Metacognition and Learning (2022) 17:1057–1077
https://doi.org/10.1007/s11409-022-09307-w

Metacognitive judgments can potentiate new learning: The role of covert retrieval

Veit Kubik[1] · **Kenneth Koslowski**[2] · **Torsten Schubert**[3] · **Alp Aslan**[4]

Received: 29 June 2021 / Accepted: 16 May 2022 / Published online: 8 June 2022

Abstract
Interim tests of previously studied information can potentiate subsequent learning of new information, in part, because retrieval-based processes help to reduce proactive interference from previously learned information. We hypothesized that an effect similar to this forward testing effect would also occur when making judgments of (prior) learning (JOLs). Previous research showed that making JOLs likely prompts covert retrieval attempts and thereby enhances memory, specifically when providing only parts of previously studied information. This study examined the forward effect of different types of JOLs (i.e., with complete or partial prior study information available) on subsequent learning of new materials, compared to restudy and retrieval practice. In a between-subjects design, participants ($N = 161$) consecutively studied five lists of 20 words with the aim to recall as many of them on a final cumulative recall test. After the presentation of each of the first four lists, participants either restudied the list, made JOLs with complete words, made JOLs with word stems, or they were tested on word stems. Compared to restudy, practicing retrieval and making JOLs with word stems, but not JOLs with complete words, facilitated the List-5 interim recall performance and attenuated the number of intrusions from prior lists. The findings suggest that, similar to overt retrieval, making JOLs with incomplete information can enhance new learning to the extent that it elicits covert retrieval attempts.

Keywords Forward testing effect · Judgment of learning · Covert retrieval · Episodic memory · JOL reactivity

Testing one's memory (e.g., by practicing retrieval in a memory test) is a powerful study technique to boost learning. Prior research has shown that retrieval practice of learned information can enhance its long-term retention more than restudying or doing nothing (Roediger & Karpicke, 2006; for reviews, see McDermott, 2021; Yang et al., 2021). For example, participants initially studied a prose text and then either practiced it by restudy or by retrieval in a free-recall test (Roediger & Karpicke, 2006). Although restudy led to a better immediate recall performance, retrieval practice led to better recall of text passages after 2 days and 1 week, indicating a reduced rate of forgetting. This *backward testing*

✉ Veit Kubik
veit.kubik@uni-bielefeld.de

Extended author information available on the last page of the article

effect (BTE) of previously learned information is of robust nature as it generalized to a wide variety of materials (e.g., visuospatial information, Carpenter & Pashler, 2007; performed actions, Kubik et al., 2018; for an overview, see Karpicke & Aue, 2015; Rowland, 2014) and settings (e.g., in the lab and classroom environments, Adesope et al., 2017; Yang et al., 2021).

Recent research has demonstrated that retrieval can also indirectly enhance memory by potentiating subsequent restudy of previously learned information (Arnold & McDermott, 2013; see also Kubik et al., 2016; Tempel & Kubik, 2017). Even more compelling, there has been a burgeoning interest in another indirect, but future-oriented benefit of retrieval practice, that is, test-potentiated learning of *new* (i.e., not-yet-presented) information (Chan et al., 2018a) or also called the *forward testing effect* (FTE, for recent overviews, see Chan et al., 2018b; Pastötter & Bäuml, 2014; Yang et al., 2018), which is the focus of the present study. As a typical experiment to examine the FTE, Szpunar et al. (2008) had participants study five separate lists of 20 words in a multiple-list paradigm, with the instruction to freely remember all of them for a final cumulative recall test. After the presentation of each of the first four lists, participants either recalled the items from the previous list (retrieval group), studied them again (restudy group), or completed a distractor task (distractor group). After studying List 5, participants received an immediate free-recall test, in which only items from List 5 were to be recalled. Critically, the retrieval group recalled more List-5 items and with fewer prior-list intrusions, compared to the restudy and distractor groups. This indirect, mnemonic benefit of retrieval practice is remarkable because it is oriented towards learning of new (i.e., not-yet-presented) information. Studies have demonstrated the FTE with various study materials (e.g., single words, paired associates, prose passages, and videos), memory tests (e.g., free recall, cued recall; for a metaanalysis, see Chan et al., 2018b; for an overview, see Yang et al., 2018), and also study populations (students, Szpunar et al., 2008; children, Aslan & Bäuml, 2016; Dang et al., in press; elderly people, Pastötter & Bäuml, 2019; clinical patients, Pastötter et al., 2013).

Multiple mechanisms can contribute to the FTE and their relative contributions likely vary as a function of the experimental setting, such as the study materials (Kliegl & Bäuml, 2021; Yang et al., in press). Using unrelated single-word lists as materials in this study, the most prominent mechanisms are release from proactive interference and reset-of-encoding. According to the *release-from-proactive-interference account* (Bäuml & Kliegl, 2013; Szpunar et al., 2008), retrieval practice alters the episodic context during study and creates a unique set of contextual features that helps to discriminate items and lists from each other during memory search of newly studied information (Lehman et al., 2014). This list segregation reduces the proactive interference during retrieval from previously learned information, resulting in fewer intrusions from previous lists. The *reset-of-encoding account* (Pastötter et al., 2011; Pastötter et al., 2018) is also based on the idea that retrieval practice drives episodic context change and list segregation "resets" encoding processes for subsequent learning of new information; memory load and inattention are then reduced, specifically for the first new items of subsequently studied information (Pastötter et al., 2018; see also Dang et al., in press).

The learning benefits of retrieval attempts emerge even when retrieval occurs covertly and an answer is not overtly produced. Covert retrieval can occur when we need to predict our own test performance in metacognitive judgments. As a consequence, when we assess our own learning, we may be able to enhance the memory performance of the particular information about which the metacognitive judgment was made due to attempting to covertly retrieve that information (Dougherty et al., 2018). Evidence that making metacognitive judgments enhances memory has been shown in studies employing judgments of

learning (JOLs; Soderstrom et al., 2015). In these studies, typically, a list of paired associates (e.g., *nail–hammer*) is presented, and participants are then provided with only the cue words (e.g., *nail*) from the paired associates to judge the likelihood of each pair that they would recall the target words (e.g., *hammer*) on a future memory test (i.e., *cue-only JOLs*). A cued-recall memory test follows after a short and long delay. Making cue-only JOLs can enhance long-term retention of paired associates relative to restudying the pairs (Jönsson et al., 2012; Putnam & Roediger, 2013). To account for this JOL-related memory benefit (Dunlosky & Nelson, 1992; Nelson & Dunlosky, 1991; Spellman & Bjork, 1992), covert retrieval attempts for the absent targets are assumed to occur prior to making the metacognitive assessment. Therefore, making cue-only JOLs can enhance learning via retrieval processes similar to those assumed to be involved in the BTE (cf. Rhodes & Tauber, 2011). In contrast, making JOLs with intact study information available—so called *cue–target JOLs*—should also enhance learning but less likely induce covert retrieval attempts. In a typical experiment involving cue–target JOLs, a list of paired associates is once presented, and participants are instructed to judge the likelihood to recall the target word in a memory test, with the paired associate present (e.g., *nail–hammer*). Critically, as the full information is present while making the judgments, participants are unlikely to engage in covert retrieval attempts from long-term memory prior to the metacognitive assessment. Nonetheless, making cue–target JOLs, compared to not making them, can enhance immediate recognition (Li et al. 2021; Myers et al., 2020) as well as later recall of target words on cued-recall tests after both brief delays (Myers et al., 2020; Soderstrom et al., 2015; but see Mitchum et al., 2016) and long delays (Witherby & Tauber, 2017). Such positive memorial effects of JOLs that are unlikely to engender covert retrieval (for a metanalysis, see Double et al., 2018) are likely due to the judgment itself encouraging participants to elaborate on the presented information, thereby strengthening associations between cues and targets (Soderstrom et al., 2015; see also Myers et al., 2020). Prior studies demonstrating these memory benefits of JOLs exclusively examined the conditions under which making metacognitive judgments affected learning of previously studied information. If, however, making metacognitive judgments can involve covert retrieval, and retrieval can enhance subsequent learning of not-yet-presented information (the FTE), then making metacognitive judgments about previously studied information should also enhance subsequent learning of not-yet-presented information. The present research will test this hypothesis and examine the degree to which the hypothesized FTE from making metacognitive judgments depends upon the format of those metacognitive judgments. It is theoretically and practically important to examine response format as a potential moderator of the FTE, that is, the way in which retrieval-based learning is prompted based on a given cue: making JOLs versus providing an interim recall test with the request of an overt reproduction of the learned materials.

To our knowledge, there is only one study by Lee and Ha (2019) that has investigated the forward effect of metacognitive judgments in comparison to retrieval practice and restudy, and it was in the context of inductive learning of artists' painting styles. In three experiments, participants studied six painting–artists pairs, and they were subsequently instructed to restudy the pairs, to take an interim cued-recall test on them (i.e., with the painting as a cue to prompt the artists' name), or, critically, to make JOLs: dependent on the experiment, participants were instructed to make item-based JOLs (i.e., rate the likelihood that they would be able to identify the artist of a specific painting on a later test; Exp. 1), to make category-based JOLs on correctly classifying the artists' painting style on new paintings (Exp. 2), or to make global metacognitive judgments on their overall performance (by prompting monitoring, evaluating, and planning processes; Experiment

3). Following a subsequent study phase of paintings from six new artists, participants received a final multiple-choice test in which they were asked to classify new exemplars of the 12 previously studied artists. The results showed that making category-based JOLs and global metacognitive judgments enhanced the classification of new materials, but not when participants were asked to provide item-based judgments. One explanation for the lack of a forward effect of these item-based JOLs could be that they asked participants to assess their learning of particular painting–artists pairs and thereby do not match with the learning goal to abstract the artists' general painting styles from the studied exemplars. However, another explanation may be that both cue (i.e., the painting) and target (i.e., artist) during the item-based JOLs were available, reducing the likelihood that participants engage in covert retrieval practice.

The aim of the present study was to generalize this work on the forward effect of item-based JOLs to verbal learning and to examine whether this potential forward effect of metacognitive judgments is contingent upon the type of JOLs. Specifically, in the present study, participants learned lists of single words (e.g., *elephant*), and we examined the forward effects of JOLs with word stems versus JOLs with complete words in comparison to restudy and retrieval practice. Item-based JOLs with word stems (e.g., *ele_____?*) do not provide the complete words and thereby encourage learners more likely to engage in covert retrieval attempts, potentiating new learning; item-based JOLs with complete words (e.g., *elephant*), in contrast, resemble restudy and render it less likely for learners to engage in covert retrieval. Using a typical multiple-list learning paradigm (cf. Szpunar et al., 2008), participants studied 5 separate lists of 20 words (i.e., nouns) with the instruction to remember all of them for a final cumulative recall test. As interim activities after Lists 1 through 4, participants were instructed to either recall the words with word stems as retrieval cues, make JOLs based on words stems, make JOLs based on intact words, or restudy the words. After studying List 5, all participants were instructed to recall only List 5.

We pursued two main research questions. First, we examined whether monitoring one's learning for words via the act of making JOLs can enhance subsequent learning of new information. Based on prior studies with paired associates (Jönsson et al., 2012; Nelson & Dunlosky, 1991; Spellman & Bjork, 1992; Tauber et al., 2015; Tauber et al., 2018), we hypothesized that making item-based JOLs for individual words can also evoke covert retrieval attempts, depending on the type of metacognitive prompt. Critically, *making JOLs with word stems*—solicited with only word stems as a cue from the prior learning episode—should evoke more frequent and effortful covert retrieval attempts from long-term memory than making JOLs with the entire word present (e.g., Carroll & Shanahan, 1997; Scarrabelotti & Carroll, 1998; Undorf et al., 2016). However, Rhodes and Tauber (2011) found a smaller delayed JOL effect for single words than for paired associates, so there is the possibility that the findings on the beneficial forward memory effects of JOLs with paired associates may not generalize to studying single words. If making JOLs with word stems leads to covert retrieval attempts, we would expect that, similar to the FTE, making JOLs with word stems enhances recall performance of the criterial List 5 and decreases the rate of prior-list intrusions (i.e., recall of words from prior Lists 1–4 on the List-5 recall test). In contrast, making *JOLs with complete words*—with intact words from the prior learning episode provided—should induce less covert retrieval attempts that do not reliably enhance the incidental retrieval rate of spaced restudy (cf., Tullis & Benjamin, 2011). Thus, for JOLs with complete words, we would not expect frequent covert retrieval attempts and do not predict an FTE-type effect to occur with enhanced List-5 recall and a decreased rate of prior-list intrusions. Similar results were expected for the final cumulative

test recall of List 5 as a function of practice type but not for Lists 1–4 (see e.g., Pastötter et al., 2020).

Second, assuming participants silently test themselves before making JOLs with word stems, we examined making JOLs with word stems is as effective for enhancing new learning as overt retrieval practice (the FTE). One possible result is that overt retrieval practice engenders larger benefits on new learning as it more frequently leads to more complete retrieval attempts. According to the two-stage process theory (Son & Metcalfe, 2005), making JOLs involves a first pre-retrieval stage of a quick familiarity assessment. For example, when cues are not evaluated as sufficiently familiar, participants truncate the retrieval attempt and give a low JOL. Otherwise, making JOLs involves a second stage in which participants attempt to fully retrieve the item and base their metacognitive judgment on retrieval fluency and accuracy. Partially due to producing (e.g., typing out) the recalled items, overt retrieval may more often lead to complete retrieval attempts (i.e., including the second stage) compared to making JOLs with word stems. In the latter case, memory search may be more often prematurely terminated and less effortful (i.e., truncated retrieval attempts in case of high and low familiarity). This prediction is consistent with evidence from research investigating the backward memory effects of retrieval practice and JOLs on learning of previously studied information. Although the effect is small, previous research has shown that delayed overt retrieval practice enhances long-term retention compared to both making cue-only JOLs (Tauber et al., 2015; Tauber et al., 2018; Tekin & Roediger, 2021) and covert retrieval practice in paired associate learning (Jönsson et al., 2014; Sundqvist et al., 2017, Exp. 3 & 4; but see Smith et al., 2013).

However, there is similar research that has reached an opposite conclusion. Several studies on the BTE in paired associate learning have not revealed a reliable benefit of retrieval practice over making cue-only JOLs in long-term retention (Putnam & Roediger, 2013; Smith et al., 2013). Thus, there is also the possibility that there is no reliable benefit of overt retrieval versus covert retrieval associated with item-based JOLs with word stems and thus response format does not affect new learning. Potentially, making JOLs based on word-stem cues immediately evokes exhaustive retrieval of the single words, even when assessing the familiarity in the pre-retrieval stage (cf. Tauber et al., 2018). Thus, making JOLs with word stems (i.e., likely evoking covert retrieval) and overt retrieval practice may lead to similar forward benefits on subsequent learning of new information.

Methods

Participants

Sample size was calculated a priori by using G*Power (Version 3.1.9.2; Faul et al., 2007). To demonstrate a reliable FTE, we determined a sufficient sample size of $n = 40$ per group (i.e., $N = 160$ in total), using an $\alpha = 0.05$, a power of $1–\beta = 0.90$, and an effect size of $d = .67$. We assumed a smaller effect size as typically found in earlier studies (e.g., $g = .75$; Chan et al., 2018b) for the following reason: JOL groups were included in the factor *practice type* in addition to the restudy and the retrieval groups. Given the novelty of the present experimental design, the size of a potential forward effect of JOLs with word stems, and even more of JOLs with complete words, was unclear and potentially smaller. As a consequence, we increased the sample size to obtain sufficient power to find the main effect of practice type across all experimental groups of practice type and reliable differences

Table 1: Participants' age as a function of practice type and gender.

Practice Type	female				male			
	n	*M*	*SD*	[*Min; Max*]	*n*	*M*	*SD*	[*Min; Max*]
Restudy	34	23.79	5.77	[18; 40]	6	26.83	6.55	[19; 38]
JOLs with complete words	33	22.97	5.59	[18; 36]	8	26.00	5.35	[19; 34]
JOLs with word stems	34	23.21	5.62	[18; 38]	6	23.50	2.17	[21; 26]
Retrieval	34	23.44	5.02	[18; 36]	6	21.67	2.58	[18; 25]

between them (consistent with prior research employing a third experimental group in addition to restudy and retrieval practice, Lehman et al., 2014; $n = 36$ per group).

Similar to the planned sample size, $N = 166$ participants took part in the present study in exchange for movie vouchers or course credit. They were undergraduate students that all were fluent in German. Five participants were excluded from the data analysis due to technical errors or participants not following the instructions, leaving a final sample of 161 participants. For the group of JOLs with complete words stems, the experimental software did not save the interim JOL data for the first nine participants; however, the interim and final cumulative recall data of these participants were complete and were included in the final analyses. Participants were randomly assigned to the four groups of practice type ($n_{\text{restudy}} = n_{\text{JOLs-with-word-stems}} = n_{\text{retrieval}} = 40$; $n_{\text{JOLs-with-complete-words}} = 41$) displaying similar participants' characteristics such as gender ratio and age between groups of practice type (all *p*s > .80; Table 1).

Materials

One hundred unrelated concrete German words (i.e., nouns) were drawn from the CELEX database (Duyck et al., 2004; using software toolbox Wordgen v1.0; cf. Pastötter et al., 2011) and were used as learning materials. All words began with a unique word stem (i.e., nouns' first 2–3 letters, depending on the word length) so that they can be distinguished in the retrieval and JOL groups with word stems. Twenty words were randomly assigned to five word lists in each of six stimulus sets, which were counterbalanced across participants for each practice type. The materials of the present study are publicly available at Open Science Framework (https://osf.io/a5ufh/).

Design

A one-factorial between-subjects design was applied, with participants being randomly assigned to the four groups of practice type: restudy, JOLs with complete words, JOLs with word stems, and retrieval. The main dependent variables were interim test recall performance (i.e., percentage of correct recall) of the criterial List 5, internal intrusions during interim test recall of List 5 (i.e., number of recalled words from prior Lists 1–4), and external (i.e., extra-experimental) intrusions.

Procedure

The present study used a typical multi-list learning paradigm to study the FTE (cf. Pastötter et al., 2011; Szpunar et al., 2008) and was run with E-Prime Software (v.2.0; Psychology Software Tools, Pittsburgh, PA; Schneider et al., 2012). Prior to the start of the experiment, participants were informed about its general procedure. They were instructed that they would study five lists of words, and they were encouraged to attentively encode these lists for a final cumulative recall test at the end of the experiment, in which they were asked to recall as many words as possible from all lists. After studying each list, participants solved a distractor task, in which they counted backwards in steps of three for 30s. Following this distractor task, participants performed one out of the following four practice types: restudying the list (restudy group), making JOLs based on intact words (JOL group with complete words), making JOLs based on word stems (JOL group with word stems), or attempting retrieval cued by word stems and writing the intact word down (retrieval group). To diminish differential expectancy effects of response format across groups, participants were informed that these interim activities were randomly determined by the computer; in fact, they were dependent on practice type, and participants kept the same type of practice consistently across Lists 1–4 within each experimental group (for a similar procedure, see Pastötter et al., 2011; Szpunar et al., 2008).

Following this general instruction, participants commenced the *learning phase*, in which five lists of 20 words were presented.

During an initial study period, 20 words were sequentially presented in the center of the computer screen for 2s (with a variable interstimulus interval of 0.8−1.2s displaying a fixation cross) followed by a 30s backward-counting task (i.e., counting backwards from a randomly drawn three-digit number in steps of 3). After this distractor task, the different experimental groups engaged in their respective practice type as interim activities. That is, for Lists 1–4 participants were instructed to either study the words again (restudy group), to make JOLs based on intact words (i.e., JOL group with complete words), to make JOLs based on word stems (i.e., JOL group with word stems), or to recall the words cued in a word-stem based cued-recall test (i.e., retrieval group) by entering them on the key board. More specifically, in the retrieval group, participants had 6s per item to recall the rest of the words cued with the word stems, followed by a variable interstimulus interval of 0.8−1.2s that was identical in all experimental groups. In the restudy group, all words were presented again in a new random order to restudy each of them one at a time for 6s (followed by a variable interstimulus interval of 0.8−1.2s). In the JOL group with complete words, participants saw the intact words from the previous study list sequentially for 6s. Within this time period, they had to judge the probability from 0 – 100% that they would remember the correct words in a few minutes and type the respective numbers on the keyboard. In the JOL group with word stems, participants were provided with sequentially presented word stems while making the JOLs within 6s. To match the practice time to the learned materials, the word stems remained on the screen for 6s in the retrieval group and JOL group with word stems. For List 5, participants in all experimental groups studied 20 words. Following a 30s backward-counting distractor task, they completed a free-recall test; participants had 3 minutes to exclusively recall the words of the criterial List 5 by entering them using a keyboard in any preferred order.

A 5-minute *distractor phase* was interleaved between the learning and final test phase. Participants completed simple addition tasks (e.g., *18 + 39 = ____ ?*) as accurate and fast as possible in a self-paced manner. At the end of the experiment, participants underwent

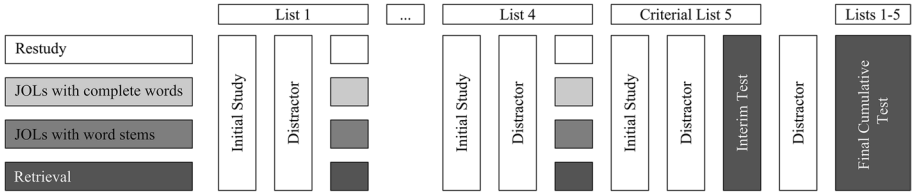

Fig. 1. Experimental procedure of the multiple-list paradigm. Participants studied five lists of 20 words. Initial study of Lists 1–4 was ensued by a 30-second distractor task (i.e., counting backwards in steps of 3). In the restudy group, complete words were presented, and participants studied Lists 1–4 again; in the two JOL groups, word stems or complete words were presented, respectively, and JOLs were solicited; in the retrieval group, word stems were presented, and participants attempted to retrieve the complete words. After initial study of List 5, all groups of practice type received a 30s backward-counting distractor task and then the 3-min interim test of free recall, in which they are instructed to only recall the words of the criterial List 5 in any preferred order. After a 5-min distractor task (i.e., solving simple addition problems), participants received a 7-min final cumulative test of free recall, in which they were instructed to attempt recalling words from all five lists in any preferred order.

a 7-min *final test phase*, in which a cumulative free-recall test was provided. Participants were instructed to attempt recalling as many of the 100 words learned from the five word lists by entering them using a keyboard in any preferred order. Overall, the processing of the tasks lasted about 40 minutes.

Scoring and data analysis

A strict scoring criterion was used in that only the test recall of original words from the study phase (but no synonyms) were scored as correct. However, it did not matter whether the words were written in single or plural forms and whether typographical errors occurred as long as they could be clearly identified. Test recall performance was thus measured as the proportion of correctly recalled words. In addition, the number of internal (prior-list) intrusions and external intrusions during interim test recall of List 5 were critical dependent variables.

For omnibus analyses of means, we ran and reported various ANOVAs using practice type as between-subjects and list as within-subjects factors. In cases when the assumption of sphericity was violated, the reported numbers were calculated using a Huynh–Feldt correction. If assumptions of the classical parametric ANOVA were in parts not met (e.g., normality, homoscedasticity), we checked the robustness of the results by also running rank-based ANOVA-type statistics (Brunner et al., 2017; see also Erceg-Hurn & Mirosevich, 2008) via the rankFD function and R package. Only consistent results of these performed multiverse analyses (Steegen et al., 2016) were reported in this study. To test the prespecified predictions of central interest, planned contrasts (one-tailed) via students *t*-tests or Welch tests in case of unequal variances were conducted. For unplanned pair-wise comparisons, post-hoc tests with Tukey's correction method for multiple testing were used. For all statistical tests, an alpha-level of .05 was used. Population effect sizes were estimated for analyses of variance (ANOVAs; omega squared [$\hat{\omega}_p^2$]), as well as for planned contrasts (Cohen's *d* with confidence intervals) (Fig. 1).

The anonymous data set and the analysis scripts of the present study are publicly available at Open Science Framework (https://osf.io/a5ufh/).

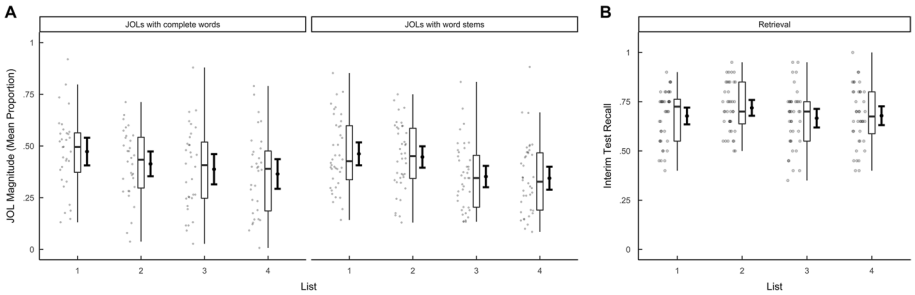

Fig. 2. Interim JOL magnitudes (i.e., mean proportion) as a function of list (1, 2, 3, vs. 4) for the JOL groups with complete words and word stems, respectively (Panel A). Interim test recall (i.e., mean proportion correct) as a function of list (1, 2, 3, vs. 4) for the retrieval group (Panel B). For each list, individual data points are presented as dots to the left; the box plot is presented in the middle, reflecting the distribution of the data points (line in the box = median; height of box = 25th to 75th percentile; whiskers = 1.5 × interquartile range below/above 1st /3rd quartile); mean proportion (correct) is presented to the right with 95% confidence intervals as errors bars.

Results

First, we report interim JOL magnitudes of Lists 1–4 for both JOL groups and interim test recall of Lists 1–4 for the retrieval group. To test our specific hypotheses, we present the main results onthe interim test recall performance (i.e., mean proportion correct) as well as the numbers of (external and internal) intrusions on the criterial List 5 as a function of practice type. Finally, we report recall performance (i.e., mean proportion correct) of the final cumulative test as a function of list and practice type.

Interim JOL Magnitudes and Interim Test Recall of Lists 1–4

Figure 2 presents the interim JOL magnitudes of Lists 1−4 for the JOL groups with complete words and word stems as well as interim test recall (i.e., word-stem based cued-recall) of Lists 1–4 for the retrieval group.

For the two JOL groups, we ran a two-factorial ANOVA on Lists 1–4 interim JOL magnitudes. The results revealed that there was a significant main effect of list, $F(2.47, 172.75) = 20.83$, $p < .001$, $\hat{\omega}_p^2 = .06$, but neither a main effect of practice type (JOLs with word stems vs. JOLs with complete words), $F(1, 70) = 0.05$, $p = .821$, $\hat{\omega}_p^2 < .001$, nor a significant interaction effect between both factors, $F(2.47, 172.75) = 1.64$, $p = .191$, $\hat{\omega}_p^2 < .001$. Post-hoc (Tukey) tests showed that JOL magnitudes differed in both JOL groups between List 1 and List 3, $t(70) = 5.14$, $p < .001$, $d = 0.61$, 95% CI [0.36, 0.87], between List 1 and List 4, $t(70) = 6.17$, $p < .001$, $d = 0.74$, 95% CI [0.47, 1], between List 2 and List 3, $t(70) = 4.41$, $p < .001$, $d = 0.53$, 95% CI [0.28, 0.78], and between List 2 and List 4, $t(70) = 4.47$, $p < .001$, $d = 0.54$, 95% CI [0.28, 0.78], reflecting the learners' experience or belief that interim test recall decreases with continued learning of new information. In addition, for the retrieval group, we ran a one-factorial ANOVA on Lists 1−4 interim test recall. The results showed that there were no significant differences in interim test recall as a function of list, as indicated by a nonsignificant main effect, $F(3, 117) = 1.73$, $p = .165$, $\hat{\omega}_p^2 < .001$.

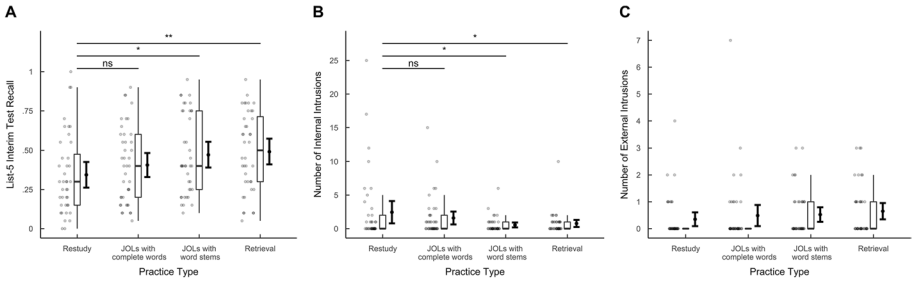

Fig. 3. Interim test recall of the criterial List 5 (i.e., mean proportion correct) as a function of practice type (Panel A). Number of internal intrusions (Panel B) and number of external intrusions (Panel C) during interim test recall of the criterial List 5. For each practice-type group, individual data points are presented as dots to the left; the box plot is presented in the middle, reflecting the distribution of each group's data points; mean proportion correct or average number of intrusions, respectively, are presented to the right with 95% confidence intervals as errors bars. Lines with asterisks indicate significance levels of planned contrast analyses between practice-type groups for interim test recall and number of internal intrusions, respectively.

Interim Test Recall of Criterial List 5

Figure 3 presents the recall measure (proportion correct, number of internal intrusions, and number of external intrusions) for the interim test (i.e., free recall) of the criterial List 5 as a function of practice type.

Proportion correct Figure 3 (Panel A) indicates that interim recall performance (i.e., proportion of correctly recalled words) of the criterial List 5 benefitted from retrieval practice and making JOLs with word stems, compared to restudy. This result pattern was supported by a between-subjects ANOVA, showing a significant main effect of practice type, $F(3, 157) = 2.83$, $p = .04$, $\hat{\omega}_p^2 = 0.03$.

First, we tested the prediction that making JOLs with word stems, but not making JOLs with complete words, in Lists 1–4 enhance the interim recall performance of List 5. Planned contrast analyses revealed that making JOLs with word stems was superior to restudy, $t(78) = 2.24$, $p = .014$, $d = 0.51$, 95% CI [0.13, ∞] (one-tailed), but that making JOLs with complete words was not, $t(79) = 1.12$, $p = .264$, $d = 0.25$, 95% CI [-0.19, 0.70]. In addition, the data showed that practicing retrieval, $t(78) = 2.59$, $p = .006$, $d = 0.59$, 95% CI [0.20, ∞] (one-tailed), trumped restudy.

Second, we tested whether making JOLs with word stems versus retrieval practice differed in the interim recall performance of the criterial List 5. A planned contrast revealed that both practice types did not significantly differ, $t(78) = 0.353$, $p = .725$, $d = 0.08$, 95% CI [-0.36, 0.52].

Number of intrusions Figure 3 presents the number of internal (i.e., prior-list) intrusions (Panel B) and external (i.e., extra-experimental) intrusions (Panel C) as a function of practice type. We report statistical analyses for internal intrusions, though not for external intrusions because the latter occurred rarely for all groups ($M = 0.503$, $SD = 0.975$). Overall, a between-subjects ANOVA revealed a significant main effect of practice type on the number of internal intrusions, $F(3, 157) = 2.97$, $p = .034$, $\hat{\omega}_p^2 = 0.04$. To test the prespecified predictions, we conducted planned contrast analyses.

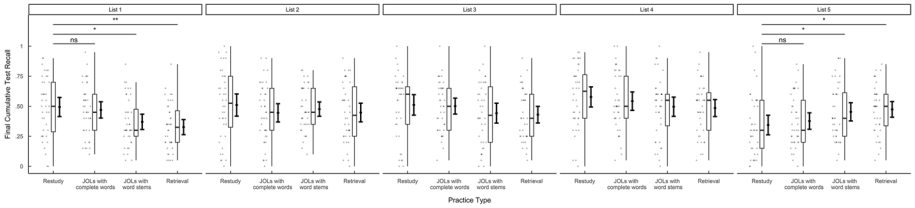

Fig. 4. Final cumulative test recall (i.e., mean proportion correct) as a function of practice type and list. For each practice-type group, individual data points are presented as dots to the left; the box plot is presented in the middle, reflecting the distribution of each group's data points; mean proportion (correct) is presented to the right with 95% confidence intervals as errors bars. Lines with asterisks indicate significance levels of planned contrast analyses between practice-type groups for List 1 and List 5.

First, we tested the prediction that making JOLs with word stems, but not making JOLs with complete words, reduce the number of internal intrusions compared to restudy. Supporting this prediction, the number of internal intrusions was decreased only for making JOLs with word stems, compared to restudy, $t(42.83) = 2.26$, $p = .015$, $d = 0.69$, 95% CI [0.17, ∞] (one-tailed), but again not for making JOLs with complete words, $t(62.38) = 0.91$, $p = .365$, $d = 0.23$, 95% CI [-0.27, 0.73]. The data demonstrated also that retrieval practice reduced the number of internal intrusions compared to restudy, $t(46.84) = 1.97$, $p = .027$, $d = 0.57$, 95% CI [0.07, ∞] (one-tailed).

Second, we tested whether making JOLs with word stems and overt retrieval practice differ in internal intrusion rates. A planned contrast analysis indicated that the number of internal intrusions did not significantly differ between both groups of practice type, $t(69.63) = 0.705$, $p = .483$, $d = 0.17$, 95% CI [-0.30, 0.64].[1]

Final Cumulative *Test* Recall

Final Cumulative Test Recall of Lists 1–5 Figure 4 shows the free-recall performance (i.e., mean proportion correct) as a function of practice type and list on the final cumulative test. An overall 4 (practice type) × 5 (list) ANOVA on mean proportion correct on the final cumulative recall test across Lists 1–5 was conducted. There was a significant main effect of list, $F(4, 628) = 13.84$, $p < .001$, $\hat{\omega}_p^2 = 0.03$, but no main effect of practice type, $F(3, 157) = 0.65$, $p = .581$, $\hat{\omega}_p^2 < 0.001$. Critically, we observed a significant Practice Type × List interaction effect, $F(12, 628) = 4.61$, $p < .001$, $\hat{\omega}_p^2 = 0.02$, indicating the effect of practice type was modulated in magnitude as a function of list. We conducted follow-up analyses for each list. Previous research (Pastötter & Bäuml, 2019) suggested that the last list is specifically susceptible to the FTE, as prior lists can proactively interfere; in addition, the final recall of List 1 may largely reflect BTE, as no lists proactively can interfere.

[1] An alternate, orthogonal set of contrast analyses revealed equivalent results for both interim test recall of criterial List 5 and the internal intrusions. For both dependent variables, we obtained significant differences (all ps $< .01$; ds > 0.4) between the practice-type groups, in which only the word stems were present (i.e., making JOLs with word stems and retrieval) versus the intact words were present (i.e., restudy and making JOLs with complete words). However, within the two combined groups, there wase neither a significant difference between the practice-type groups of JOLs with word stems and retrieval nor between the groups of restudy and JOLs with complete words, respectively (all ps $> .26$).

Given this, we analyzed final cumulative recall performance of List 1 and List 5 separately, because we assumed them to reflect different testing effects. Consistent with previous research (Roediger & Karpicke, 2006; Rowland, 2014), we expected after the short retention interval a restudy advantage for List 1 (as typical for the BTE), but a List-5 recall advantage on the final cumulative test in favor of retrieval practice (cf., Yang et al., 2018), which we tested with contrast analyses. As Lists 2–4 are likely susceptible to the combined influences of FTE and BTE, we did not have any specific predictions and tested potential differences with post-hoc tests.

Final Cumulative Test Recall of List 5 Similar to the results for the List-5 interim test recall, planned contrast analyses demonstrated that making JOLs with word stems, $t(78) = 2.0$, $p = .025$, $d = 0.45$, 95% CI [0.07, ∞] (one-tailed), and retrieval practice, $t(78) = 2.52$, $p = .007$, $d = 0.57$, 95% CI [0.19, ∞] (one-tailed), trumped restudy in final cumulative test recall performance of List 5. However, making JOLs with complete words did not significantly differ from restudy in final cumulative recall performance, $t(79) = 0.63$, $p = .531$, $d = 0.14$, 95% CI [-0.30, 0.58]). Post-hoc (Tukey) tests showed that there were no other significant differences between practice-type groups (all ps > .14).

Final Cumulative Test Recall of List 2–4 As we did not specify any prediction, we ran post-hoc (Tukey) tests, revealing no reliable differences between all four practice types (all ps > .33) on the final cumulative test for Lists 2–4.

Final Cumulative Test Recall of List 1 Planned contrast analyses showed significant differences between restudy and making JOLs with word stems, $t(78) = 2.46$, $p = .016$, $d = 0.56$, 95% CI [0.10, 1.01], and retrieval practice, $t(78) = 3.34$, $p = .001$, $d = 0.76$, 95% CI [0.30, 1.21], exhibiting a restudy advantage on the final cumulative test for List 1. Again, there was no difference between making JOLs with complete words and restudy, $t(79) = 0.47$, $p = .640$, $d = 0.11$, 95% CI [-0.34, 0.55]). Post-hoc (Tukey) tests showed a significant recall advantage of making JOLs with complete words over retrieval practice, $t(157) = 2.99$, $p = .009$, $d = 0.48$, 95% CI [0.16, 0.79]). There were no other significant differences between types of practice (all ps > .09).

General Discussion

The aim of the present study was to examine the effects of making metacognitive judgments, specifically JOLs, and retrieval practice, compared to restudy, on subsequent learning of new information. First, we investigated whether making JOLs has an FTE-type effect, and if so, whether this forward-oriented learning benefit differs for JOLs with word stems and JOLs with complete words, with the potential that participants attempt covert retrieval with the word stems prior to making JOLs. Second, if any JOL-related forward effects emerged, we compared their magnitude to the FTE produced with overt retrieval practice.

Does Making JOLs Potentiate New Learning?

The present results revealed that, relative to restudying Lists 1–4, making JOLs with word stems after each of Lists 1–4 enhanced recall performance of List 5. In contrast, making JOLs with complete words did not produce any significant learning benefit over restudy for List-5 recall performance. These findings extend the results by a previous study on the forward effect of JOLs in inductive learning (Lee & Ha, 2019): item-based JOLs can potentiate new learning when partial study information is available (i.e., covert retrieval attempts are likely) but not when complete study information is available (i.e., covert retrieval attempts are unlikely).

Rather than attempting covert retrieval when making a JOL with all of the study information is available, it is possible that participants engage in elaboration (Craik & Tulving, 1975). Although elaboration generally enhances learning when studying or restudying previously information, elaboration of single words (e.g., semantic generation of a target word based on a word-stem cue) has not been found to have a forward-oriented effect the way that retrieval seems to (Lehman et al., 2014). That is, elaboration has not been shown to be beneficial for learning new, subsequently presented single words. Nevertheless, there may be conditions under which elaborating on the complete study information produces a forward-oriented effect, and these are the conditions under which we would predict making a JOL with the complete study information available would benefit new learning. Indeed, prior research on the backwards effects of making JOLs has found that making JOLs with complete study information available only enhances learning of previously-studied related word pairs (Soderstrom et al., 2015) but not unrelated word pairs (Janes et al., 2018; Myers et al., 2020; Soderstrom et al., 2015) or single words (e.g., Myers et al., 2020). Future research should examine if making JOLs with related word pairs, rather than single words as in the present study, can enhance new learning, even when the complete study information (i.e., the cue–target pair) is present at the time the JOL is made.

Extending the results of Lee and Ha (2019) to verbal learning, this study showed that making item-based JOLs with word stems can potentiate new learning, relative to restudy, an effect that we have attributed to covert retrieval. As participants have not the full information available during JOLs with word stems, more frequent and effortful retrieval attempts and also retrieval failures are likely evoked prior to making the JOLs. Consistent with our predictions, this results pattern hint to the possibility that making cue-only JOLs and word-stem based JOLs can potentiate new learning to the extent that they trigger covert retrieval and that they rely on similar mechanisms as overt retrieval practice. Given this line of reasoning, future research should explore whether a forward effect of item-based JOLs occurs in inductive learning when only the cue (i.e., the painting) but not the corresponding target name of the artist is available.

Given that we used lists of unrelated single words in the present study, the forward effect of interim tests and JOLs with word stems on new learning is likely due to the encoding and reinstatement of contextual features during (covert) retrieval (Kliegl & Bäuml, 2021; Yang et al., in press). This retrieval-based context change facilitates list segregation and thereby protects retrieval of new information against proactive interference from previously studied lists. Support for this hypothesis comes from the finding that both making JOLs with word stems and retrieval practice after each of Lists 1–4 reduced the number of prior-list intrusions on the List-5 recall test compared to restudy (cf. Szpunar et al., 2008). This finding is also consistent with the account that retrieval-based context change "resets" subsequent encoding and provides more capacity of encoding and storage of new information

(Pastötter et al., 2018). In contrast, in support of our hypothesis that making JOLs with complete words does not engender covert retrieval to the same degree as JOLs with incomplete words, the number of prior-list intrusions on the List-5 recall test was similar in the groups involving restudy or JOLs with the complete words.

In addition to the proactive-interference and reset-of-encoding accounts of the FET, the encoding-effort and strategy-change accounts of the FET could also be extended to the forward-oriented effects of making JOLs with only partial study information available. It is conceivable that, just as with retrieval practice, the experience of retrieval failure when making JOLs with word stems could motivate learners to put forth more effort (e.g., Cho et al., 2017; for a metanalytic review, see Chan et al., 2018b) and/or employ more effective strategies to process semantic and/or temporal interitem information when encoding new, subsequently presented material (Chan et al., 2018a; Cho et al., 2017; Cho & Powers, 2019; Yang et al., in press). Although plausible, we argue that encoding effort and strategy shift play a minor role, if at all, in explaining the results of the current study. First, due to the usage of lists of unrelated words, there were only few semantic interitem relations, if at all, that participants could strategically use. In addition, tests with word-stem cues do not require strategic processing of semantic information (because the word can be generated from lexical knowledge) and thereby do likely not facilitate processing of semantic interitem relations in subsequent encoding. Second, participants were also not encouraged to strategically process the temporal or idiosyncratic interitem relations for List 5, because complete words and stem words were presented in a uniquely randomized order during initial study, restudy, and interim phases of Lists 1–4, respectively. Thus, the Lists 1–4 phases of the experiment likely discouraged participants to remember the individual words based on their temporal order. Finally, any potential benefits of developing a particular processing strategy based on expectations of the test format were likely relatively ineffective because the interim tests for Lists 1–4 were stem-based recall but the List-5 test was a free recall test. The List-5 test was the first test that allowed participants to use any recall order but also offered no retrieval support. Thus, it is unlikely that qualitative changes in encoding strategies that participants pursued because of experience with the interim tasks after Lists 1–4 would have significantly improved recall on the List-5 test due to the change in recall format.

Is Making JOLs with Word Stems as Effective as Overt Retrieval Practice?

The present results showed that, relative to making JOLs with word stems, overt retrieval did not further enhance List-5 test recall and further reduce prior-list intrusions. To the degree that making JOLs with word stems engenders covert retrieval, the present results suggest that covert versus overt cued-recall practice with word stems has no reliable influence on potentiating new learning of single-word lists. Thus, response format may not differ in terms of the completeness of the retrieval process, as it would be assumed according to the two-stage process theory (Son & Metcalfe, 2005). Instead, JOLs with word stems may immediately trigger complete retrieval when the materials are single words—potentially this already occurs when assessing the cue familiarity (cf. Tauber et al., 2018). These results are in line with the majority of prior research on the BTE, which has generally found that there is no benefit of overt compared to covert retrieval (e.g., Jönsson et al., 2012; Putnam & Roediger, 2013, Exp. 1 & 3; Smith et al., 2013) or, at most, a very small benefit (e.g, Jönsson et al., 2012, Exp. 1 & 2; Smith et al., 2013; Sundqvist et al., 2017; Tauber et al., 2015, Exp. 2 & 3; Tauber et al., 2018). The present study adds to this

literature, hinting that response format, at least with simple lists of words or word materials, is negligible for the forward testing effect as well.

Although List-5 recall performance was the primary measure of interest, performance on the final cumulative test recall suggests some interesting possibilities about differences in the encoding and recall dynamics induced by restudying, making JOLs, and engaging in overt retrieval. Critically, the results of the final cumulative recall test are consistent with our hypothesis that making JOLs with words stems involves covert retrieval.The effect of (covert and overt) retrieval practice relative to restudy was reversed as a function of list, as indicated by a significant interaction effect between practice type and list. On the final cumulative test, participants in the restudy group recalled more List-1 words than participants in both the JOL-group with word stems and the retrieval group. In contrast, participants in both the JOL group with word stems and the retrieval group recalled more List-5 words on the final cumulative recall test than participants in the restudy group. However, there were no differences among these groups in terms of words recalled from Lists 2–4 on the final cumulative test. Recent research reported a similar results pattern for both younger participants (Pastötter et al., 2020, 2022, but see Yang et al., in press) and older participants (Pastötter & Bäuml, 2019). We attribute the finding that recall of List-5 words on the final cumulative test was lowest in the restudy group to a significant FTE resulting from overt retrieval in the retrieval group (see also Pastötter et al., 2020) and covert retrieval associated with making JOLs with word stems. The finding that recall of List-1 words on the final cumulative test was greatest in the restudy group largely mirrors existing research on the BTE; that is, restudy is typically superior to retrieval practice after short delays (i.e., several minutes), but retrieval practice often leads to enhanced final test recall only after longer delays (e.g., several days; Roediger & Karpicke, 2006; Rowland, 2014). Finally, memory for words from Lists 2–4 might not favor one group because recall of these lists on the final cumulative test can be regarded as a mixed index of the FTE and BTE (cf. Pastötter et al., 2020, 2022).

Together, the findings suggest that the act of making item-based JOLs, similar to overt retrieval practice, can indirectly potentiate new learning when only partial study information is available when making the JOLs. We speculate, although we cannot firmly conclude, that making JOLs with word stems has likely evoked covert retrieval processes. This findings support prior research showing that making cue-only JOLs can also enhance long-term retention of previously learned information (Carpenter et al., 2006; Jönsson et al., 2012; Putnam & Roediger, 2013, Exp. 2; Putnam & Roediger, 2013, Exp. 2 & 3; Smith et al., 2013, Exp. 3 & 4; but see Jönsson et al., 2014; Putnam & Roediger, 2013, Exp. 1; Tekin & Roediger, 2021). Future studies may also involve brain imaging studies to examine more specifically the underlying retrieval processes and potential differences between JOL- and test-based learning (e.g., Jonsson et al., 2020; Vestergren & Nyberg, 2013).

Future Work

Given that a large deal of prior studies on learning and memory has used paired associates, future research should extend the present study on the JOL-based benefits on new learning with single words to paired associates. As suggested by the finding that the delayed JOL effect is smaller for single words than for paired associates (Rhodes & Tauber, 2011), the processes involved in making delayed JOLs for single words versus paired associates may not be identical and potential benefits might be even larger for paired associates. Furthermore, it is a fruitful research avenue to generalize JOL-based benefits in future research

with more complex materials, such as learning text passages or key-term definitions (e.g., Tauber et al., 2018). Beyond the practical relevance, it is also theoretically informative when materials contain multiple idea units: it requires a more effortful and extensive attempt to fully retrieve all the sought-after information. Such exhaustive retrieval of the information is more likely to be triggered when instructing participants to overtly produce the recalled responses. In contrast, if the response is evaluated as familiar, people likely truncate the memory search when making JOLs on materials including more target information that is to be recalled (Son & Metcalfe, 2005). Thus, unlike the results of the present study, when learning more complex materials, the response format of overt versus covert retrieval may significantly affect new learning. Indeed, recent evidence on the BTE showed that overt retrieval of key-term definitions enhanced learning relative to both covert retrieval and restudy in long-term retention, even when encouraging and teaching students how to engage in a more exhaustive covert retrieval attempt (Tauber et al., 2018). In fact, with key-term definitions, covert retrieval did not enhance long-term more than restudying. Thus, it is a promising research avenue for the future to explore the conditions under which response format matters in more complex materials, and whether overt retrieval also enhances subsequent learning of new information more than making JOLs based only on a cue, which presumably evokes covert retrieval. Recent research showed that more effective response formats such as producing an overt response (e.g., enacted retrieval, Kubik et al., 2020) can trump covert retrieval in long-term retention when using a BTE paradigm. Future work is encouraged to further examine the moderating role of response format on potentiating learning of new materials and different interim tests which would facilitate a better theoretical understanding on the workings of JOL- and retrieval-based learning and their educational implications.

Concluding Comments

It is of theoretical, practical, and methodological interest to study the learning benefits conferred by different types of JOLs. Theoretically, making JOLs can potentiate new learning to the extent that people engage in covert retrieval attempts prior to making the JOLs. This is much more likely the case when JOLs are delayed and based on partial cues such as word stems. JOLs with word stems apparently triggered the same, complete retrieval process as overt retrieval practice, thereby potentially encouraging a reset of encoding and combatting proactive interference to a similar degree as overt retrieval practice. Practically, overt retrieval practice helps to sustain the learning outcome throughout a prolonged study material. Similar learning benefits can also be achieved by making JOLs, and they may have similar or even greater utility, considering the circumstance that students often need to silently self-test themselves. Indeed, in higher education, students typically need to acquire the study materials in a library or in a quiet room together with other learners, which makes overt retrieval less applicable, at least in terms of orally rehearsing learning materials. Instead, learners and instructors can seek opportunities to interpolate metacognitive judgments or covert retrieval attempts during longer study periods to maintain their learning outcome over time. Methodologically, researchers need to be more considerate of the reactivity effects of JOLs, and metacognitive ratings in general, when designing experiments. As shown in the present experiment, making JOLs is not a neutral act of assessing one's learning; in fact, JOLs do not only affect previous learning but may also potentiate subsequent learning of new information. Thus, researchers typically need to include

a judgment-free condition to ensure an unbiased assessment of memory performance (cf. Soderstrom et al., 2015).

Author's Contributions Veit Kubik was responsible for the study conception and design, with contributions of Alp Aslan and Torsten Schubert. Material preparation and data collection were performed mainly by Veit Kubik. Analysis scripts and open science materials were made by Kenneth Koslowski and Veit Kubik. The first draft of the manuscript was written by Veit Kubik, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL. T.S. was supported by a grant of the German Research Council Schu 1397 / 7-2 and by a grant of the German Federal Ministry of Research and Education, No. 01PL17065.

Declarations

Conflict of interest The authors declare that they do not have any conflict of interest.

Ethics Approval The study was carried out in accordance with the recommendations of the American Psychological Association's Ethical Principles of Psychologists and Code of Conduct. All participants gave written informed consent in accordance with the Declaration of Helsinki (2013) before participating in the study, with the understanding that they could quit at any time. The study was carried out in accordance with the recommendations of the Ethics Committee of the Faculty of Medicine at Martin Luther University of Halle-Wittenberg, Germany.

References

Arnold, K. M., & McDermott, K. B. (2013). Test-potentiated learning: Distinguishing between direct and indirect effects of tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*, 940–945. https://doi.org/10.1037/a0029199

Aslan, A., & Bäuml, K.-H. T. (2016). Testing enhances subsequent learning in older but not in younger elementary school children. *Developmental Science, 19*(6), 992–998. https://doi.org/10.1111/desc.12340

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research, 87*, 659–701. https://doi.org/10.3102/0034654316689306

Bäuml, K.-H. T., & Kliegl, O. (2013). The critical role of retrieval processes in release from proactive interference. *Journal of Memory and Language, 68*, 39–53. https://doi.org/10.1016/j.jml.2012.07.006

Brunner, E., Konietschke, F., Pauly, M., & Puri, M. L. (2017). Rank-based procedures in factorial designs: Hypotheses about nonparametric treatment effects. *Journal of the Royal Statistical Society: Series B, 79*(5), 1463–1485. https://doi.org/10.1111/rssb.12222

Carpenter, S. K., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued recall test? *Psychonomic Bulletin & Review, 13*, 826–830. https://doi.org/10.3758/BF03194004

Carroll, M., & Shanahan, C. (1997). The effect of context and metamemory judgments on automatic processes in memory. *Acta Psychologica, 97*(3), 219–234. https://doi.org/10.1016/s0001-6918(97)00032-2

Carpenter, S. K., & Pashler, H. (2007). Testing beyond words: Using tests to enhance visuospatial map learning. *Psychonomic Bulletin & Review, 14*(3), 474–478. https://doi.org/10.3758/BF03194092

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

Chan, J., Meissner, C., & Davis, S. (2018b). Retrieval potentiates new learning: A theoretical and metaanalytic review. *Psychological Bulletin. 144*(11), 1111–1146. https://doi.org/10.1037/bul0000166

Cho, K. W., Neely, J. H., Crocco, S., & Vitrano, D. (2017). Testing enhances both encoding and retrieval for both tested and untested items. *Quarterly Journal of Experimental Psychology, 70*, 1211–1235. https://doi.org/10.1080/17470218.2016.1175485

Cho, K. W., & Powers, A. (2019). Testing enhances both memorization and conceptual learning of categorical materials. *Journal of Applied Research in Memory & Cognition, 8*(2), 166–177. https://doi.org/10.1016/j.jarmac.2019.01.003

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General, 104*(3), 268–294. https://doi.org/10.1037/0096-3445.104.3.268

Dang, X., Yang, C., Che, M., Chen, Y., & Yu, X. (in press). Developmental trajectory of the forward testing effect: The role of reset-of-encoding. *European Journal of Developmental Psychology, 59*, 101079. https://doi.org/10.1080/17405629.2021.1986386

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*(4), 558–565. https://doi.org/10.3758/s13421-018-0791-y

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*, 373–380.

Duyck, W., Desmet, T., Verbeke, L. P. C., et al. (2004). WordGen: A tool for word selection and nonword generation in Dutch, English, German, and French. *Behavior Research Methods, Instruments, & Computers, 36*, 488–499. https://doi.org/10.3758/BF03195595

Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *American Psychologist, 63*(7), 591–601. https://doi.org/10.1037/0003-066X.63.7.591

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175–191. https://doi.org/10.3758/BF03193146

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

Jonsson, B., Wiklund-Hörnqvist, C., Stenlund, T., Andersson, M., & Nyberg, L. (2020). A learning method for all: The testing effect is independent of cognitive ability. *Journal of Educational Psychology. Advance online publication.* https://doi.org/10.1037/edu0000627

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 (Formerly Zeitschrift Für Experimentelle Psychologie), 59*(5), 251–257. https://doi.org/10.1027/1618-3169/a000150

Jönsson, F. U., Kubik, V., Larsson Sundqvist, M., Todorov, I., & Jonsson, B. (2014). How crucial is the response format for the testing effect? *Psychological Research, 78*, 623–633. https://doi.org/10.1007/s00426-013-0522-8

Karpicke, J. D., & Aue, W. R. (2015). The testing effect is alive and well with complex materials. *Educational Psychology Review, 27*(2), 317–326. https://doi.org/10.1007/s10648-015-9309-3

Kliegl, O., & Bäuml, K. H. T. (2021). When retrieval practice promotes new learning–The critical role of study material. *Journal of Memory and Language, 120*, 104253. https://doi.org/10.1016/j.jml.2021.104253

Kubik, V., Jönsson, F. U., de Jonge, M., & Arshamian, A. (2020). Putting testing into action. Enacted retrieval practice benefits long-term retention more than covert retrieval retention. *Quarterly Journal of Experimental Psychology, 73*(12), 2093–2105. https://doi.org/10.1177/1747021820945560

Kubik, V., Jönsson, F. U., Knopf, M., & Mack, W. (2018). The direct testing effect is pervasive in action memory. Analyses of recall accuracy and recall speed. *Frontiers in Psychology, 9*, 1632. https://doi.org/10.3389/fpsyg.2018.01632

Kubik, V., Olofsson, J. K., Nilsson, L.-G., & Jönsson, F. U. (2016). Putting action memory to the test: Testing affects subsequent restudy but not long-term forgetting of action events. *Journal of Cognitive Psychology, 28*(2), 209–219. https://doi.org/10.1080/20445911.2015.1111378

Lee, H. S., & Ha, H. (2019). Metacognitive judgments of prior material facilitate the learning of new material: The forward effect of metacognitive judgments in inductive learning. *Journal of Educational Psychology, 111*(7), 1189–1201. https://doi.org/10.1037/edu0000339

Lehman, M., Smith, M. A., & Karpicke, J. D. (2014). Toward an episodic context account of retrieval-based learning: Dissociating retrieval practice and elaboration. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*, 1787–1794. https://doi.org/10.1037/xlm0000012

Li, B., Zhao, W., Zheng, J., Hu, X., Su, N., Fan, T., Yin, Y., Liu, M., Yang, C., & Luo, L. (in press). Soliciting judgments of forgetting reactively enhances memory as well as making judgments of learning: Empirical and meta-analytic tests. *Memory and Cognition.* https://doi.org/10.3758/s13421-021-01258-y

McDaniel, M. A., Roediger, H. L., & McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review, 14*(2), 200–206. https://doi.org/10.3758/BF03194052

McDermott, K. (2021). Practicing retrieval facilitates learning. *Annual Review of Psychology, 72*, 609–633. https://doi.org/10.1146/annurev-psych-010419-051019

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. https://doi.org/10.1037/a0039923

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

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*, 267–270. https://doi.org/10.1111/j.1467-9280.1991.tb00147.x

Pastötter, B., & Bäuml, K.-H. T. (2014). Retrieval practice enhances new learning: The forward effect of testing. *Frontiers in Psychology, 5*. https://doi.org/10.3389/fpsyg.2014.00286

Pastötter, B., & Bäuml, K.-H. T. (2019). Testing enhances subsequent learning in older adults. *Psychology and Aging, 34*, 242–250. https://doi.org/10.17605/osf.io/yfqw9

Pastötter, B., Engel, M., & Frings, C. (2018). The forward effect of testing: Behavioral evidence for the reset-of-encoding hypothesis using serial position analysis. *Frontiers in Psychology, 9*, 1197. https://doi.org/10.3389/fpsyg.2018.01197

Pastötter, B., Schicker, S., Niedernhuber, J., & Bäuml, K.-H. T. (2011). Retrieval during learning facilitates subsequent memory encoding. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 37*, 287–297. https://doi.org/10.1037/a0021801

Pastötter, B., Urban, J., Lötzer, J., & Frings, C. (2022). Retrieval practice enhances new learning but does not affect performance in subsequent arithmetic tasks. *Journal of Cognition, 5*(1), 22. https://doi.org/10.5334/joc.216

Pastötter, B., von Dawans, B., Domes, G., & Frings, C. (2020). The forward testing effect is immune to acute psychosocial encoding/retrieval stress. *Experimental Psychology, 67*, 112–122. https://doi.org/10.1027/1618-3169/a000472

Pastötter, B., Weber, J., & Bäuml, K.-H. T. (2013). Using testing to improve learning after severe traumatic brain injury. *Neuropsychology, 27*(2), 280–285. https://doi.org/10.1037/a0031797

Postman, L., & Keppel, G. (1977). Conditions of cumulative proactive inhibition. *Journal of Experimental Psychology: General, 106*, 376–403. https://doi.org/10.1037/0096-3445.106.4.376

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. https://doi.org/10.3758/s13421-012-0245-x

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. https://doi.org/10.1037/a0021705

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests Improves long-term retention. *Psychological Science, 17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140*(6), 1432–1463. https://doi.org/10.1037/a0037559

Scarrabelotti, M., & Carroll, M. (1998). Awareness of remembering achieved through automatic and conscious processes in multiple sclerosis. *Brain and Cognition, 38*(2), 183–201. https://doi.org/10.1006/brcg.1998.1028

Schneider, W., Eschman, A., & Zuccolotto, A. (2012). *E-Prime user's guide*. Psychology Software Tools, Inc.

Smith, M. A., Roediger, 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

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*(2), 553–558. https://doi.org/10.1037/a0038388

Spellman, B. A., & Bjork, R. A. (1992). When predictions create reality: Judgments of learning may alter what they are intended to assess. *Psychological Science, 3*(5), 315–317. https://doi.org/10.1111/j.1467-9280.1992.tb00680.x

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science, 11*(5), 702–712. https://doi.org/10.1177/1745691616658637

Sundqvist, M. L., Mäntylä, T., & Jönsson, F. U. (2017). Assessing boundary conditions of the testing effect: On the relative efficacy of covert vs. overt retrieval. *Frontiers in Psychology, 8*. https://doi.org/10.3389/fpsyg.2017.01018

Szpunar, K. K., McDermott, K. B., & Roediger, H. L. (2008). Testing during study insulates against the buildup of proactive interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(6), 1392–1399. https://doi.org/10.1037/a0013082

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

Tauber, S. K., Witherby, A. E., Dunlosky, J., Rawson, K. A., Putnam, A. L., & Roediger, H. L. (2018). Does covert retrieval benefit learning of key-term definitions? *Journal of Applied Research in Memory and Cognition, 7*(1), 106–115. https://doi.org/10.1016/j.jarmac.2016.10.004

Tempel, T., & Kubik, V. (2017). Test-potentiated learning of motor sequences. *Memory, 25*(3), 326–334. https://doi.org/10.1080/09658211.2016.1171880

Tullis, J. G., & Benjamin, A. S. (2011). On the effectiveness of self-paced learning. *Journal of Memory and Language, 64*, 109–118. https://doi.org/10.1016/j.jml.2010.11.002

Underwood, J. (1957). Interference and forgetting. *Psychological Review, 64*(1), 12.

Undorf, M., Böhm, S., & Cüpper, L. (2016). Do judgments of learning predict automatic influences of memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*(6), 882–896. https://doi.org/10.1037/xlm0000207

Vestergren, V., & Nyberg, L. (2013). Testing alters brain activity during subsequent restudy: Evidence for test-potentiated encoding. *Trends in Neuroscience and Education, 3*(2), 69–80. https://doi.org/10.1016/j.tine.2013.11.001

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. https://doi.org/10.1016/j.jarmac.2017.08.004

Yang, C., Luo, L., Vadillo, M. A., Yu, R., & Shanks, D. R. (2021). Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review. *Psychological Bulletin, 147*(4), 399–435. https://doi.org/10.1037/bul0000309

Yang, C., Potts, R., & Shanks, D. R. (2018). Enhancing learning and retrieval of new information: A review of the forward testing effect. *Npj Science of Learning, 3*(1). https://doi.org/10.1038/s41539-018-0024-y

Yang, C., Zhao, W., Luo, L., Sun, B., Potts, R., & Shanks, D. R. (in press). *Testing potential mechanisms underlying test-potentiated new learning*. *Learning, Memory, and Cognition*. https://doi.org/10.1037/xlm0001021

Authors and Affiliations

Veit Kubik[1] · Kenneth Koslowski[2] · Torsten Schubert[3] · Alp Aslan[4]

Kenneth Koslowski
k.koslowski@phb.de

Torsten Schubert
torsten.schubert@psych.uni-halle.de

Alp Aslan
alp.aslan@th-rosenheim.de

1 Department of Psychology, Bielefeld University, 33501 Bielefeld, Germany

2 Department of Psychology, Psychologische Hochschule Berlin, 10179 Berlin, Germany

3 Department of Psychology, Martin-Luther-University Halle-Wittenberg, 06108 Halle, Saale, Germany

4 Faculty of Social Sciences, Department of Psychology, Rosenheim Technical University of Applied Sciences, 83024 Rosenheim, Germany