Memory & Cognition (2023) 51:1547–1561
https://doi.org/10.3758/s13421-023-01417-3

Judgments of learning enhance recall for category-cued but not letter-cued items

Michelle L. Rivers[1,2] · John Dunlosky[2] · Jessica L. Janes[2] · Amber E. Witherby[1,3] · Sarah K. Tauber[1]

Accepted: 19 March 2023 / Published online: 12 May 2023

Abstract
Making immediate judgments of learning (JOLs) during study can influence later memory performance, with a common outcome being that JOLs improve cued-recall performance for related word pairs (i.e., positive reactivity) and do not impact memory for unrelated pairs (i.e., no reactivity). The *cue-strengthening hypothesis* proposes that JOL reactivity will be observed when a criterion test is sensitive to the cues used to inform JOLs (Soderstrom et al., *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41* (2), 553-558, 2015). Across four experiments, we evaluated this hypothesis with category pairs (e.g., *A type of gem – Jade*) and letter pairs (e.g., *Ja – Jade*). Participants studied a list comprised of both pair types, made (or did not make) JOLs, and completed a cued-recall test (Experiments 1a/b). The cue-strengthening hypothesis predicts greater positive reactivity for category pairs than for letter pairs, because making a JOL strengthens the relationship between the cue and target, which is more beneficial for material with an a priori semantic relationship. Outcomes were consistent with this hypothesis. We also evaluated and ruled out alternative explanations for this pattern of effects: (a) that they arose due to overall differences in recall performance for the two pair types (Experiment 2); (b) that they would also occur even when the criterion test is *not* sensitive to the cues used to inform JOLs (Experiment 3); and (c) that JOLs only increased memory strength for the targets (Experiment 4). Thus, the current experiments rule out plausible accounts of reactivity effects and provide further, converging evidence for the cue-strengthening hypothesis.

Keywords Judgments of learning · Measurement reactivity · Metamemory · Monitoring

Introduction

On a scale from 0% to 100%, judge how likely you are to remember the word pair *table – chair* on a later test. This judgment, often used in metamemory research to investigate people's ability to monitor their own learning, is referred to as a *judgment of learning* (JOL; for a review, see Rhodes, 2016). Research suggests that under certain conditions, making JOLs influences the representation of the material being monitored, ultimately affecting later retention (e.g., Ariel et al., 2021; Double & Birney, 2019; Double et al., 2017; Halamish & Undorf, 2022; Janes et al., 2018; Li et al., 2022; Maxwell & Huff, 2022; Mitchum et al., 2016; Myers et al., 2020; Rivers et al., 2021; Senkova & Otani, 2021; Shi et al., 2022; Soderstrom et al., 2015; Tauber & Witherby, 2019; Tekin & Roediger, 2020; Witherby & Tauber, 2017; Yang et al., 2015; Zhao et al., 2022; for a review of the effects on learning when JOLs are made after a delay, see Rhodes & Tauber, 2011). Our primary aim in the present research was to answer the question: What processes underlie this reactive effect of making immediate JOLs on retention (i.e., *JOL reactivity*)?

Researchers have investigated JOL reactivity using similar methodology (Janes et al., 2018; Mitchum et al., 2016; Myers et al., 2020; Rivers et al., 2021; Soderstrom et al., 2015). In particular, participants studied a mixed list of cue-target word pairs, half of which were related (e.g., *railroad – train*) and half of which were unrelated (e.g., *traffic – soap*). Some participants made JOLs during the presentation of each pair (i.e., estimate the likelihood of successfully recalling each pair on a 0–100% scale), and others did

✉ Michelle L. Rivers
mlrivers3@gmail.com

[1] Department of Psychology, Texas Christian University, TCU Box #298920, 2800 S. University Dr, Fort Worth, TX 76129, USA

[2] Department of Psychology, Kent State University, Kent, OH, USA

[3] Department of Psychology, Creighton University, Omaha, NE, USA

not. Performance on a final cued-recall test (e.g., *traffic – ?*) was then compared for participants who made JOLs relative to those who did not. Recall performance for related pairs consistently demonstrated *positive reactivity* – recall was significantly higher for related pairs that were judged compared to those that were not judged. By contrast, recall for unrelated pairs tended to be statistically equivalent for judged versus non-judged pairs or in some cases, unrelated pairs showed impaired recall for judged versus non-judged pairs (*negative reactivity*).

Multiple hypotheses have been proposed to explain these observed recall patterns (and we return to such hypotheses in the *General discussion*), but most relevant to the current research, the *cue-strengthening hypothesis* (Soderstrom et al., 2015) proposes that if the cue used to inform JOLs about particular items is relevant to a future criterion test, positive reactivity will be observed for those items. In the case of related and unrelated pairs, participants use the associative relatedness of the two words in a pair to inform their judgments and typically make substantially higher JOLs for related than unrelated pairs (e.g., Dunlosky & Matvey, 2001; Koriat, 1997; Mueller et al., 2013). According to this hypothesis, making JOLs (vs. not making them) increases people's processing of the cue-target relationships, which presumably improves performance on a cued-recall test for related pairs because they have an a priori relationship that could benefit from further processing and the test is sensitive to this relationship.

Evidence for this hypothesis is based largely on the multiple replications of positive reactivity effects for related pairs but not for unrelated pairs (Halamish & Undorf, 2022; Janes et al., 2018; Mitchum et al., 2016; Myers et al., 2020; Rivers et al., 2021; Soderstrom et al., 2015), but related and unrelated pairs can differ on other dimensions, such as the ease of processing the words within the pair. Thus, in the current research, we further evaluated the cue-strengthening hypothesis by using another manipulation of a priori association, with the goal being to provide a greater breadth of converging evidence that a priori association moderates JOL reactivity effects. In particular, we used a levels-of-processing manipulation (Craik & Lockhart, 1972; cf. Maxwell & Huff, 2022; Tekin & Roediger, 2020) in which participants made JOLs for target items paired with category cues (e.g., *A type of gem – Jade*) or letter cues (e.g., *Ja – Jade*). Like related pairs, category cues share a semantic relationship with the target and hence further processing invoked by making a JOL may strengthen this semantic relationship. In contrast, letter-cued targets, like unrelated pairs, are not expected to benefit from cue-target relational processing because they do not share a semantic relationship.

An important assumption of the cue-strengthening hypothesis is that making a JOL can result in a strengthening of the cues used as a basis for the judgment (i.e., semantic relatedness). That is, participants should make higher judgments for category than letter pairs. With respect to this assumption, prior research is mixed, showing either significantly higher JOLs for items with category cues than letter cues or trends in that direction (Bieman-Copland & Charness, 1994; Matvey et al., 2002; Mueller et al., 2015). Note, however, that these mixed outcomes were obtained after a single study cycle. When people undergo multiple study-test cycles (with new items on each cycle), they gain test experience on the first cycle and update their knowledge that memory is better for category than letter cues. Accordingly, on the second study-test cycle, JOLs are higher for items with category than letter cues (e.g., Mueller et al., 2015). Thus, to ensure that participants would use semantic relatedness to inform their JOLs, we used two study-test cycles for our initial experiments. If the same pattern of reactivity occurs for category and letter-cued items (as for related vs. unrelated pairs), it would provide further confidence in the conclusion that the critical variable driving JOL reactivity is a priori semantic relatedness (and not other, idiosyncratic factors that differ across the various manipulations).

According to the aforementioned rationale, the cue-strengthening hypothesis makes the following predictions: Assuming that JOLs are greater for category-cued items than letter-cued items (which may occur on the first study-test cycle but is expected on the second study-test cycle), then (a) positive reactivity will occur for targets with category cues and (b) the reactive effect will be larger for targets with category cues relative to letter cues (which we expect to be small, if it occurs at all). A key idea here is that strengthening of semantic relationships (in the case of category pairs) will be relevant for boosting performance on a subsequent cued-recall test because the semantic relationship will provide a distinctive cue for reconstructing the target from the cue. In contrast, for letter pairs, making JOLs is not expected to boost performance for a couple reasons. First, letter cues and their target words would not obviously benefit from semantic processing because their relationship is not inherently semantic. Thus, the (lack of a) semantic relationship would not be relevant to making JOLs. Second, any extra processing from making a JOL could not further strengthen a lexical association between the first letter of a well-known word and the word itself (e.g., the fact that "jade" begins with "ja" is already overlearned).

To foreshadow, in Experiment 1, we observed outcomes consistent with the cue-strengthening hypothesis – positive reactivity for category-cued items, and no reactivity for letter-cued items. Thus, in Experiment 2, we evaluated an alternative explanation for reactive effects. In particular, positive reactivity is confounded with the level of mean recall performance (i.e., it is higher for related pairs than unrelated ones in prior research and for category-cued items than for letter-cued items in the current research). Our materials allowed us

to eliminate this confound so as to evaluate whether reactivity is more due to a priori semantic relatedness or to higher levels of recall performance. Finally, we evaluated another prediction of the cue-strengthening hypothesis (that positive reactivity for category-cued items should only occur if the criterion test relies on the same cues that inform JOLs) in Experiments 3 and 4, and explored whether making JOLs increased the memory strength of targets of letter-cued items in Experiment 4.

Experiments 1a and 1b

Our primary goal in Experiments 1a and 1b was to investigate the cue-strengthening hypothesis using a levels-of-processing manipulation. As discussed above, participants were presented with a list of category-cued items (i.e., category pairs) and letter-cued items (i.e., letter pairs; pairs were randomly intermixed) and were randomly assigned to either make JOLs for each pair or to simply study each pair. Following study, participants took a cued-recall test on the pairs. Participants then repeated this procedure for another list of pairs. Given that Experiment 1a comprises the first investigation of JOL reactivity using category and letter pairs, Experiment 1b was conducted to directly replicate these outcomes with a different student population (for discussion of the importance of direct replication, see Simons, 2014).

Method

Participants

We used the software program G*Power (Faul et al., 2007) to conduct a power analysis for an analysis of variance (ANOVA). Our goal was to obtain .80 power at the standard alpha error probability (.05) to detect a medium effect (f = .25; based on prior research with related and unrelated pairs), which yielded a target sample size of 128 participants. For all experiments, timeslots were posted on a weekly basis until the target sample size was reached. In Experiment 1a, 140 undergraduates participated in exchange for partial credit in their Psychology course. Two participants were excluded from analysis due to a computer error during data collection. Participants were randomly assigned to the JOL (n = 68) and no-JOL (n = 70) groups.

Experiment 1b was an independent replication of Experiment 1a. Thus, the method was identical except for the participant sample. Whereas participants in Experiment 1a were undergraduates from Kent State University (KSU), participants in Experiment 1b were 133 undergraduates (n = 66 and 67 randomly assigned to the JOL and no-JOL groups, respectively) from Texas Christian University (TCU). Informed consent was obtained from all individual participants included in the experiments.

Design

We used a 2 (judgment group: JOL vs. no JOL) × 2 (pair type: letter vs. category) × 2 (study-test cycle: first vs. second) mixed design, with pair type and study-test cycle manipulated within participants and judgment group manipulated between participants.

Materials

Materials were 88 paired associates from Bieman-Copland and Charness (1994). The targets were randomly assigned to be presented with either a category cue (e.g., *A type of gem – Jade*) or a letter cue (e.g., *Ja – Jade*) for each participant. Two lists were formed such that half of the pairs were presented during each cycle, with 22 pairs of each type presented during each cycle. Within each list, the letter and category cues were unique for each target (e.g., not more than one type of gem was presented). The presentation order of the two lists was counterbalanced across cycles between participants. Materials and item-level data for all experiments reported here can be accessed at the following link: https://osf.io/84q2p/.

Procedure

Participants were run in small groups of up to six, and each participant was seated at an individual terminal with a computer that displayed instructions and stimuli. The experiment was coded with LiveCode. Participants were instructed to study the pairs for an upcoming cued-recall test (i.e., "On the test you will receive the clue and be asked to recall the word it was presented with") and given an example of each pair type. Pairs were presented individually for 8 s each. Halfway through the presentation of each pair (i.e., after 4 s), participants in the JOL group were prompted to type a JOL into a text box (i.e., indicate the likelihood of remembering the pair on a later test on a scale from 0 to 100) while the pair remained on screen (i.e., participants had the remaining 4 s to make their JOL). Participants in the no-JOL group made no such judgments. After presentation of all pairs, participants engaged in a self-paced test in which they were given the category or letter cue (depending on the pair type) and were asked to recall the target (e.g., *A type of gem – ?*). Across experiments 1–3, participants took approximately 5 s attempting cued recall for each target. The order of presentation of the pairs during both study and test were randomized anew for each participant, constrained to only allow three of the same type of pair to be presented in a row. After completing the first cycle, participants followed the same

Table 1 Mean magnitudes of judgments of learning (JOLs) in Experiments 1a, 1b, 2, 3, and 4

Experiment/manipulation	Category pairs	Letter pairs	Difference
1a, Cycle 1	58.61 (1.94)	47.79 (2.39)	$t(67) = 5.06$, $p < .001$, 95% CI [6.55, 15.10], $g_{av} = 0.60$, $BF_{10} = 4188.12$
1a, Cycle 2	60.77 (2.38)	48.71 (2.38)	$t(67) = 5.03$, $p < .001$, 95% CI [7.28, 16.85], $g_{av} = 0.61$, $BF_{10} = 3827.07$
1b, Cycle 1	57.14 (2.19)	47.74 (2.45)	$t(65) = 4.62$, $p < .001$, 95% CI [5.33, 13.46], $g_{av} = 0.49$, $BF_{10} = 865.87$
1b, Cycle 2	57.81 (2.12)	47.70 (2.51)	$t(65) = 5.35$, $p < .001$, 95% CI [6.33, 13.88], $g_{av} = 0.53$, $BF_{10} = 11574.20$
2, Category-Advantaged	55.00 (2.96)	43.51 (3.55)	$t(33) = 4.69$, $p < .001$, 95% CI [6.51, 16.47], $g_{av} = 0.59$, $BF_{10} = 494.45$
2, No-Advantage	43.62 (2.67)	47.88 (3.49)	$t(35) = -1.54$, $p = .13$, 95% CI [-9.88, 1.36], $g_{av} = -0.22$, $BF_{01} = 2.51$
3, Cued Recall	51.75 (2.36)	48.31 (2.74)	$t(35) = 1.26$, $p = .22$, 95% CI [-2.10, 8.99], $g_{av} = 0.22$, $BF_{01} = 3.61$
3, Letter-Cued Recall	46.20 (2.41)	45.50 (3.05)	$t(36) = 0.26$, $p = .80$, 95% CI [-4.76, 6.16], $g_{av} = 0.04$, $BF_{01} = 7.57$
4, Free Recall	60.25 (1.94)	53.22 (2.40)	$t(64) = 3.10$, $p = .003$, 95% CI [2.50, 11.56], $g_{av} = 0.39$, $BF_{10} = 8.07$

Mean values of judgments of learning for each pair type as a function of various manipulations across experiment (i.e., cycle, recall group, test group) across Experiments 1a, 1b, 2, 3, and 4. Standard error of the mean is in parentheses. Difference represents statistics from a paired-samples *t*-test (two-tailed) comparing mean JOLs for category and letter pairs

procedure for a novel set of pairs in cycle 2.[1] The entire procedure took approximately 25 min.

Results and discussion

For the focal analyses, we report the *p* value, a standardized measure of effect size (Hedges' *g* or η_p^2; formulas from Lakens, 2013), and the Bayes factor (*BF;* Kruschke, 2013). BFs quantify the strength of the evidence in favor of the alternative hypothesis (in the current investigation, JOL reactivity) relative to the null hypothesis (i.e., no JOL reactivity). The BF is a ratio of the likelihood of the data given the alternative hypothesis to the likelihood of the data given the null hypothesis (BF_{10}). That is, a BF of 1 means that the data are equally likely under the alternative and null hypotheses. Bayes factors can indicate that the null hypothesis is more probable than the alternative hypothesis (i.e., when $BF_{10} < 1$), and is reported as the reciprocal BF_{01}. We used the Jeffrey-Zellner-Siow (JZS) prior (with the *r* scale parameter set at 1) because it requires the fewest prior assumptions about the range of the true effect size (Rouder et al., 2009). BFs were calculated using the R Package BayesFactor (Morey & Rouder, 2018).

JOL magnitudes for category and letter pairs for all experiments are presented in Table 1. For both Experiment 1a and Experiment 1b, JOLs were significantly higher for category than letter pairs. Recall performance is presented in Fig. 1. We conducted 2 (judgment group) × 2 (pair type) × 2 (study-test cycle) mixed ANOVAs on cued-recall performance for both Experiment 1a and Experiment 1b. Patterns of recall for both experiments were consistent with the cue-strengthening hypothesis – we found positive reactivity for category pairs, and no reactivity for letter pairs.

Experiment 1a

Collapsed across the other variables, recall was higher for category pairs ($M = .78$, $SE = .01$) than for letter pairs ($M = .39$, $SE = .01$), $F(1, 136) = 792.37$, $p < .001$, $\eta_p^2 = 0.85$; recall was also higher for judged pairs ($M = .61$, $SE = .02$) than for non-judged pairs ($M = .56$, $SE = .02$), $F(1, 136) = 4.68$, $p = .032$, $\eta_p^2 = 0.03$. The judgment group by cycle interaction was marginally significant, $F(1, 136) = 3.87$, $p = .051$, $\eta_p^2 = 0.03$, and the pair type by cycle interaction was significant[2], $F(1, 136) = 4.17$, $p = .043$, $\eta_p^2 = 0.03$.

Critically, the judgment group by pair type interaction was significant, $F(1, 136) = 26.91$, $p < .001$, $\eta_p^2 = 0.17$. Collapsing across the two cycles, independent samples *t*-tests (two-tailed) revealed that recall for category pairs was significantly higher for the JOL group ($M = .84$, $SE = .01$) than for the no-JOL group ($M = .71$, $SE = .02$), $t(136) = 4.56$, $p < .001$, 95% CI [.07, .17], $g_s = 0.77$, $BF_{10} = 1430.63$, whereas recall for letter pairs was not significantly different for the JOL group ($M = .38$, $SE = .02$) versus the no-JOL group ($M = .40$, $SE = .02$), $t(136) = -0.75$, $p = .45$, 95% CI [-.07, .03], $g_s = -0.13$, $BF_{01} = 5.77$. The three-way interaction was not significant, $F(1, 136) = 0.36$, $p = .55$, $\eta_p^2 = 0.003$, suggesting the pattern of JOL reactivity (i.e., positive reactivity for category pairs and no reactivity for letter pairs) did not differ across the two learning cycles.

[1] Before and after Experiment 1b, we also administered a questionnaire to assess participants' effectiveness ratings for different encoding strategies, but due to a coding error, these data were not saved.

[2] This interaction was driven by the fact that recall for letter pairs significantly increased across cycles, whereas recall for category pairs did not (for both Experiment 1a and Experiment 1b), perhaps because some learners demonstrated a "learning to learn" effect across cycles (as in deWinstanley & Bjork, 2004; Storm et al., 2016).

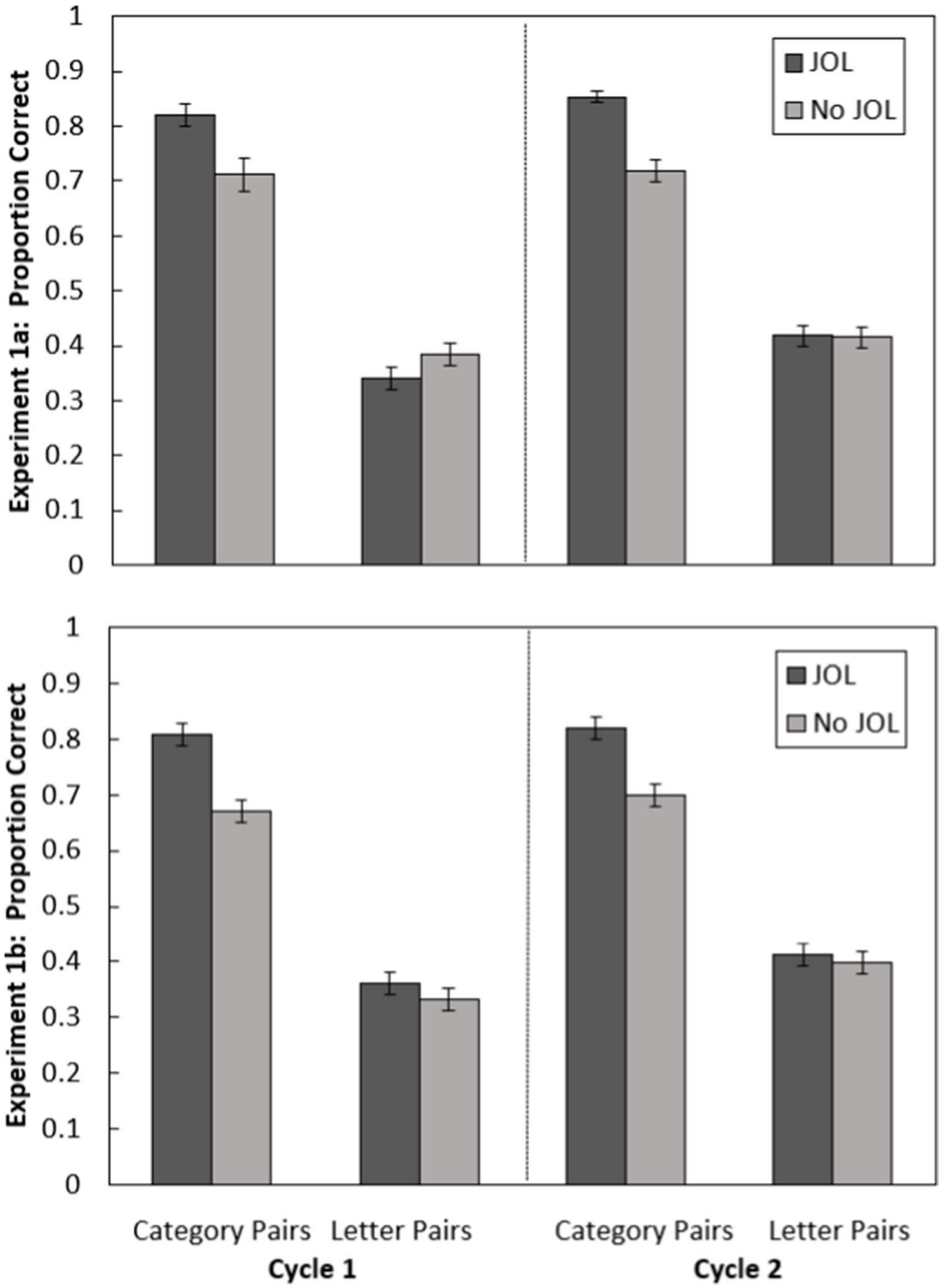

Fig. 1 Recall performance as a function of judgment group, pair type, and study-test cycle in Experiment 1a (**top**) and Experiment 1b (**bottom**). *Note*. JOL = judgment of learning. Error bars reflect the standard error of each mean

Experiment 1b

Patterns of results in Experiment 1b replicated Experiment 1a. Collapsed across the other variables, recall was higher in cycle 2 (M = .58, SE = .01) than in cycle 1 (M = .54, SE = .01), $F(1, 131) = 14.00$, $p < .001$, $\eta_p^2 = 0.10$; recall was also higher for category pairs (M = .75, SE = .01) than for letter pairs (M = .38, SE = .01), $F(1, 131) = 1029.94$, $p < .001$, $\eta_p^2 = 0.89$; finally, recall was higher for judged pairs (M = .60, SE = .02) than for non-judged pairs (M = .53, SE = .02), $F(1, 131) = 11.33$, $p < .001$, $\eta_p^2 = 0.08$. The judgment group by cycle interaction was not significant, $F(1, 131) = 0.70$, $p = .41$, $\eta_p^2 = 0.01$. The pair type by cycle interaction was significant, $F(1, 131) = 6.38$, $p = .013$, $\eta_p^2 = 0.05$.

Critically, the judgment group by pair type interaction was significant, $F(1, 131) = 21.38$, $p < .001$, $\eta_p^2 = 0.14$. Collapsing across the two cycles, independent samples t-tests

(two-tailed) revealed that recall for category pairs was significantly higher for the JOL group ($M = .81$, $SE = .02$) than for the no-JOL group ($M = .68$, $SE = .02$), $t(131) = 4.81$, $p < .001$, 95% CI [.08, .18], $g_s = 0.83$, $BF_{10} = 3750.67$, whereas recall for letter pairs did not significantly differ for the JOL group ($M = .39$, $SE = .02$) and the no-JOL group ($M = .37$, $SE = .01$), $t(131) = 0.92$, $p = .36$, 95% CI [-.02, .07], $g_s = 0.16$, $BF_{01} = 4.97$. The three-way interaction was not significant, $F(1, 131) = 0.001$, $p = .98$, $\eta_p^2 < .001$.

Experiment 2

The recall patterns observed with both related and unrelated word pairs (in prior research) and category and letter pairs (in Experiments 1a and 1b) confirmed a key prediction of the cue-strengthening hypothesis, which explains these differential effects based on the qualitative nature of the cues used to inform judgments. In particular, positive reactivity occurs for category cues because JOLs presumably strengthen the semantic relationship between either (a) two semantically related words in a pair (prior research) and (b) the category cue and corresponding target (current research).

Another (albeit less interesting) explanation is that reactivity occurs in these cases because of a quantitative difference in the level of recall performance. According to this hypothesis, reactivity presumably occurs for category pairs and not for letter pairs because the former are recalled at a higher rate, even when JOLs are not elicited. The same is true for prior manipulations of associative relatedness, with performance being substantially higher for related than unrelated pairs (Janes et al., 2018; Mitchum et al., 2016; Myers et al., 2020; Rivers et al., 2021; Soderstrom et al., 2015). To provide a more complete characterization of these relationships, Fig. 2 depicts JOL reactivity effects as a function of mean recall performance (across JOL and no-JOL groups), with contributing data from (a) prior research with related and unrelated pairs, and (b) Experiments 1a and 1b of the current investigation. As illustrated in this figure, positive reactivity effects are confounded with higher levels of mean recall performance. By virtue of beginning higher on the performance scale, extra processing accrued by making a JOL may be more likely to improve performance.

One reason this effect may occur pertains to the psychophysical function relating memory to objective performance. As depicted in Fig. 3 (solid line), some plausible functions (with a flatter function at the lower end of the memory scale and steeper one at the higher end) would produce such an outcome. Note that the boost that occurs in memory for both kinds of cue is similar, yet given the psychophysical function relating memory to performance, this actual change in memory yields an effect on cued-recall performance for items (in this case, those with category cues) that begin higher on the scale but no difference

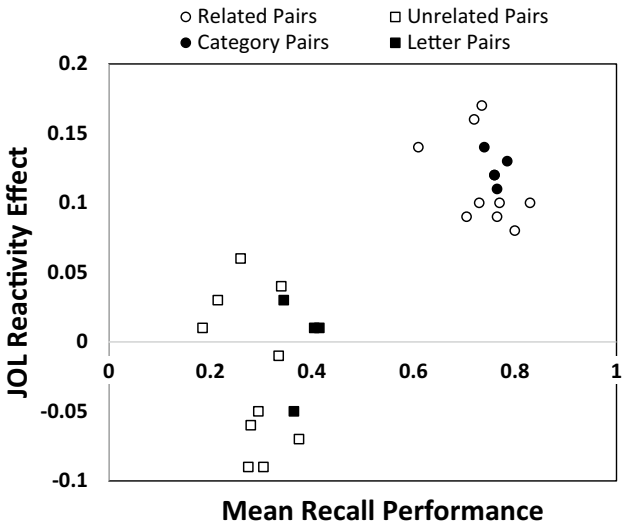

Fig. 2 Judgment of learning (JOL) reactivity as a function of mean recall performance. *Note.* JOL reactivity effect = Proportion of items recalled for JOL group – Proportion of items recalled for no-JOL group. Mean recall performance = Average proportion recalled for both the JOL and no-JOL groups. Contributing articles for related and unrelated pair points are noted by an asterisk (*) in the References section. All contributing studies used a similar methodology: a mixed list of related and unrelated word pairs, a between-participant manipulation of JOL and no-JOLs, experimenter-paced study, and a cued-recall criterion test. Category and letter pair data points come from Experiments 1a and 1b of the current investigation

for items that begin lower on the scale (for a detailed discussion, see Loftus, 1978; Wagenmakers et al., 2012). Of course, such psychophysical functions for memory are unknown (i.e., the strength of a given item in memory cannot be directly measured; Soderstrom & Bjork, 2015), and other plausible functions would yield different outcomes, such as a steeper function near the lower end of the memory scale (Fig. 3, dashed line). If the former is the case (Fig. 3, solid line), however, the reactivity observed in the present experiments may be attributable to differences in level of recall performance rather than to the qualitative differences in the cues used to make JOLs.

In Experiment 2, we evaluated the possibility that different levels of memory strength contribute to the pattern of reactivity effects presented in Fig. 1 (i.e., the *memory strength account*). In particular, based on outcomes from Experiments 1a and 1b, we experimentally manipulated the level of recall performance for category and letter pairs such that a random half of participants demonstrated a recall advantage for category pairs (category-advantaged group) as in Experiments 1a and 1b, and the other half demonstrated approximately equivalent recall for category and letter pairs when no JOLs were made (no-advantage group; cf. Storm et al., 2016, Experiment 2).

The cue-strengthening hypothesis predicts that only a specific type of processing (i.e., for a cued-recall test, one

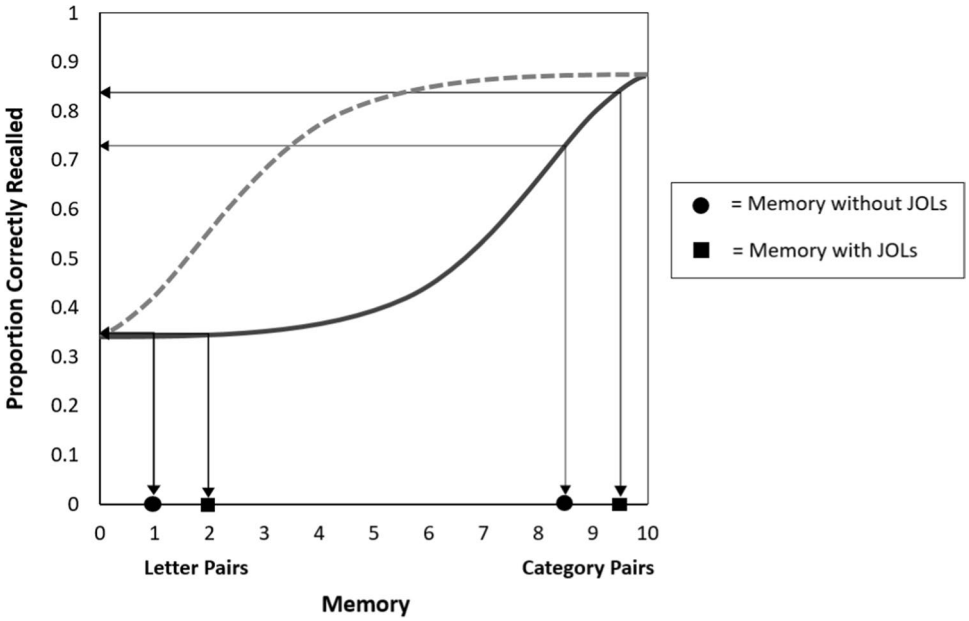

Fig. 3 Unobserved psychophysical functions relating memory (with or without judgments of learning (JOLs)) to cued-recall performance (proportion correctly recalled) for letter and category pairs. *Note.* Scale for memory is arbitrary except that higher values indicate higher levels of memory. Solid line represents a function that would only lead to JOL reactivity for items that begins higher on the memory scale but no JOL reactivity for items that begin lower on the scale. Dashed line represents a function that would only lead to JOL reactivity for items that begins lower on the memory scale but no JOL reactivity for items that begin higher on the scale

that leads to strengthening of a cue-target relationship) will lead to positive JOL reactivity, regardless of the level of underlying memory strength or recall performance. Thus, this hypothesis predicts greater positive reactivity for category than letter pairs for both the category-advantaged and no-advantage groups. According to the memory strength account, however, when recall is approximately equivalent (and high) for category- and letter-cued items, the observed level of JOL reactivity will be similar for category and letter pairs. That is, learning material that is above a certain memory strength will benefit from the requirement to make JOLs regardless of the type of processing that JOLs invoke.

Finally, we expected JOLs to be higher for category pairs than for letter pairs in the category-advantaged group, as in Experiments 1a and 1b. However, even if this cue is consistently a basis for JOLs, it may not be observed for the no-advantage group. In particular, prior research has demonstrated that item difficulty can influence JOLs (Begg et al., 1989), and hence the impact of pair type (with JOLs being higher for category than letter pairs) may trade off with item difficulty (with items being normatively easier for letter pairs). Accordingly, we evaluated whether both groups held the same belief that category cues are more effective than letter cues, so we also administered a questionnaire about encoding preferences before and after the experiment.

Method

We used a 2 (judgment group: JOL vs. no JOL) × 2 (pair type: letter vs. category) × 2 (recall group: category-advantaged vs. no-advantage) mixed design, with pair type manipulated within participants and judgment group and recall group manipulated between participants. Our target sample size was 128 participants. Based on sensitivity analyses using G*Power with alpha of .05 and .80 power, this sample size afforded sufficient power to observe the positive reactivity effect ($g \geq 0.58$) observed for the JOL versus no-JOL recall for the category pairs in Experiments 1a and 1b. A total of 138 KSU undergraduates participated in exchange for partial credit in their Psychology course. Participants were randomly assigned to the four groups: JOL, category-advantaged recall ($n = 34$); no JOL, category-advantaged recall ($n = 33$); JOL, no-advantage recall ($n = 36$); no JOL, no-advantage recall ($n = 35$).

The materials were nearly identical to those used in our prior experiments. However, the list was constructed to make either category or letter pairs "advantaged" and thus easier to recall. Using data from Experiments 1a and 1b, we identified the 15 target words that were easiest to recall (average recall: 73%), and the 15 target words that were most difficult to recall (average recall: 44%), regardless of encoding condition (i.e., whether they were paired with category or letter

Table 2 Mean ratings of encoding strategy effectiveness in Experiments 2, 3, and 4

		Category-advantaged group			No-advantage group	
	Category	Letter	Rhyme	Category	Letter	Rhyme
Expt 2, Before	7.81 (.25)	5.06 (.28)	7.07 (.27)	7.69 (.27)	4.62 (.30)	6.83 (.25)
Expt 2, After	9.00 (.16)	2.69 (.21)	5.46 (.25)	7.76 (.25)	5.28 (.34)	5.66 (.25)
		Cued-recall group			Letter-cued recall group	
	Category	Letter	Rhyme	Category	Letter	Rhyme
Expt 3, Before	7.56 (.27)	5.14 (.33)	6.74 (.29)	8.19 (.20)	4.66 (.30)	6.64 (.26)
Expt 3, After	8.78 (.24)	2.84 (.28)	5.53 (.27)	7.68 (.26)	2.97 (.28)	6.09 (.25)
		Free recall				
	Category	Letter	Rhyme			
Expt 4, Before	7.94 (.18)	5.11 (.24)	6.97 (.19)			
Expt 4, After	6.71 (.22)	5.26 (.23)	5.41 (.21)			

Mean ratings on the strategy effectiveness questionnaire administered before and after the experiment, averaged across the judgment of learning (JOL) and no-JOL groups in Experiments 2–4. Ratings were made on a scale ranging from 1 (least effective) to 10 (most effective) for each cue type (category, letter, and rhyme). Standard error of the mean is in parentheses. Expt = Experiment

cues). Then, for participants in the category-advantaged condition, the targets for category pairs all came from the easy set of pairs whereas the targets for the letter pairs all came from the difficult set of items. By contrast, for participants in the no-advantage condition, the targets for letter pairs all came from the easy set of items, whereas the targets for the category pairs all came from the difficult set of items.

The procedure was similar to prior experiments. Because we found similar reactivity effects on the tests of cycles 1 and 2 in Experiments 1a and 1b, we included only one cycle in which participants studied a list of 30 pairs. Before and after the experiment, we administered an encoding strategy effectiveness questionnaire to assess participants' perception of category cues, letter cues, and rhyme cues (modified from Hertzog and Dunlosky's (2004) "personal encoding preferences" questionnaire). Participants were given an example of each cue type and asked to rate how effective they think each cue is for learning word pairs. They responded on a scale of 1–10 ranging from "least effective" to "most effective."

Results and discussion

Ratings of encoding strategy effectiveness

Average ratings for the effectiveness of category cues, letter cues, and rhyme cues are presented in the top portion of Table 2. Participants in both the category-advantaged and no-advantage groups rated category cues significantly more effective than letter cues both before and after the experiment (all $gs > .70$), and ratings did not significantly differ between the no-advantage and category-advantaged groups before the experiment. Ratings for the JOL and no-JOL groups did not significantly differ, with the exception of the category-advantaged group's ratings made after the experiment. For this group, effectiveness ratings for category cues were higher for participants who made JOLs ($M = 9.35$, $SE = 0.16$) compared to those who did not make JOLs ($M = 8.64$, $SE = 0.28$), $t(65) = 2.24$, $p = .029$, 95% CI [0.08, 1.36], $g_s = 0.54$, $BF_{10} = 1.73$.

For the no-advantage group, JOLs did not significantly differ between category and letter pairs (Table 1), possibly because JOLs may have been influenced by multiple cues, including the fact that letter pairs were normatively easier than category pairs (Begg et al., 1989). Although we did not find evidence that participants in the no-advantage JOL group were using pair type to inform their JOLs, they still rated category cues as more effective than letter cues on the encoding strategy effectiveness questionnaire both before the experiment; $t(35) = 4.17$, $p < .001$, 95% CI [1.32, 3.84], $g_{av} = 1.01$, $BF_{10} = 125.94$; and after the experiment; $t(35) = 5.02$, $p < .001$, 95% CI [1.75, 4.14], $g_{av} = 1.26$, $BF_{10} = 1359.55$.

Recall performance

Recall performance is presented in Fig. 4. We conducted a 2 (judgment group) × 2 (pair type) × 2 (recall group) mixed ANOVA on cued-recall performance. If reactivity only occurs for items with high levels of recall performance, then we would expect a 2 (judgment group) × 2 (pair type) interaction for the category-advantaged recall group but a smaller or non-significant 2 × 2 interaction for the no-advantage recall group. We observed no such three-way interaction, $F(1, 134) = 0.05$, $p = .83$, $\eta_p^2 < .001$. Our results support predictions from the cue-strengthening hypothesis – we observed positive reactivity for category pairs regardless of recall group.

Collapsed across the other variables, recall was higher for category pairs ($M = .80$, $SE = .02$) than for letter pairs (M

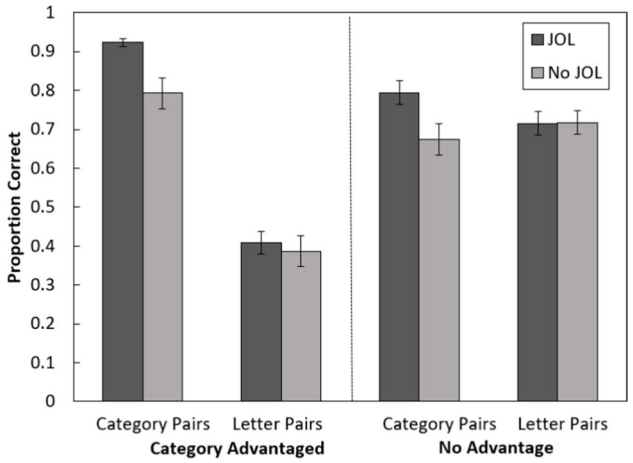

Fig. 4 Recall performance as a function of judgment group, pair type, and recall group in Experiment 2. *Note.* JOL = judgment of learning. Error bars reflect the standard error of each mean

= .56, *SE* = .02), $F(1, 134) = 230.38$, $p < .001$, $\eta_p^2 = 0.63$; recall was also higher for judged pairs (*M* = .71, *SE* = .02) than for non-judged pairs (*M* = .64, *SE* = .02), $F(1, 134) = 6.56$, $p = .012$, $\eta_p^2 = 0.05$. The judgment group by recall group interaction was not significant, $F(1, 134) = 0.11$, $p = .74$, $\eta_p^2 = 0.001$. The pair type by recall group interaction was significant, $F(1, 134) = 197.63$, $p < .001$, $\eta_p^2 = 0.60$ – not surprisingly given our recall group manipulation (i.e., recall was higher for category pairs than letter pairs in the category-advantaged group, and recall did not significantly differ for letter pairs than category pairs in the no-advantage group).

Critically, the judgment group by pair type interaction was significant, $F(1, 134) = 13.46$, $p < .001$, $\eta_p^2 = 0.09$. Collapsing across the two recall groups, independent-samples *t*-tests revealed that recall for category pairs was significantly higher for the JOL group (*M* = .86, *SE* = .02) than the no-JOL group (*M* = .73, *SE* = .03), $t(136) = 3.96$, $p < .001$, 95% CI [.06, .19], $g_s = 0.67$, $BF_{10} = 164.06$, whereas recall for letter pairs was not significantly different for the JOL group (*M* = .57, *SE* = .03) and the no-JOL group (*M* = .56, *SE* = .03), $t(136) = 0.22$, $p = .82$, 95% CI [-.07, .09], $g_s = 0.04$, $BF_{01} = 7.39$.

Given our manipulation of recall performance, item effects may be contributing to the interactions observed for the category-advantaged or no-advantage groups in Fig. 4. Note that even when the same items are compared (e.g., by comparing recall for the category pairs of the category-advantaged groups with the letter pairs of the no-advantage groups), a similar interaction was observed – positive reactivity for category pairs and no reactivity for letter pairs. Additionally, one could argue that by using normatively easy items, perhaps participants would be more inclined to attend to features of the target words themselves (e.g., their high concreteness or frequency) rather than the cue-target relationship. If so, we would not expect any reactivity effects, and the fact that we found positive reactivity for category pairs at least partially rules out this possibility.

Experiment 3

Results from Experiment 2 ruled out the memory-strength account. In the case of items cued by their first two letters, JOLs do not invoke additional processing – or at least not the type of (semantic) processing that would benefit performance on a cued-recall test – even when memory strength is increased. In contrast, for items cued with their category label, positive reactivity occurred regardless of the level of underlying memory strength or recall performance. Thus, results from Experiment 2 suggest that only a specific type of processing invoked by making JOLs (i.e., a strengthening of the semantic relationship between cues and targets) leads to positive reactivity for cued-recall tests. Another possibility – one which we attempted to rule out in Experiment 3 – is that positive reactivity would occur for category pairs even when the processing at encoding does not match requirements of the criterion test.

Recall that the cue-strengthening hypothesis has two components: (1) the associative information strengthened by the cues used to make JOLs and (2) the relevance of the strengthened information for boosting performance on the criterion test. In our next two experiments, we attempted to elucidate the encoding-retrieval dynamics for the category- and letter-cued items by investigating JOL reactivity with different criterion tests. The cue-strengthening hypothesis predicts that JOL reactivity will only occur on tests that are sensitive to the cues that inform JOLs. As an initial investigation into this idea, Myers et al. (2020) investigated JOL reactivity using a mixed list of related and unrelated word pairs. Half of the participants made JOLs during study, and half did not. According to the cue-strengthening hypothesis, JOL reactivity should only occur if the criterion test relies on the same cues that inform JOLs (i.e., cue-target semantic relationships). Thus, they predicted positive reactivity for related pairs on tests that are sensitive to cue-target relatedness (e.g., cued recall), but not for tests that are less sensitive to cue-target relatedness (e.g., free recall). Consistent with this hypothesis, positive reactivity occurred for related pairs on a cued-recall test, but no such reactivity on a free recall test – presumably because the strengthening of cue-target associations (from making JOLs) is less beneficial in the absence of cues (e.g., Rivers & Dunlosky, 2021).

In Experiment 3, we investigated JOL reactivity for category and letter pairs on a criterion test that either relies on the semantic associations between cues and targets or a test that does not rely on such associations. That is, after

studying a list of category and letter pairs, half of participants received a cued-recall test in which they received the cue they had studied the target with (as in prior experiments). The other half of participants always received a letter-cued recall test for all items, including the target items they had initially studied with category cues. As in prior experiments, some participants made JOLs for all items, whereas others did not. Based on the cue-strengthening hypothesis, we predicted positive reactivity for category pairs on the (category) cued-recall test (replicating the pattern of results from prior experiments). However, because letter-cued recall does not rely on a strengthening of the semantic association between cues and targets, the cue-strengthening hypothesis predicts that reactivity will not occur for category-cued targets on the letter-cued recall test.

Method

We used a 2 (judgment group) × 2 (pair type) × 2 (test group: cued recall vs. letter-cued recall) mixed design, with pair type manipulated within participant and judgment group and test group manipulated between participants. As in Experiment 2, our target sample size was 128 participants. A total of 146 undergraduates from TCU were randomly assigned to one of four groups: JOL, cued recall (n = 36); no JOL, cued recall (n = 36); JOL, letter-cued recall (n = 37); no JOL, letter-cued recall (n = 37).

Participants studied a list of 44 paired associates (the set of target words were randomly selected from Experiment 1). The procedure was similar to prior experiments in that participants studied pairs for an upcoming cued-recall test (and made JOLs, if in the JOL groups). Unlike prior experiments, participants completed a 3-min distractor task (i.e., list as many states of the United States as you can) between study and test. On the final test, participants in the cued-recall groups received the category or letter cue (depending on how they had studied the pair) and were asked to recall the corresponding target word (as in prior experiments; i.e., "When you studied the words, you sometimes received a letter clue (e.g., *wa – wave*) and you sometimes received a category clue (e.g., *something in the ocean – wave*). During this test, you will receive those same clues"). Meanwhile, participants in the letter-cued recall group always received the first two letters of the target words, regardless of how they were studied (i.e., "During this test, you will receive a letter clue for all of the words you studied"). Both test types were self-paced. Participants also completed the encoding strategy effectiveness questionnaire before and after the experiment (for details, refer to Experiment 2).

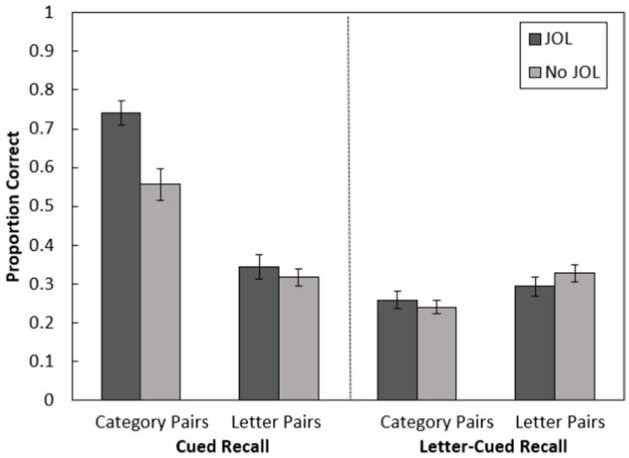

Fig. 5 Recall Performance as a Function of Judgment Group, Pair Type, and Test Group in Experiment 3. *Note.* JOL = judgment of learning. Error bars reflect the standard error of each mean

Results and discussion

Ratings of encoding strategy effectiveness

Average ratings for the effectiveness of category cues, letter cues, and rhyme cues are presented in the middle portion of Table 2. Participants both in the cued recall and letter-cued recall groups rated category cues as significantly more effective than letter cues both before and after the experiment (all gs > .95). The pattern of ratings did not significantly differ for the two judgment groups or the two test groups. For both test groups, JOLs did not significantly differ for category and letter pairs (Table 1).

Recall performance

Recall performance is presented in Fig. 5. We conducted a 2 (judgment group) × 2 (pair type) × 2 (test group) mixed ANOVA on cued-recall performance. Consistent with inspection of Fig. 5, the three-way interaction approached significance, $F(1,142) = 3.60$, $p = .06$, $\eta_p^2 = 0.03$, suggesting the pattern of reactivity varied by test type. Thus, we next present analyses of recall performance based on individual ANOVAs for each test group.

For the cued-recall group, recall was significantly higher for category pairs ($M = .65$, $SE = .03$) than letter pairs ($M = .33$, $SE = .02$), $F(1, 70) = 183.03$, $p < .001$, $\eta_p^2 = 0.72$. And, participants recalled more judged pairs ($M = .54$, $SE = .03$) than non-judged pairs ($M = .44$, $SE = .03$), $F(1, 70) = 7.25$, $p = .009$, $\eta_p^2 = 0.09$. These main effects were qualified by a significant interaction, $F(1, 70) = 11.30$, $p = .001$, $\eta_p^2 = 0.14$. Independent-samples t-tests

revealed that recall was significantly higher for category pairs in the JOL than for the no-JOL groups, $t(70) = 3.55$, $p < .001$, 95% CI [.08, .29], $g_s = 0.83$, $BF_{10} = 40.30$, whereas recall did not significantly differ for letter pairs in the JOL and no-JOL groups, $t(70) = 0.69$, $p = .49$, 95% CI [-.05, .10], $g_s = 0.16$, $BF_{01} = 4.48$. Thus, the pattern of recall results observed in Experiment 1 replicated.

For the letter-cued recall group, recall was significantly lower for category pairs ($M = .25$, $SE = .01$) than for letter pairs ($M = .31$, $SE = .02$), $F(1, 72) = 16.57$, $p < .001$, $\eta_p^2 = 0.19$. Recall for participants who made JOLs ($M = .28$, $SE = .02$) did not differ from those who did not make JOLs ($M = .28$, $SE = .02$), $F(1, 72) = 0.10$, $p = .76$, $\eta_p^2 = 0.001$. The interaction between pair type and judgment group was not significant, $F(1, 72) = 3.13$, $p = .08$, $\eta_p^2 = 0.04$. Thus, consistent with the prediction from the cue-strengthening hypothesis, no reactivity was observed on a test that did not require semantic associations between cues and targets.

Experiment 4

In Experiment 3, we found that reactivity did not occur for category-cued targets on a letter-cued recall test. To further evaluate the cue-strengthening hypothesis, we investigated JOL reactivity on another test that does not rely on semantic relationships between cues and targets – a free-recall test. Note that the prior investigation of JOL reactivity on a free-recall test (Myers et al., 2020) did not find reactivity for related pairs, presumably because strengthening the cue-target semantic relationship would not be relevant to the memory strength of targets relevant to recalling them without the corresponding cues. Accordingly, we did not expect JOL reactivity for category-cued targets for the free-recall test. By contrast, one possible reason that JOL reactivity has not arisen for letter pairs in the prior experiments reported here is that making JOLs enhances the memory strength of the target alone (given that the cue is simply a portion of the target itself), which a measure of associative strength (i.e., cued recall) would not be sensitive to reveal. If making a JOL does boost target strength in this case, then a prediction – which does not follow from the cue-strengthening account that concerns strengthening the cue-target association – is that JOL reactivity will occur for the letter-cued targets on the free-recall test.

Method

Our target sample size was 102 participants, which was determined with an a priori power analysis for an independent-samples *t*-test for the JOL versus no-JOL groups. We set power at .80 and $\alpha = .05$ to detect a medium effect of $g = 0.50$. A total of 129 KSU undergraduates participated ($n = 65$ and 64 randomly assigned to the JOL and no-JOL groups, respectively).

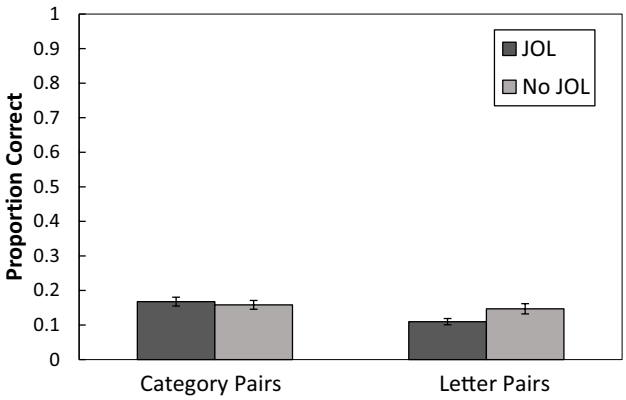

Fig. 6 Free-recall performance as a function of judgment group and pair type in Experiment 4. *Note*. JOL = judgment of learning. Error bars reflect the standard error of each mean

Participants studied a list of 44 targets (randomly selected from Experiment 1), a random half of which were presented with a category cue and the other half were presented with the first two letters as a cue. The procedure and instructions to participants were similar to prior experiments, except after presentation of the pairs (and a 2-min distractor task that involved listing country names), participants completed a free-recall test of the target words. The test was self-paced and participants were instructed to recall as many of the words on the right-hand side of the pairs as possible in the order that they came to mind. Participants spent an average of 2.47 min ($SE = 0.15$) attempting free recall.

Results and discussion

Ratings of encoding strategy effectiveness

Average ratings for the effectiveness of category cues, letter cues, and rhyme cues are presented in the bottom portion of Table 2. Participants rated category cues as significantly more effective than letter cues both before and after the experiment (*g*s > .57). And, JOLs were significantly higher for category than letter pairs (Table 1).

Recall performance

Free-recall performance is presented in Fig. 6. A 2 (judgment group) × 2 (pair type) mixed ANOVA revealed that recall was higher for category pairs ($M = .16$, $SE = .01$) than letter pairs ($M = .13$, $SE = .01$), $F(1, 127) = 12.38$, $p < .001$, $\eta_p^2 = 0.09$. The main effect of judgment group was not significant, $F(1, 127) = 0.89$, $p = .35$, $\eta_p^2 = 0.01$. The judgment group by pair type interaction was significant, $F(1, 127) = 5.60$, $p = .019$, $\eta_p^2 = 0.04$. Independent-samples

t-tests revealed free recall for category pairs did not significantly differ for the JOL and no-JOL groups, $t(127) = 0.52$, $p = .60$, 95% CI [-.03, .05], $g_s = 0.09$, $BF_{01} = 6.43$. However, recall for letter pairs was significantly *lower* for the JOL versus no-JOL group, $t(127) = -2.15$, $p = .035$, 95% CI [-.07, -.003], $g_s = 0.38$, $BF_{10} = 1.19$.

Patterns of recall suggest that JOLs did not enhance performance on a free-recall test, consistent with the hypothesis that reactivity depends on the overlap between cues used to inform JOLs and to retrieve items on a later criterion test (cf. Myers et al., 2020). That is, outcomes from this experiment provide preliminary support for the idea that (a) the positive reactivity effect for category pairs is primarily due to a strengthening of cue-target semantic associations rather than the strengthening of the representations of targets alone, and (b) that the lack of reactivity for letter pairs on cued recall tests (Experiments 1a, 1b, 2, and 3) does not necessarily result from JOLs boosting target strength for letter-cued targets (that the cued-recall tests would not reveal). However, because performance on free recall tests is influenced by factors other than memory for targets alone (e.g., inter-target relationships; Rivers & Dunlosky, 2021), future research should investigate reactivity with other test types (e.g., item recognition tests as in Myers et al., 2020).

General discussion

JOLs enhance cued recall for category-cued but not letter-cued items

Prior research investigating memory reactivity with JOLs has used related and unrelated word pairs as stimuli. Using these materials, a common recall pattern is positive reactivity for material with a sematic relationship (i.e., related pairs) and negative or no reactivity for those without such a relationship (i.e., unrelated pairs). In the present research, our goal was to provide critical, converging evidence by using another manipulation of a priori semantic relatedness. We used a levels-of-processing manipulation – words cued by either their category label or their first two letters – to investigate JOL reactivity and found a remarkably similar recall pattern: positive reactivity for material with a sematic relationship (i.e., category pairs) and no reactivity for those without such a relationship (i.e., letter pairs).

Experiments 1a and 1b investigated JOL reactivity across multiple study-test cycles, which allowed us to examine whether patterns of reactivity change after participants gain test experience. For example, perhaps participants in the no-JOL group could adopt enhanced encoding strategies for category pairs in anticipation of a cued-recall test (e.g., enhanced cue-target relational processing; Rivers & Dunlosky, 2021) after gaining test experience on cycle 1, and if so, recall for the no-JOL group would be comparable to the JOL group on cycle 2. We did not find evidence for this possibility – patterns of reactivity were similar across cycles.

In Experiment 2, we ruled out an important alternative explanation for reactive effects. Namely, that reactivity occurs only for conditions that produce a high level of recall performance (see Fig. 3), such as category pairs (as in the current experiments) or related pairs (as in prior research). However, even when recall performance was approximately equated and relatively high both for category and letter pairs (that were not judged), reactivity only occurred for category pairs (Fig. 4).

Finally, in Experiments 3 and 4, we administered criterion tests that did not rely on a strengthening of a semantic relationship between cues and targets. On both a letter-cued recall test (Fig. 5) and a free-recall test (Fig. 6), no reactivity occurred for category pairs. Thus, positive reactivity for category pairs seems to be driven by a strengthening of cue-target associations rather than the strengthening of the representations of targets alone. And, for the letter pairs, this evidence provides preliminary support for the conclusion that JOLs do not strengthen memory for the targets. Overall, these outcomes are consistent with the cue-strengthening hypothesis, which provides a parsimonious explanation for reactivity effects – at least for material with semantic relationships.

Status of hypotheses for JOL reactivity effects

To what degree can the cue-strengthening hypothesis account for all the data published in this growing area of JOL-reactivity research, and how do the results from the present experiments inform other hypotheses presented in the literature? Much of the research to date, including the current set of experiments, supports the cue-strengthening hypothesis. Moreover, recent evidence from Halamish and Undorf (2022) suggests that making JOLs can benefit cued-recall performance for identical word pairs (e.g., *beach – beach*), and improves accuracy for judgments made about whether a cue appeared with an unrelated, related, or identical target during study. The authors argue these results are consistent with a *relatedness-processing assumption* of the cue-strengthening hypothesis, which states that people process cue-target relatedness more when making JOLs than when they do not make JOLs. Specifically, making JOLs for semantically related material (e.g., related pairs, category pairs) improves cued-recall performance through cue-strengthening and relatedness processing, whereas identical pairs benefit only from relatedness processing (Halamish & Undorf, 2022).

However, as mentioned in the *Introduction*, some research has found negative reactivity for unrelated pairs, although these effects are always smaller than the positive reactivity observed for related pairs (refer to Fig. 2). In attempt to explain this negative reactivity, the *changed-goal hypothesis* and the *dual-task hypothesis* have been proposed (Mitchum et al., 2016).

The changed-goal hypothesis proposes that judging future memory performance leads participants to consider that some items will be remembered and some will not. As a result, participants change their learning goal from trying to learn all of the items presented during study and instead focus on learning only the easy items at the expense of the more difficult items. Thus, this hypothesis predicts positive reactivity for normatively easier learning material and negative reactivity for normatively more difficult material. In contrast to this prediction, we did not find any trends of negative reactivity for letter pairs in Experiments 1a and 1b. Although speculative, in Experiment 2, this hypothesis might predict an exaggerated changed-goal pattern (i.e., particularly robust positivity for category pairs and negativity for letter pairs) for the category-advantaged groups (because category pairs were normatively easier and letter pairs were normatively more difficult), or a reversed pattern of reactivity for the no-advantage groups (i.e., positive reactivity for the normatively easier letter pairs and negative reactivity for the category pairs). We did not find support for this prediction (Fig. 4). Of course, study time was experimenter-paced in the current investigation (rather than self-paced), so we could not investigate whether learners adopted different learning agendas depending on the to-be-learned material (e.g., spending less study time on more difficult items as in Janes et al., 2018 and Mitchum et al., 2016).

The dual-task hypothesis proposes that eliciting JOLs could interfere with the primary task of memorizing to-be-learned material, particularly when the material is difficult to learn. In the case of related and unrelated pairs, this hypothesis predicts that unrelated pairs are processed less fully when JOLs are elicited (compared to when they are not), thus explaining the negative reactivity – or at the very least, the lack of positive reactivity – observed for those pairs. In the present research, we observed no negative reactivity for letter pairs as this hypothesis would predict (except in Experiment 4 when we administered a free-recall test), but perhaps the processing demands are lighter for letter pairs than they are for unrelated pairs.

Another approach to evaluating hypotheses of JOL reactivity is to investigate effects across a variety of material types. In the current research, we critically extended the typical findings with related and unrelated word pairs – positive JOL reactivity for material with a semantic relationship (i.e., category pairs), but only when the test was sensitive to such strengthening. Researchers have also investigated JOL reactivity with lists of single words (e.g., Halamish, 2018; Li et al., 2022; Senkova & Otani, 2021; Tauber & Rhodes, 2012; Tekin & Roediger, 2020; Yang et al., 2015; Zhao et al., 2022), image pairs (Shi et al., 2022), and educational text passages (Ariel et al., 2021). For example, Senkova and Otani (2021) found that making JOLs enhanced item-specific processing of individual words, which improved memory performance for categorized lists by promoting distinctiveness processing (e.g., Hunt & Einstein, 1981). In the current Experiment 4, JOL reactivity for recalling single target words was not found, but our lists were not words from a set of categories, so such distinctive processing would be less likely to occur.

In research by Tekin and Roediger (2020), participants learned a list of single words (e.g., *apple*) and completed various orienting tasks based on letter case, category membership, or rhyme (e.g., *Does the word rhyme with "chapel"?*). Half of the participants made item-by-item JOLs for each word predicting future recognition performance, and all participants completed an immediate old/new recognition test. Positive reactivity occurred for judged words, but the reactivity effect was larger for items processed shallowly (e.g., through perceptual orienting tasks) compared to those processed deeply (e.g., through semantic orienting tasks). The authors argue that making JOLs "improves retention especially if [they] strengthen information that is not strengthened otherwise" (p. 288). Because category-based orienting tasks already promote semantic processing, the memorial benefit of making JOLs is not as strong as what is found for tasks that involve more shallow processing – at least on recognition tests. These results appear to be inconsistent with predictions of the cue-strengthening hypothesis and results found in the present research. However, given the different methodologies used between our research and Tekin and Roediger (2020) – materials being word pairs versus single words, studying pairs versus performing an orienting task, completing recall versus recognition tests, etc. – future research should aim to replicate and compare various methods to develop a comprehensive account of JOL reactivity across all material and test types.

Yet another approach to evaluating hypotheses of JOL reactivity is to investigate individual differences (e.g., Tauber & Witherby, 2019; Zhao et al., 2022). For example, Tauber and Witherby (2019) investigated JOL reactivity with older and younger adults. Participants were randomly assigned to either study alone or to study and make JOLs for a list of related pairs. Across five experiments, the authors found positive JOL reactivity for younger adults, but no reactivity for older adults. The authors propose that JOLs may not have strengthened the cue-target association to the same degree across age groups because of a processing deficit for older adults. That is, given that older adults presumably have more difficulties in associating words in a pair (e.g., Naveh-Benjamin, 2000), the small benefit to cue strengthening (i.e., between a cue and its target) may not be sufficient to overcome the age-related processing deficiency. Perhaps most important to emphasize here given that research in this area is in its infancy, the aforementioned hypotheses are not mutually exclusive. That is, all the mechanisms (e.g., cue strengthening, goal changing) may contribute to reactivity patterns, and perhaps some are active in one context whereas others are active in another.

Accordingly, a challenge for future research will involve developing methods to estimate the contribution of each mechanism to JOL reactivity across a variety of contexts and individual differences. Myers et al. (2020) propose that Jenkins' (1979; see also Roediger, 2008) tetrahedral model of memory is a useful perspective for guiding future investigations. Specifically, considering a combination of factors – participants, materials, encoding conditions, and retrieval conditions – and how they influence JOLs, memory, and JOL reactivity may aid in developing a comprehensive account of JOL reactivity.

Concluding comment

In summary, the present research provides the first demonstration of JOL reactivity using category and letter pairs. Converging with prior research using related and unrelated word pairs as stimuli (e.g., Soderstrom et al., 2015), we found positive reactivity for material with an a priori sematic relationship – but only on tests that were sensitive to the strengthening of this relationship – which is consistent with a core prediction of the cue-strengthening hypothesis.

Acknowledgements The authors gratefully acknowledge Rachel Hall, Hannah Barringer, Alex Knopps, Emily Moore, Abby O'Brien, Jacob O'Connor, Bailey Patouhas, and research assistants in the Tauber Lab for their assistance with data collection.

Code availability Not applicable.

Funding The authors did not receive support from any organization for the submitted work.

Data Availability The datasets generated and analyzed during the current study are available on the Open Science Framework at https://osf.io/84q2p/.

Declarations

Competing interests The authors have no relevant financial or non-financial interests to disclose.

Ethics approval All experiments reported received approval from the Institutional Review Boards at Kent State University and Texas Christian University.

Consent to participate Informed consent was obtained from all individual participants included in the experiments.

Consent for publication Not applicable.

References

Articles that contributed values for Figure 2 are preceded by an asterisk (*).

Ariel, R., Karpicke, J. D., Witherby, A. E., & Tauber, S. K. (2021). Do judgments of learning directly enhance learning of educational materials? *Educational Psychology Review, 33*, 693–712. https://doi.org/10.1007/s10648-020-09556-8

Begg, I., Duft, S., Lalonde, P., Melnick, R., & Sanvito, J. (1989). Memory predictions are based on ease of processing. *Journal of Memory and Language, 28*(5), 610–632. https://doi.org/10.1016/0749-596X(89)90016-8

Bieman-Copland, S., & Charness, N. (1994). Memory knowledge and memory monitoring in adulthood. *Psychology and Aging, 9*(2), 287–302. https://doi.org/10.1037/0882-7974.9.2.287

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*(6), 671–684. https://doi.org/10.1016/S0022-5371(72)80001-X

deWinstanley, P. A., & Bjork, E. L. (2004). Processing strategies and the generation effect: Implications for making a better reader. *Memory & Cognition, 32*(6), 945–955. https://doi.org/10.3758/BF03196872

Double, K. S., & Birney, D. P. (2019). Reactivity to measures of metacognition. *Frontiers in Psychology, 10*, 2755. https://doi.org/10.3389/fpsyg.2019.02755

Double, K. S., Birney, D. P., & Walker, S. A. (2017). A meta-analysis and systematic review of reactivity to judgements of learning. *Memory, 26*(6), 741–750. https://doi.org/10.1080/09658211.2017.1404111

Dunlosky, J., & Matvey, G. (2001). Empirical analysis of the intrinsic–extrinsic distinction of judgments of learning (JOLs): Effects of relatedness and serial position on JOLs. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(5), 1180–1191. https://doi.org/10.1037/0278-7393.27.5.1180

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*(2), 175–191. https://doi.org/10.3758/BF03193146

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

Halamish, V., & Undorf, M. (2022). Why do judgments of learning modify memory? Evidence from identical pairs and relatedness judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition.* https://doi.org/10.1037/xlm0001174

Hertzog, C., & Dunlosky, J. (2004). Aging, metacognition, and cognitive control. In B. H. Ross (Ed.), *The psychology of learning and motivation: Advances in research and theory* (pp. 215–252). Elsevier Academic Press.

Hunt, R. R., & Einstein, G. O. (1981). Relational and item-specific information in memory. *Journal of Verbal Learning and Verbal Behavior, 20*(5), 497–514. https://doi.org/10.1016/S0022-5371(81)90138-9

* 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

Jenkins, J. J. (1979). Four points to remember: A tetrahedral model of memory experiments. In L. S. Cermak & F. I. M. Craik (Eds.), *Levels of processing in human memory* (pp. 429–446). Psychology Press.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126*(4), 349–370. https://doi.org/10.1037/0096-3445.126.4.349

Kruschke, J. K. (2013). Bayesian estimation supersedes the *t* test. *Journal of Experimental Psychology: General, 142*(2), 573–603. https://doi.org/10.1037/a0029146

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs.

Frontiers in Psychology, 4, 863. https://doi.org/10.3389/fpsyg.2013.00863

Li, B., Zhao, W., Zheng, J., Hu, X., Su, N., Fan, T., ... & Luo, L. (2022). Soliciting judgments of forgetting reactively enhances memory as well as making judgments of learning: Empirical and meta-analytic tests. *Memory & Cognition*, *50*(5), 1061-1077. https://doi.org/10.3758/s13421-021-01258-y

Loftus, G. R. (1978). On interpretation of interactions. *Memory & Cognition, 6*(3), 312–319. https://doi.org/10.3758/BF03197461

Matvey, G., Dunlosky, J., Shaw, R. J., Parks, C., & Hertzog, C. (2002). Age-related equivalence and deficit in knowledge updating of cue effectiveness. *Psychology and Aging, 17*(4), 589–597. https://doi.org/10.1037/0882-7974.17.4.589

Maxwell, N. P., & Huff, M. J. (2022). Reactivity from judgments of learning is not only due to memory forecasting: Evidence from associative memory and frequency judgments. *Metacognition and Learning*, 1–37. https://doi.org/10.1007/s11409-022-09301-2

* 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

Morey, R. D., & Rouder, J. N. (2018). BayesFactor: Computation of Bayes Factors for Common Designs (R Package Version 0.9.12-4.2) [Computer software]. Retrieved from https://CRAN.R-project.org/package=BayesFactor.

Mueller, M. L., Tauber, S. K., & Dunlosky, J. (2013). Contributions of beliefs and processing fluency to the effect of relatedness on judgments of learning. *Psychonomic Bulletin & Review, 20*(2), 378–384. https://doi.org/10.3758/s13423-012-0343-6

Mueller, M. L., Dunlosky, J., & Tauber, S. K. (2015). Why is knowledge updating after task experience incomplete? Contributions of encoding experience, scaling artifact, and inferential deficit. *Memory & Cognition, 43*(2), 180–192. https://doi.org/10.3758/s13421-014-0474-2

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

Naveh-Benjamin, M. (2000). Adult age differences in memory performance: Tests of an associative deficit hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(5), 1170–1187. https://doi.org/10.1037/0278-7393.26.5.1170

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

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

Rivers, M. L., & Dunlosky, J. (2021). Are test-expectancy effects better explained by changes in encoding strategies or differential test experience? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 47*(2), 195–207. https://doi.org/10.1037/xlm0000949

* Rivers, M. L., Janes, J. L., & Dunlosky, J. (2021). Investigating memory reactivity with a within-participant manipulation of judgments of learning: Support for the cue-strengthening hypothesis. *Memory, 29*(10), 1342–1353. https://doi.org/10.1080/09658211.2021.1985143

Roediger, H. L. (2008). Relativity of remembering: Why the laws of memory vanished. *Annual Review of Psychology, 59*, 225–254. https://doi.org/10.1146/annurev.psych.57.102904.190139

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*(2), 225–237. https://doi.org/10.3758/PBR.16.2.225

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

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

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science, 9*(1), 76–80. https://doi.org/10.1177/1745691613514755

Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science, 10*(2), 176–199. https://doi.org/10.1177/1745691615569000

* 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

Storm, B. C., Hickman, M. L., & Bjork, E. L. (2016). Improving encoding strategies as a function of test knowledge and experience. *Memory & Cognition, 44*(4), 660–670. https://doi.org/10.3758/s13421-016-0588-9

Tauber, S. K., & Rhodes, M. G. (2012). Measuring memory monitoring with judgements of retention (JORs). *Quarterly Journal of Experimental Psychology, 65*(7), 1376–1396. https://doi.org/10.1080/17470218.2012.656665

Tauber, S. K., & Witherby, A. E. (2019). Do judgments of learning modify older adults' actual learning? *Psychology and Aging, 34*(6), 836–847. https://doi.org/10.1037/pag0000376

Tekin, E., & Roediger, H. L. (2020). Reactivity of judgments of learning in a levels-of-processing paradigm. *Zeitschrift für Psychologie, 228*(4), 278–290. https://doi.org/10.1027/2151-2604/a000425

Wagenmakers, E. J., Krypotos, A. M., Criss, A. H., & Iverson, G. (2012). On the interpretation of removable interactions: A survey of the field 33 years after Loftus. *Memory & Cognition, 40*(2), 145–160. https://doi.org/10.3758/s13421-011-0158-0

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

Zhao, W., Li, B., Shanks, D. R., Zhao, W., Zheng, J., Hu, X., ... & Yang, C. (2022). When judging what you know changes what you really know: Soliciting metamemory judgments reactively enhances children's learning. *Child Development, 93*(2), 405-417. https://doi.org/10.1111/cdev.13689

Open practices statement The data and materials for all experiments are available at https://osf.io/84q2p/. None of the experiments were preregistered.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.