidence for a two-stage process. *Memory & Cognition, 33*(6), 1116–1129.
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*(5), 315–317.
Soderstrom, N. C., Clark, C. T., Halamish, V., & Bjork, E. L. (2015). Judgments of learning as memory modifiers. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*, 553–558. https://doi.org/10.1037/a0038388.
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*(4), 254–263. https://doi.org/10.1027/1618-3169/a000296.
Tekin, E., & Roediger III, H. L. (2020). Reactivity of judgments of learning in a levels-of-processing paradigm. *Zeitschrift für Psychologie, 228,* 278-290.
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*(4), 496–503.