Psychonomic Bulletin & Review (2018) 25:2356–2364
https://doi.org/10.3758/s13423-018-1463-4

BRIEF REPORT

The influence of making judgments of learning on memory performance: Positive, negative, or both?

Jessica L. Janes[1] · Michelle L. Rivers[1] · John Dunlosky[1]

Published online: 2 April 2018

Abstract

A common measure of memory monitoring——judgments of learning (JOLs)——has recently been shown to have reactive effects on learning. When participants study a list of related and unrelated word pairs, they recall more related than unrelated pairs. This *relatedness effect* is larger when people make JOLs than when they do not make them. Evidence is mixed concerning whether this *increased* relatedness effect arises because JOLs help memory for related pairs, hurt it for unrelated pairs, or do both. In three experiments, we investigated (1) the nature of the increased relatedness effect (i.e., does it arise from positive reactivity for related pairs, negative reactivity for unrelated pairs, or both?) and (2) the mechanisms underlying the effect. Participants studied cue–target word pairs and either did (or did not) make immediate JOLs and then completed a cued-recall test. When participants studied a mixed list consisting of related and unrelated pairs, the increased relatedness effect was largely driven by positive reactivity. When participants studied pure lists consisting solely of related or unrelated word pairs (Experiment 2 only), the increased relatedness effect was minimized. These and other findings suggest that making JOLs helps learning more than hurts it, and that this reactive effect partly occurs because making JOLs changes people's learning goals.

Keywords Metamemory · Reactivity · Judgments of learning · Monitoring

Can prompting students to judge their learning change the ongoing learning process, ultimately impacting memory? If so, does this act help or harm memory, and what are the practical implications of such reactivity? For instance, should students be told to judge their learning frequently during study because doing so enhances learning, or should they be told that doing so may harm their learning? Relevant to these questions, the potential reactivity of metamemory judgments have been scrutinized ever since Spellman and Bjork (1992) argued that delayed *judgments of learning* (JOLs) enhance subsequent memory (for a review of the reactive effects of delayed JOLs, see Rhodes & Tauber, 2011). More recently, the possible reactive effects of *immediate* JOLs have been investigated (Mitchum, Kelley, & Fox, 2016; Soderstrom, Clark, Halamish, & Bjork, 2015). In a typical experiment, participants are presented with some to-be-learned items, and after studying each item, they are prompted to judge the likelihood (typically on a 0%–100% scale) that they will be able to recall that item on a future test. According to Ericsson and Simon (1980), verbal reports like these will be reactive when participants attend to information to which they would not otherwise attend. For example, when participants are prompted to judge their learning, they tend to make inferences about how various factors within the learning context will influence memory (Koriat, 1997). This inferential processing (which presumably would not occur in the absence of JOLs) may lead to changes in how the to-be-learned items are encoded and subsequent recall performance.

Consistent with this possibility, two studies have recently demonstrated that making JOLs can influence memory performance (Mitchum et al., 2016; Soderstrom et al., 2015). However, it is unclear exactly *how* making JOLs influences memory (e.g., do they have a positive or negative impact?), and *why* making JOLs influences memory (e.g., do they cause learners to approach the task differently?). Accordingly, in three experiments, we aimed (1) to clarify the size and nature of any reactive effects produced by JOLs and (2) to test theoretical accounts that have been proposed to explain JOL reactivity. For the remainder of the

✉ Jessica L. Janes
jjanes1@kent.edu

[1] Department of Psychological Sciences, Kent State University, Kent, OH 44240, USA

introduction, we review prior work examining JOL reactivity, focusing on the conflicting findings that have emerged from the aforementioned studies.

Despite the potential for reactivity, JOLs are commonly employed in metamemory research without the inclusion of a no-JOL control group. When a control group is included, JOLs are considered reactive when recall significantly differs between those who judged their learning and those who did not. Some studies have reported higher recall for judged items over nonjudged items, which is referred to as *positive reactivity* (e.g., Dougherty, Scheck, Nelson, & Narens, 2005; Zechmeister & Shaughnessy, 1980); others, however, have reported lower recall for judged items compared to nonjudged items, which is referred to as *negative reactivity* (Mitchum et al., 2016); and others have reported no reactive effects (Kelemen & Weaver, 1997; Kornell & Bjork, 2008b; Tauber & Rhodes, 2012). In these studies, the main goal was not to investigate reactivity, so other variables may have covaried with making (vs. not making) JOLs. For instance, time on task was not always held constant between JOL versus no-JOL groups, which undermines the interpretation of any reactive effects.

Whether making immediate JOLs helps or harms memory remained an open question until studies by Mitchum et al. (2016) and Soderstrom et al. (2015) were specifically designed to estimate such an influence. In both studies, participants studied a list of cue–target word pairs, half of which were related (e.g., *feathers–bird*) and half of which were unrelated (e.g., *practice–tree*). Some participants made JOLs following the study of each pair (JOL group) and some did not (no-JOL group). Presentation time was held constant for the two groups. Outcomes from both studies revealed a similar interaction between judgment group and relatedness. The no-JOL group demonstrated a classic *relatedness effect* on memory, such that recall was higher for related pairs than for unrelated pairs. More important, the JOL group demonstrated an *increased* relatedness effect, such that the difference in recall between related and unrelated pairs was larger for the JOL than for the no-JOL group. Consider one explanation for why the relatedness effect increased for the JOL group. According to the *changed-goal hypothesis* (Mitchum et al., 2016), making JOLs encourages participants to notice that some pairs will be remembered and some will not. As a consequence, participants shift their learning goal away from mastering all pairs and instead focus on learning the relatively easy pairs at the expense of the more difficult ones. This shift then translates into a larger difference in recall between related and unrelated pairs for the JOL than no-JOL group (i.e., the increased relatedness effect). As discussed further below, we conducted new tests of the changed-goal hypothesis, but another goal was to further establish the nature of the increased relatedness effect.

In particular, although outcomes from both studies revealed the increased relatedness effect, the nature of the interactions may have been different. Soderstrom et al. (2015) reported that it was largely driven by positive reactivity, wherein recall for related pairs was higher for the JOL than for the no-JOL group; recall differences between groups were minimal for unrelated pairs. By contrast, Mitchum et al. (2016) reported that the increased relatedness effect was largely driven by negative reactivity, wherein recall for unrelated pairs was lower for the JOL group; some trends toward positive reactivity for related pairs occurred in Mitchum et al. (2016), but they were not significant.

Given these discrepancies, one question emerges: Why might the increased relatedness effect be driven by positive reactivity in one case and negative reactivity in the other? One possibility is a simple difference in methods. Whereas participants could pace their study in Mitchum et al. (2016; but see Experiment 5), participants' study time was experimenter-paced (8 s total to study and judge each pair) in Soderstrom et al. (2015). Another possibility, however, is that the difference in reported outcomes across these prior studies is more apparent than real, especially given that Mitchum et al. (2016) found some trends toward positive reactivity.

Assuming that making JOLs results in an increased relatedness effect, does this impact of JOLs arise from positive reactivity for related pairs, negative reactivity for unrelated pairs, or both? A primary goal of the present research was to answer this question by estimating the relative contributions of both positive and negative reactivity to the increased relatedness effect. Because the contributions of positive and negative reactivity may have been different in prior research (Mitchum et al., 2016; Soderstrom et al., 2015), Experiment 1 was designed to replicate these outcomes using procedures from both studies; that is, one group had experimenter-paced study and another group self-paced their study. Given the importance of replication (e.g., Simmons, Nelson, & Simonsohn, 2011), Experiment 2 was designed to replicate and extend the findings from the experimenter-paced groups as well as to test theoretical accounts of the effect. In particular, we evaluated competitive predictions made from the changed-goal hypothesis and those from a dual-mechanism account. Finally, in Experiment 3, we increased our power to better estimate the relative contributions of both positive and negative reactivity to the increased relatedness effect.

Experiment 1

Our main goal of Experiment 1 was to explore JOL reactivity using procedures from both Soderstrom et al. (2015) and Mitchum et al. (2016). Participants studied related (e.g., *railroad–train*) and unrelated (e.g., *practice–tree*) word pairs, and after studying each pair, half of the participants made a JOL and half did not. Study was either experimenter-paced or self-paced. If pace was contributing to differences in the prior

studies, then the outcomes from the experimenter-paced groups should correspond to those from Soderstrom et al. (2015), and outcomes from the self-paced groups should correspond to those from Mitchum et al. (2016).

Method

Design and participants Experiment 1 used a 2 (judgment group: JOL vs. no-JOL) × 2 (pace group: experimenter-paced vs. self-paced) × 2 (relatedness: related vs. unrelated) mixed design, with judgment group and pace manipulated between participants, and relatedness manipulated within participant. A sample size of 144 individuals was determined by a power analysis using an effect size of d = .69 (from the positive reactivity reported by Soderstrom et al., 2015, Experiment 1b), with the power set at .80 and alpha at .05. Participants were college students who participated for partial course credit. Data were excluded for two participants in the experimenter-paced group who failed to follow instructions and one participant in the self-paced group for which the computer malfunctioned.

Materials Participants were run in small groups on computers. Stimuli consisted of 60 cue–target word pairs (from Soderstrom et al., 2015), half of which were related (mean forward associative strength = 0.57) and half of which were unrelated (according to Nelson, McEvoy, & Schreiber, 2004).

Procedure For experimenter-paced study, the following procedure was adopted from Soderstrom et al. (2015). In the no-JOL group (n = 35), each pair was presented for 8 s, after which the screen would advance to the next pair. In the JOL group (n = 35), each pair was also presented for 8 s, but halfway through the presentation of that pair (i.e., after 4 s), participants were prompted to make their JOL (i.e., judge the likelihood, on a 0%–100% scale, that they could correctly recall the target when presented with a cue).

For self-paced study, the following procedure was adopted from Mitchum et al. (2016). In the no-JOL group (n = 35), participants studied each pair for as long as they wished and clicked the "next" button to advance to the next pair. A blank screen was inserted in between the presentation of each pair for .5 s (Mitchum et al. included this interstimulus interval to try to equate for the length of the study phases between the two groups). In the JOL group (n = 36), participants were prompted to make their JOL after studying each pair. The pair did not remain on the screen while participants made their JOL.

Following study, all participants engaged in a 3-min arithmetic task and then completed a self-paced cued-recall test, with the order of the pairs being randomized anew for each participant.

Results

As noted above, we based sample size on outcomes from a power analysis aimed at detecting positive reactivity for related pairs (Soderstrom et al., 2015)——to foreshadow, we powered for negative reactivity in Experiment 3. Our analytic approach attempted to replicate outcomes from Soderstrom et al. (2015), who used experimenter-paced study, and outcomes from Mitchum et al. (2016), who used self-paced study. Thus, we present analyses of these groups separately, but for interested readers, we present results from the omnibus analysis of variance (ANOVA) including both pace groups in the Appendix.

Experimenter-paced recall performance Mean recall for the experimenter-paced groups is presented in Fig. 1. A 2 (judgment group) × 2 (relatedness) ANOVA revealed a significant main effect of associative relatedness, with recall being higher for related pairs than unrelated pairs, $F(1, 68) = 557.27$, $p < .001$, $\eta_p^2 = .89$, but no main effect of judgment group, $F(1, 68) = .01$, $p = .91$. More important, an increased relatedness effect was found: A significant interaction occurred between relatedness and judgment group, $F(1, 68) = 22.03$, $p < .001$, $\eta_p^2 = .25$, such that the difference in recall between the related and unrelated pairs was larger for the JOL group (d = 3.81) than for the no-JOL group (d = 1.63). Planned comparisons revealed that recall for related pairs was higher for the JOL than no-JOL group, $t(68) = 2.55$, $p = .01$, $d = .61$, and that recall for unrelated pairs tended to be lower for the JOL group, $t(68) = 1.77$, $p = .08$, $d = .42$.

Self-paced recall performance A 2 (judgment group) × 2 (relatedness) mixed ANOVA revealed a significant main effect of relatedness (Fig. 2), with recall being higher for related pairs than unrelated pairs, $F(1, 69) = 199.78$, $p < .001$, $\eta_p^2 = .74$, but no main effect of judgment group, $F(1, 69) = .99$, $p = .91$. The difference in recall between the related and unrelated pairs was larger for the JOL group (d = 1.91) than for the no-JOL group (d = 1.35), consistent with demonstrating an

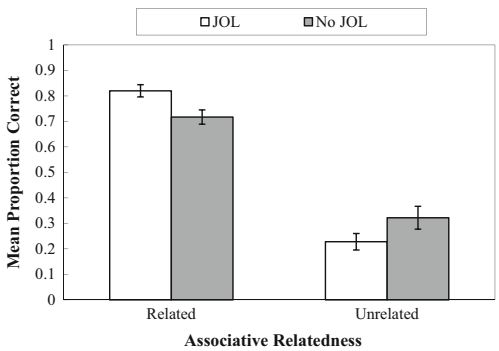

Fig. 1 Mean recall performance for related and unrelated pairs for the experimenter-paced groups in Experiment 1. Error bars represent the standard error of each mean

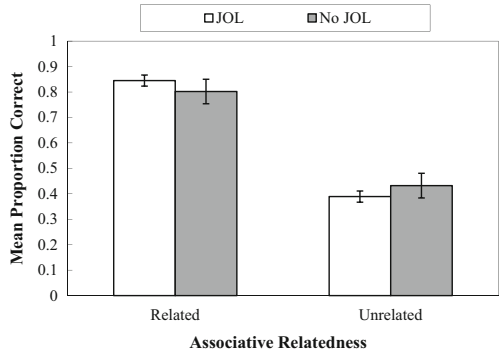

Fig. 2 Mean recall performance for related and unrelated pairs for the self-paced groups in Experiment 1. Error bars represent the standard error of each mean

increased relatedness effect. Nevertheless, the interaction did not reach conventional significance, $F(1, 69) = 2.21$, $p = .14$. Because our intent was to estimate the relative contributions of positive and negative reactivity to the increased relatedness effect, we still conducted planned comparisons. As shown in Fig. 2, recall for related pairs tended to be higher for the JOL than no-JOL group, $t(69) = 1.38$, $p = .17$, $d = .33$; differences in recall for unrelated pairs were negligible between the judgment groups, $t(69) = .64$, $p = .52$, $d = .15$.

Study time As in Mitchum et al. (2016), we also analyzed self-paced study times. According to the changed-goal hypothesis, when people study and do not make JOLs, they adopt a mastery learning approach wherein study time is driven by item difficulty (e.g., Koriat, Ma'ayan, & Nussinson, 2006). In this case, the no-JOL group would be expected to show a negative correlation between item difficulty and study time, as participants should allocate more time studying the more difficult, unrelated pairs. However, if making JOLs highlights that some items will not be remembered, then those in the JOL group should de-emphasize mastery. Thus, Mitchum et al. (2016) argued that the JOL group (relative to the no-JOL group) will demonstrate a weaker negative correlation. We evaluated this prediction (as in Mitchum et al., 2016) using latent semantic analysis (LSA) that results in an objective index of cue–target similarity (Landauer, Foltz, & Laham, 1998). We then computed the correlation between LSA similarity scores and study times. Means across participant's correlations are presented in Table 1 along with mean study times. The correlations were significantly negative (less than zero; ps $< .05$), but they did not differ significantly between groups, $t(68) = .18$, $p = .86$. A 2 (judgment group) × 2 (relatedness) ANOVA on study times revealed that more time was allocated to unrelated than related pairs, $F(1, 69) = 35.38$, $p < .001$, $\eta_p^2 = .34$. The main effect of judgment group, $F(1, 69) = 1.15$, $p = .29$, and the interaction, $F(1, 69) = 2.58$, $p = .11$, were not significant.

Table 1 Self-paced study times for the JOL and no-JOL groups in Experiment 1

		Group	
Group	Correlation	Mean Related	Unrelated
JOL	−0.18* (.03)	4.17 (.37)	6.13 (.68)
No-JOL	−0.17* (.03)	4.70 (.74)	8.11 (1.39)

Correlation = mean correlation across individual participant's correlations between self-paced study time and latent semantic analysis similarity scores (for details, see text). *Correlation is significantly different from zero ($p < .05$). Related/Unrelated = mean self-paced study time in seconds on each pair. Values in parentheses refer to the standard error of each mean

Discussion

Given that the nature of the increased relatedness effect may have been different in prior studies, Experiment 1 was designed to replicate findings from both. We replicated the increased relatedness effect using both procedures, but the increase in this effect was not as robust with self-paced study (JOL group, $d = 1.91 \geq$ no-JOL group, $d = 1.35$) as it was with experimenter-paced study (JOL group, $d = 3.81 >$ no-JOL group, $d = 1.63$). For the experimenter-paced groups where the increased relatedness effect was significant, positive reactivity ($d = .61$) was a larger contributor than was negative reactivity ($d = .42$). Even for the self-paced groups, the trend toward an increased relatedness effect appeared to be driven more by positive ($d = .33$) than negative ($d = .15$) reactivity.

One unexpected outcome was that the increased relatedness effect was less robust with self-paced than experimenter-paced study, even though our self-paced procedure was modeled after Mitchum et al. (2016), who consistently demonstrated the effect. Nevertheless, the changed-goal hypothesis provides an explanation for the reduction in the increased relatedness effect with self-paced study. It assumes that JOLs cause participants to shift their learning goal away from mastery, and prior research has shown that limited study time also leads participants to abandon a mastery approach (Metcalfe & Kornell, 2003). For the experimenter-paced groups, making JOLs under limited study time produces a strong shift toward learning easier items, which presumably yields a large increased relatedness effect. By contrast, participants in the self-paced groups may be less likely to shift their goal if they believe that the unlimited study time would allow them to master all items (Son & Metcalfe, 2000; Thiede & Dunlosky, 1999). Consistent with this possibility, the lack of difference in correlations (between study time and difficulty) for the JOL and no-JOL groups suggests that making JOLs had a minimal influence on participants' goals—if so, one would also not expect a major impact of JOLs on recall (as per Fig. 2).

Experiment 2

Results from Experiment 1 suggest that the increased relatedness effect arises more from positive than negative reactivity. Although the changed-goal hypothesis can explain these findings, another possibility is that positive and negative reactivity arise from two separate mechanisms, which we will refer to as the *dual-mechanism account*. Specifically, positive reactivity may be driven by an increased association between the cue and the target words studied, whereas negative reactivity may be driven by the difficulty of performing a judgment and memory task concurrently. According to Soderstrom et al. (2015), the act of making a JOL strengthens the cues (in this case, relatedness) used to make that judgment. If a later test is then sensitive to those cues, recall will be higher when JOLs are made than when they are not made. Positive reactivity then occurs for related pairs (but not unrelated pairs) because the cue–target relationship is strong, allowing participants to strengthen an already meaningful relationship. Negative reactivity results from dual-task costs that arise from the requirement to monitor one's learning while attempting to learn difficult pairs. When two tasks are performed concurrently, performance is impaired to the extent that their joint demands exceed available resources. Accordingly, making JOLs impairs recall for unrelated pairs (more so than related ones) because unrelated pairs are more difficult to learn.

The main goal of Experiment 2 was to evaluate competitive predictions made from these two hypotheses. The procedure was similar to Experiment 1 with two critical changes. First, the study phase was experimenter-paced for all participants. Second, and more important, whereas some participants studied a mixed list of related and unrelated word pairs (as in Experiment 1), we also included groups that studied pure lists comprised solely of related pairs or unrelated pairs. Both the changed-goal hypothesis and dual-mechanism account would predict an increased relatedness effect for the JOL group on the mixed list, but the two accounts differ in their predictions for the JOL groups on the pure lists. In particular, the changed-goal hypothesis predicts minimal JOL reactivity on pure lists, as changing one's goal is contingent on people noticing variation in item difficulty. By contrast, the dual-mechanism account predicts an increased relatedness effect such that positive reactivity will occur on the pure related list and negative reactivity will occur on the pure unrelated list.

Method

Design and participants A 2 (judgment group) × 3 (list type: mixed, related only, unrelated only) between-participant design was used. A power analysis was conducted using a medium effect size (based on the interaction for the experimenter-paced groups in Experiment 1), with the power set at .80 and alpha at .05. A total of 192 participants (n = 55, mixed list; n = 71, related pure list; n = 66, unrelated pure list) were recruited through Mechanical Turk and were compensated 50 cents each. Participation was restricted to the United States with an approval rating of 95% or higher.

Materials and procedure The mixed list consisted of the 60 cue–target word pairs used in Experiment 1. Two pure lists were created using only the 30 related pairs or the 30 unrelated pairs. The pairs were randomly ordered during study for each participant. The remainder of the procedure was identical to that of Experiment 1, except that all groups received experimenter-paced study.

Results

Similar to Experiment 1, our analytic plan was to analyze recall separately for the mixed-list groups and the pure-lists groups and to forgo an omnibus ANOVA given the unbalanced design. Nevertheless, we report the outcomes from the omnibus ANOVA (which yielded a significant three-way interaction) in the Appendix.

Mixed-list recall performance Mean recall for participants who received the mixed list are presented in Fig. 3. A 2 (judgment group) × 2 (relatedness) ANOVA revealed a significant main effect of relatedness, with recall being higher for related than unrelated pairs, $F(1, 53) = 221.56$, $p < .001$, $\eta_p^2 = .81$, and no main effect of judgment group, $F(1, 53) = .78$, $p = .38$. More important, a significant interaction occurred between relatedness and judgment group, $F(1, 53) = 15.36$, $p < .001$, $\eta_p^2 = .23$, such that the difference in recall between the related and unrelated pairs was larger for the JOL (d = 2.83) than no-JOL (d = 1.09) group. Planned comparisons revealed that recall for related pairs was significantly higher for the JOL than no-JOL group, $t(53) = 2.65$, $p = .01$, $d = .72$, and recall for unrelated pairs was lower for the JOL group, although this trend was not significant, $t(53) = .87$, $p = .39$, $d = .23$.

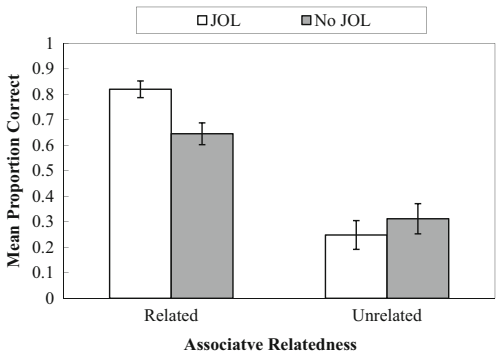

Fig. 3 Mean recall performance for related and unrelated word pairs for participants who received the mixed list in Experiment 2. Error bars represent the standard error of each mean

Pure-list recall performance Mean recall for participants who received the pure lists are presented in Fig. 4. A 2 (judgment group) × 2 (relatedness) ANOVA revealed a significant main effect of relatedness, with recall being higher for the related list than the unrelated list, $F(1, 133) = 124.48$, $p < .001$, $\eta_p^2 = .48$, and no main effect of judgment group, $F(1, 133) = .66$, $p = .42$. Critically, no interaction occurred between relatedness and judgment group, $F(1, 133) = .02$, $p = .88$, $\eta_p^2 = .23$.

Discussion

The main goal of Experiment 2 was to evaluate competitive predictions from the changed-goal hypothesis versus the dual-mechanism account. Consistent with both accounts, we replicated the increased relatedness effect on the mixed list. On the pure lists, the increased relatedness effect did not occur, which (as discussed above) provides more competitive support for the changed-goal hypothesis. Another goal of Experiment 2 was to estimate the relative contributions of positive and negative reactivity, and the increased relatedness effect obtained for the mixed list was driven primarily by positive reactivity.

Experiment 3

Outcomes from Experiments 1 and 2 converge on the conclusion that positive reactivity contributes more than negative reactivity to the increased relatedness effect. Given the consistent but nonsignificant trends toward negative reactivity, an issue arises as to whether it significantly contributes at all. Using the mixed-list design, we conducted a high-powered replication study to better estimate the relative contribution of negative reactivity. A power analysis estimated a sample of 300 participants to detect an effect size of .29 (based on estimates of negative reactivity from Experiments 1 and 2), with the power set at .80 and alpha at .05.

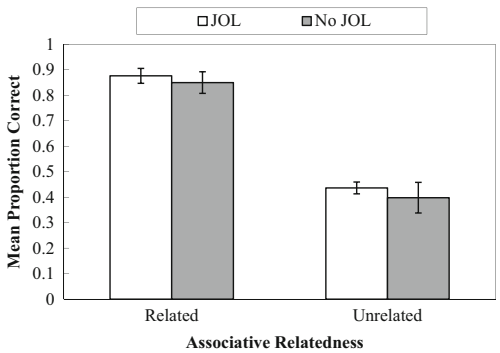

Fig. 4 Mean recall performance for related and unrelated word pairs for participants who received the pure lists in Experiment 2. Error bars represent the standard error of each mean.

Method

Design, participants, materials, and procedure A 2 (judgment group) × 2 (relatedness) mixed design was used, with judgment group manipulated between participants and relatedness manipulated within participant. Participants were recruited through Mechanical Turk and were compensated 50 cents. Restrictions on participation were the same as in Experiment 2. Data from 11 participants who participated in Experiment 2 were not analyzed, so the final sample consisted of 289 individuals (JOL group, $n = 147$; no-JOL group, $n = 142$). The procedure was identical to that of the mixed-list, experimenter-paced groups in Experiment 2.

Results and discussion

Mean recall is presented in Fig. 5. A 2 (judgment group) × 2 (relatedness) ANOVA revealed a significant main effect of relatedness, with recall being higher for related pairs than unrelated pairs, $F(1, 287) = 1077.86$, $p < .001$, $\eta_p^2 = .79$, and no main effect of judgment group, $F(1, 287) = .66$, $p = .42$. A significant interaction occurred between relatedness and judgment group, $F(1, 287) = 29.92$, $p < .001$, $\eta_p^2 = .09$, such that the difference in recall between the related and unrelated pairs was larger for the JOL group ($d = 2.12$) than for the no-JOL group ($d = 1.29$). Planned comparisons revealed that recall performance for related pairs was significantly higher for the JOL than for the no-JOL group, $t(287) = 3.35$, $p = .001$, $d = .39$. Although recall for unrelated pairs was lower for the JOL than no-JOL group, this trend was again not significant, $t(287) = 1.55$, $p = .12$, $d = .18$.

General discussion

What is the nature of JOL reactivity?

A main goal of the present research was to better reveal the nature of JOL reactivity in the context of the increased

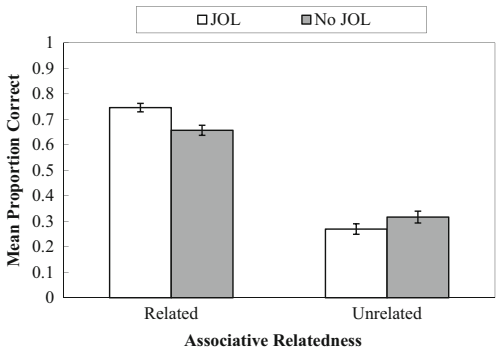

Fig. 5 Mean recall performance for related and unrelated word pairs in Experiment 3. Error bars represent the standard error of each mean

relatedness effect. In particular, given the conflicting reports about the source driving the effect in prior research, our goal was to investigate the extent to which the effect is driven by positive reactivity for related pairs or by negative reactivity for unrelated pairs. When the increased relatedness effect was obtained, positive reactivity consistently contributed to the effect, whereas negative reactivity did not. In fact, even when sufficient power was available to detect the relative contribution of negative reactivity (Experiment 3), its contribution was still not significant. Nevertheless, given that a consistent trend for negative reactivity occurred across experiments (see Figs. 1, 2, 3, 4, and 5), we conducted a continuously cumulating meta-analysis (CCMA) across our three experiments to further investigate the potential presence of negative reactivity (Braver, Thoemmes, & Rosenthal, 2014). Results indicated that the effect was small (d = .18) and approached conventional significance (p = .06), which confirms the conclusion that the increased relatedness effect is primarily driven by positive reactivity.

Consistent with these findings are those from a meta-analysis examining the reactive effects of immediate JOLs on memory. Double, Birney, and Walker (2017) reported that significant positive reactivity emerged for related pairs, whereas no reactivity emerged for unrelated pairs. The consistency with which positive reactivity is observed suggests that making JOLs is beneficial rather than detrimental to learning, although future research will need to identify the particular contexts in which negative reactivity might arise. Double et al. (2017) also reported positive reactivity for single-noun word lists, suggesting that JOLs may benefit memory in other learning contexts and with different to-be-learned materials.

What are the mechanism(s) of JOL reactivity?

Our data provide preliminary evidence that the increased relatedness effect arises from a single mechanism—for example, a changed learning goal—rather than from multiple mechanisms, as per the dual-mechanism account. According to the changed-goal hypothesis, JOLs direct participants to notice variation in item difficulty, which then causes them to focus more on learning the pairs they perceive to be easy (Mitchum et al., 2016). With the mixed lists (Experiments 1–3), this shifted study goal was marked by the larger difference in recall between related and unrelated pairs for the JOL than for the no-JOL group. With the pure lists (Experiment 2), however, the increased relatedness effect was eliminated, as predicted by the changed-goal hypothesis. By contrast, the dual-mechanism account predicted an increased relatedness effect even on the pure lists, as positive reactivity should have resulted from the strengthening of the cue–target pairs for the related list, and negative reactivity from dual-task costs on the unrelated list.

Although our data are consistent with the changed-goal hypothesis, Mitchum et al. (2016) argued that strong evidence for a changed learning goal would be (a) negative reactivity for unrelated pairs and (b) a weaker negative correlation between item difficulty and self-paced study time for the JOL than for the no-JOL group. Concerning the former, the minor contribution of negative reactivity in the present experiments suggests that changing one's goal to focus on related pairs can impair memory for unrelated pairs, although such an impairment makes a minor contribution to the overall increased relatedness effect (as per the CCMA reported above). Concerning the latter, we did not find the reduced correlation between study time and item difficulty for the JOL than for the no-JOL group, which may explain why the increased relatedness effect was so small for the self-paced group. Note, however, given that even large changes in study time can have little influence on subsequent memory performance (Nelson & Leonesio, 1988), any impact of changing one's goal on performance is likely not mediated by changes in study time. Instead, making JOLs may alter participants' goals that lead to changes in the strategies used to learn related and unrelated pairs. Accordingly, investigating the extent to which JOLs lead participants to strategically process items differently will be an important challenge for future research.

Finally, one possibility is that the changed-goal hypothesis can only account for outcomes on mixed lists consisting of related and unrelated word pairs. The relatedness manipulation may have made variations in item difficulty more obvious than other manipulations, and so further research will need to investigate JOL reactivity outside the context of the increased relatedness effect. Moreover, Witherby and Tauber (2017) recently found that participants who made JOLs on a pure list of related pairs demonstrated positive reactivity relative to participants who did not make JOLs—both immediately and after a 2-day delay. This outcome is inconsistent with the changed-goal hypothesis and suggests that JOLs can directly enhance learning, perhaps through some attentional mechanism that acts either globally or on an item-by-item basis. Similarly, Dougherty et al. (2005) noted that "one interesting hypothesis is that making a metacognitive judgment forces participants to process the to-be-remembered item more thoroughly than they would if no judgment was made. Thus, the act of making the judgment may affect how well the item is stored in memory" (p. 1110). An attentional account is not mutually exclusive with a changed learning goal, and we suspect that it would be premature (given that research on immediate JOL reactivity has just recently begun using appropriate research designs) to firmly conclude that any particular mechanism is absolutely responsible for reactivity effects. Instead, future research will need to investigate the extent to which making JOLs influences attention, processing, and/or how participants approach the learning task.

Practical implications and concluding remarks

Regardless of the mechanism(s) underlying JOL reactivity, outcomes from the present experiments offer an optimistic conclusion with respect to education. In particular, our findings suggest that instructing students to judge their learning during study could improve their retention of the materials. Whether a memorial benefit would persist in more authentic educational contexts (e.g., with more complex materials and for high-stakes exams) remains an open question, so discovering the particular contexts in which making JOLs boosts learning will be an important avenue for future research.

From a measurement perspective, we would argue that metamemory researchers should err on the side of caution when using JOLs by including a no-JOL group whenever possible. In fact, given that the reactive effects of metacognitive judgments as a whole are still poorly understood, including a no-judgment group regardless of the judgment being investigated will provide key outcomes regarding memory reactivity. In summary, we agree with Rhodes' (2016) commentary that the mere potential for reactivity "suggests an agenda for future research (a) to include appropriate control conditions to assess the impact of prediction on memory and (b) to provide a viable explanation of such reactivity" (p. 76). In the present experiments, doing so revealed positive reactivity for related word pairs and offered further evidence for the potential contribution of learning goals to JOL reactivity.

Appendix

For interested readers, we present results from the omnibus factorial ANOVA for both Experiments 1 and 2. Mean recall performance for related and unrelated pairs across these groups can be found in Tables 2 and 3, respectively.

For Experiment 1, a 2 (judgment group: JOL vs. no-JOL) × 2 (associative relatedness: related vs. unrelated) × 2 (pacing: experimenter-paced vs. self-paced) mixed ANOVA revealed a significant interaction between judgment group and associative relatedness, such that recall performance differed more between the JOL and no-JOL groups for related pairs than for unrelated pairs, $F(1, 137) = 15.44$, $p < .001$, $\eta_p^2 = .10$. Follow-up tests indicated that recall for related pairs was higher for the JOL groups than for the no-JOL groups, $t(139) = 3.12$, $p = .002$, $d = .47$, whereas recall for unrelated pairs did not differ significantly between judgment groups, $t(139) = 1.58$, $p = .12$, $d = .26$. Next, the interaction between pacing and associative relatedness was also significant, such that recall performance differed more between the experimenter-paced and self-paced groups for unrelated than for related pairs, $F(1, 137) = 5.00$, $p = .03$, $\eta_p^2 = .04$. Follow-up tests indicated that recall was higher for the experimenter-paced groups than for the self-paced groups for related pairs, $t(139) = 2.09$, $p = .04$, $d = .35$, and for unrelated pairs, $t(139) = 3.12$, $p = .002$, $d = .53$. Finally, the interaction between pacing and judgment condition was not significant, $F(1, 137) = .007$, $p = .94$, nor was the three-way interaction, $F(1, 137) = 2.31$, $p = .13$.

For Experiment 2, we conducted a 2 (judgment group: JOL vs. no-JOL) × 2 (associative relatedness: related vs. unrelated) × 2 (list composition: mixed vs. pure) ANOVA. Given that we manipulated associative relatedness both between and within participants, we could not treat it as a repeated measure (as we had in Experiment 1). Instead, we treated associative relatedness as a between-subjects variable, which may have reduced our power and resulted in a negatively biased F statistic (Erlebacher, 1977). Even with this conservative estimate, we still obtained a significant three-way interaction, $F(1, 239) = 3.97$, $p = .047$, $\eta_p^2 = .02$. Follow-up tests revealed a significant interaction between judgment group and associative relatedness for participants who received the mixed list, but not for participants who received the pure lists (refer to the Results section of Experiment 2).

Table 2 Mean proportion of word pairs recalled for all groups in Experiment 1

Group	Related	Unrelated
Experimenter-paced		
JOL	.82 (.02)	.23 (.03)
No-JOL	.72 (.03)	.32 (.04)
Self-paced		
JOL	.84 (.02)	.39 (.05)
No-JOL	.80 (.02)	.43 (.05)

Values in parentheses refer to the standard error of each mean

Table 3 Mean proportion of word pairs recalled for all groups in Experiment 2

Group	Related	Unrelated
Mixed list		
JOL	.82 (.03)	.25 (.04)
No-JOL	.65 (.06)	.31 (.06)
Pure lists		
JOL	.88 (.03)	.44 (.04)
No-JOL	.85 (.02)	.40 (.06)

Values in parentheses refer to the standard error of each mean

References

Braver, S. L., Thoemmes, F. J., & Rosenthal, R. (2014). Continuously cumulating meta-analysis and replicability. *Perspectives on Psychological Science*, *9*, 333–342.

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

Dougherty, M. R., Scheck, P., Nelson, T. O., & Narens, L. (2005). Using the past to predict the future. *Memory & Cognition, 33*, 1096–1115.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, *87*, 215–251.

Erlebacher, A. (1977). Design and analysis of experiments contrasting the within-and between subjects manipulation of the independent variable. *Psychological Bulletin, 84,* 212–219.

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

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

Koriat, A., Ma'ayan, H., & Nussinson, R. (2006). The intricate relationships between monitoring and control in metacognition: Lessons for the cause-and-effect relation between subjective experience and behavior. *Journal of Experimental Psychology: General*, *135*, 36–69.

Kornell, N., & Bjork, R. A. (2008). Optimising self-regulated study: The benefits—and costs—of dropping flashcards. *Memory, 16*, 125–136.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes, 25,* 259–284.

Metcalfe, J., & Kornell, N. (2003). The dynamics of learning and allocation of study time to a region of proximal learning. *Journal of Experimental Psychology: General, 132*, 530–542.

Mitchum, A. L., Kelley, C. M., & Fox, M. C. (2016). When asking the question changes the ultimate answer: Metamemory judgments change memory. *Journal of Experimental Psychology: General*, *145*, 200–219.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*, 402–407.

Nelson, T. O. & Leonesio, J. (1988). Allocation of self-paced study time and the "labor-in-vain effect" *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 676–686.

Rhodes, M. G. (2016). Judgments of learning: Methods, data, and theory. In J. Dunlosky & S. K. Tauber (Eds.), The Oxford Handbook of Metamemory (65 – 80). New York: 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*, 131–148.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366.

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.

Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 204–221.

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), 315–316.

Tauber, S. K., & Rhodes, M. G. (2012). Measuring memory monitoring with judgements of retention (JORs). *The Quarterly Journal of Experimental Psychology*, *65*(7), 1376–1396.

Thiede, K. W., & Dunlosky, J. (1999). Toward a general model of self-regulated study: An analysis of selection of items for study and self-paced study time. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 1024–1037.

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.

Zechmeister, E. B., & Shaughnessy, J. J. (1980). When you know that you know and when you think that you know but you don't. *Bulletin of the Psychonomic Society*, *15*, 41–44.