UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Undorf, M., Schäfer, F., & Halamish, V. (2024). Making Judgments of Learning Either Enhances or Impairs Memory: Evidence From 17 Experiments With Related and Unrelated Word Pairs. *Collabra: Psychology*, *10*(1). https://doi.org/10.1525/collabra.117108

Cognitive Psychology

Making Judgments of Learning Either Enhances or Impairs Memory: Evidence From 17 Experiments With Related and Unrelated Word Pairs

Monika Undorf[1][a], Franziska Schäfer[1], Vered Halamish[2]

[1] Department of Psychology, Technical University of Darmstadt, Darmstadt, Germany, [2] Faculty of Education, Bar-Ilan University, Ramat Gan, Israel

Keywords: metamemory, judgments of learning, reactivity

https://doi.org/10.1525/collabra.117108

Collabra: Psychology

Vol. 10, Issue 1, 2024

Published studies found that predicting one's future memory during learning (judgments of learning, JOLs) consistently improved cued-recall performance for related word pairs. In contrast, making JOLs had inconsistent effects on memory for unrelated pairs, with most studies finding null effects and some finding detrimental effects. This study reports data from 17 experiments in which participants either made or did not make JOLs for related and unrelated word pairs in their everyday language. Making JOLs increased the difference in memory performance between related and unrelated pairs in every experiment. Although almost all experiments showed numerically positive JOL reactivity for related pairs and numerically negative JOL reactivity for unrelated pairs, either effect was reliable in just half of the experiments. Small-scale meta-analyses revealed small-to-moderate positive reactivity for related pairs and small-to-moderate negative reactivity for unrelated pairs. Language of word pairs (German, Hebrew, or English) moderated positive reactivity and experimental setting (controlled or unsupervised online) moderated negative reactivity, but none of the other moderators we examined—presence of an additional pair type, study time, total number of pairs—impacted reactivity. Experiments that showed positive reactivity for related pairs tended not to show negative reactivity for unrelated pairs, and vice versa. Overall, these findings indicate that negative JOL reactivity for unrelated pairs is similarly large and robust as positive reactivity for related pairs. They favor the cue-strengthening hypothesis with dual-task costs over other accounts and raise the practically relevant possibility that monitoring could have detrimental effects on learning in educational settings. All data are freely available online.

Imagine students studying for a foreign language vocabulary test. Some students are instructed to closely self-monitor their learning by assessing their chances of remembering each studied vocabulary pair. Will these students outperform others who are not instructed to monitor their learning? Research on metamemory has examined this question in experiments in which some but not all participants made predictions of their future memory performance (judgments of learning, JOLs) while studying related and unrelated word pairs (e.g., *fish – salmon* vs. *fish – talent*) for a cued-recall test. Results have consistently shown that making JOLs improved memory for related pairs (positive JOL reactivity; e.g., Janes et al., 2018; Maxwell & Huff, 2024; Mitchum et al., 2016; Myers et al., 2020; Rivers et al., 2021; Soderstrom et al., 2015). In contrast, the effects of making JOLs on memory for unrelated pairs were inconsistent. The majority of studies found that making JOLs did not affect memory for unrelated pairs (Janes et al., 2018; Myers et al., 2020; Soderstrom et al., 2015), whereas some found that making JOLs impaired memory for unrelated pairs (negative JOL reactivity; Mitchum et al., 2016; Rivers et al., 2021). In a meta-analysis by Double et al. (2018), making JOLs improved cued-recall performance for related word pairs across 11 experiments, Hedges's g = .32, but did not affect cued-recall performance for unrelated word pairs, Hedges's g = -.01.

Published research thus suggests that positive JOL reactivity for related pairs is a robust phenomenon, whereas negative JOL reactivity for unrelated pairs is not. For instance, Rivers et al. (2021, Experiment 1, between-partic-

a Correspondence concerning this article should be addressed to Monika Undorf, Department of Psychology, Technical University of Darmstadt, Alexanderstraße 10, 64283 Darmstadt, Germany. Email: monika.undorf@tu-darmstadt.de

ipants manipulation of JOLs) reported a direct replication of an experiment by Soderstrom et al. (2015, Experiment 1b). The two experiments used the same word pairs, the same study times, the same retention intervals, and the same number of participants, but differed with respect to the setting (Soderstrom et al., 2015: unsupervised online experiment; Rivers et al., 2021: lab experiment), the sample population (Soderstrom et al., 2015: a diverse sample from Amazon's Mechanical Turk; Rivers et al., 2021: undergraduate students), and the distractor task (Soderstrom et al., 2015; Tetris, Rivers et al., 2021: paper-and-pencil arithmetic problems). Both experiments showed medium-sized positive reactivity for related pairs. In contrast, the effect of making JOLs on cued-recall performance for unrelated words differed substantially across the two experiments: Soderstrom et al. (2015) found nearly identical memory performance in participants who made JOLs and participants who did not make JOLs, whereas Rivers et al. (2021) found negative reactivity.

To our surprise, and contrary to the majority of published work, we have repeatedly found negative reactivity for unrelated pairs in experimental conditions that conformed to variations of the standard paradigm of JOL reactivity research introduced by Soderstrom et al. (2015). These experimental conditions were parts of experiments that addressed specific hypotheses about the mechanisms underlying JOL reactivity, the consequences of JOL reactivity, or individual differences in JOL reactivity. In this study, we report all published and unpublished experimental conditions that either author ran until January 2022 and that included groups of participants who were or were not instructed to make JOLs while studying related and unrelated word pairs at an experimenter-paced rate. We make these data available online as a public resource.

Reporting these data is important for several reasons. Determining whether judging one's own learning has detrimental effects or, alternatively, no effects on memory for unrelated word pairs has theoretical relevance. At a very general level, JOL reactivity challenges the common assumption that the monitoring processes captured in JOLs occur spontaneously and inevitably during learning (Halamish & Undorf, 2020; Mitchum et al., 2016), and robust negative reactivity for unrelated pairs would challenge this assumption specifically for this type of material. Consequently, one would have to suspect that soliciting JOLs alters learning and memory and inflates measures of JOL accuracy for unrelated word pairs (e.g., Bröder & Undorf, 2019; Halamish, 2018; Mitchum et al., 2016; Soderstrom et al., 2015). In contrast, evidence against JOL reactivity for unrelated pairs would be consistent with ubiquitous memory monitoring for unrelated pairs and would therefore alleviate the abovementioned concerns as far as this type of material is concerned.

As importantly, clarifying the effect of making JOLs on memory performance for unrelated pairs is informative for theories of JOL reactivity. Three major accounts have been proposed to explain JOL reactivity: the changed-goal hypothesis (Mitchum et al., 2016), the attentional reorienting account (Rivers et al., 2021; Tauber & Witherby, 2019; Zhao et al., 2021), and the cue-strengthening hypothesis (Soderstrom et al., 2015).

The changed-goal hypothesis (Mitchum et al., 2016) assumes that making JOLs draws attention to differences in item difficulty. Learners who make JOLs therefore tend to shift their goals from mastering all items towards focusing on the easy items and invest more study time and more effort in the easy items. This change in goals may produce positive reactivity for related pairs and, at the same time, negative reactivity for unrelated pairs, because people selectively focus on related pairs at the expense of unrelated pairs (Janes et al., 2018; Mitchum et al., 2016). The changed-goal hypothesis is therefore consistent with robust negative reactivity for unrelated pairs.

According to the attentional reorienting account (Rivers et al., 2021; Tauber & Witherby, 2019; Zhao et al., 2021), positive reactivity for related pairs is due to the JOL prompts reorienting participants' attention to the pairs at study. Thus, while the attention of participants in the no-JOL group wanes as the presentation of a study pair progresses, the JOL prompt refreshes the JOL group's attention and, consequently, increases this group's engagement with the pairs. By assuming that mind wandering is reduced for unrelated pairs to begin with, or that higher engagement does not improve memory for unrelated pairs, attentional reorienting can also account for null effects of making JOLs on memory for unrelated pairs. In contrast, robust negative reactivity for unrelated pairs is inconsistent with attentional reorienting. The current findings thus indicate that attentional reorienting alone cannot account for JOL reactivity for unrelated pairs.

The cue-strengthening hypothesis (Soderstrom et al., 2015) assumes that making JOLs strengthens the cues JOLs rely on (Koriat, 1997) and therefore improves performance on memory tests that are sensitive to these cues (Morris et al., 1977). When studying word pairs, the associative relationship between cue and target is a central cue for JOLs (Dunlosky & Matvey, 2001; Koriat, 1997) and processing the existing cue-target relationship in related pairs enhances cued-recall performance. In contrast, making JOLs does not improve cued-recall performance for unrelated pairs, because these pairs do not have an existing cue-target relationship. Consequently, cue strengthening alone predicts that making JOLs does not affect memory for unrelated pairs. Robust negative reactivity for unrelated pairs thus indicates that the cue-strengthening hypothesis needs to be supplemented with dual-task costs, that is, the assumption that making JOLs while studying interferes with learning, and particularly so when the learning task is demanding, as has been suggested by Janes et al. (2018; also see Mitchum et al., 2016).

Taken together, robust negative JOL reactivity for unrelated pairs would inform accounts of JOL reactivity, because it is consistent with the changed-goal hypothesis (Mitchum et al., 2016), inconsistent with the attentional reorienting account (Rivers et al., 2021; Tauber & Witherby, 2019; Zhao et al., 2021), and indicates that the cue-strengthening hypothesis (Soderstrom et al., 2015) needs to be supple-

mented with dual-task costs (Janes et al., 2018; Mitchum et al., 2016).

Potentially negative effects of making JOLs on memory for unrelated word pairs are also of practical relevance. When considering educational implications of JOL reactivity, researchers have focused on the possibility that eliciting JOLs enhances learning and academic achievement (Double & Birney, 2019b; Janes et al., 2018; Tekin & Roediger, 2020; Witherby & Tauber, 2017). Negative JOL reactivity with unrelated pairs, however, would raise the possibility that making JOLs could have detrimental effects for learning educationally relevant material (but see Ariel et al., 2021; Schäfer & Undorf, 2024).

Another important reason for reporting the current data is that they can help to examine potential boundary conditions of negative JOL reactivity for unrelated word pairs. An obvious question is: Is negative JOL reactivity due to peculiarities in the materials or the design of specific experiments? As importantly, do study characteristics that increase dual-task costs, a common explanation for negative JOL reactivity (e.g., Janes et al., 2018; Mitchum et al., 2016), foster negative JOL reactivity for unrelated pairs?

To address these questions, we considered five potential moderators of JOL reactivity for unrelated pairs. First, almost all published studies on JOL reactivity involved participants from the US who studied and recalled English word pairs. While this is also true for some of our experiments, more than half of our experiments involved participants from Germany who studied and recalled German word pairs or, alternatively, participants from Israel who studied and recalled Hebrew word pairs. It was therefore possible that JOL reactivity for unrelated pairs varies across languages, with negative reactivity being more pronounced with or even limited to German- or Hebrew-speaking participants. Second, while most published studies included only related and unrelated word pairs, some of our experiments additionally comprised a third pair type (e.g., identical pairs such as *fish – fish* or related pairs printed in an alternating font format such as *fIsH – sAlMoN*). This raised the possibility that the presence of another pair type produced or increased negative JOL reactivity for unrelated pairs (also see Mitchum et al., 2016). Third, while most published studies used study times of 8 s per word pair, some of our experiments used shorter study times. It is plausible that shorter study times increase dual-task costs and, consequently, increase negative reactivity for unrelated pairs and decrease positive reactivity for related pairs. We therefore examined whether JOL reactivity varied between experiments with study times of 8 s and experiments with shorter study times. Fourth, based on the assumption that participants might be less involved or less motivated to do well in unsupervised online experiments than in more tightly controlled experiments (e.g., Clifford & Jerit, 2014; Gould et al., 2015; Peer et al., 2021), we reasoned that dual-task costs might be more pronounced in the former than in the latter. We therefore tested whether negative reactivity varied between unsupervised online experiments and tightly controlled experiments. Fifth, we examined whether variations in the length of study lists across our experiments are related to negative reactivity. This was based on the assumption that long study lists with many to-be-studied word pairs might increase dual-task costs and, consequently, increase negative reactivity. In the small-scale meta-analyses reported below, we therefore tested whether language, the presence of an additional item type, experimental setting, study time, and the total number of to-be-studied word pairs moderated JOL reactivity.

Moreover, the current data enable us to examine the relationship between (a) the effect of making JOLs on memory for related word pairs and (b) the effect of making JOLs on memory for unrelated word pairs, which may inform theories of JOL reactivity. Importantly, the JOL reactivity accounts that are compatible with robust negative reactivity make different predictions for this relationship. The changed-goal hypothesis (Mitchum et al., 2016) predicts a positive association between positive and negative reactivity across experiments. The reason for this is that particularly strong strategy shifts towards related pairs should strongly impair memory for unrelated pairs and strongly improve memory for related pairs. Consequently, experiments that yield strong positive JOL reactivity for related pairs should yield strong negative JOL reactivity for unrelated pairs. In contrast, the cue-strengthening hypothesis with dual-task costs (Janes et al., 2018; Soderstrom et al., 2015) predicts inverse relations between positive and negative reactivity across experiments. The reason for this is that particularly high dual-task costs should strongly reduce the benefits of cue-strengthening for related pairs and strongly impair memory for unrelated pairs. Consequently, experiments that yield strong positive JOL reactivity for related pairs should yield little or no negative JOL reactivity for unrelated pairs and experiments that yield strong negative reactivity for unrelated pairs should yield reduced or no positive reactivity for related pairs.

Finally, other researchers might use the data presented here for addressing further research questions.

The Current Study

This study reports 17 separate experiments in which participants who were instructed to make JOLs (JOL group) and participants who were not instructed to make JOLs (no-JOL group) studied related and unrelated word pairs for a cued-recall test. All participants studied pairs in their everyday language, which was German (Experiments 1 to 6), Hebrew (Experiments 7 to 10), or English (Experiments 11 to 17). These 17 experiments are all published and unpublished data sets either of the authors ran until January 2022 and that included groups of participants who were or were not instructed to make JOLs while studying related and unrelated word pairs at an experimenter-paced rate.

In addition to analyzing each individual experiment separately, we report a series of small-scale meta-analyses that synthesize data from the three groups of experiments and examine the relationship between reactivity for unrelated and related pairs across experiments. This approach allowed us to (1) evaluate whether making JOLs has a detrimental effect or, alternatively, no effect on memory for unrelated word pairs, (2) compare the robustness and the size

of positive JOL reactivity for related pairs to that of negative JOL reactivity for unrelated pairs, (3) examine potential moderators of JOL reactivity, and (4) explore the relationship between JOL reactivity for related and unrelated pairs across experiments.

Experiments 1 to 6

Participants in Experiments 1 to 6 came from the University of Mannheim community and studied and recalled word pairs in their everyday language German.

Method

Design

Each experiment conformed to the standard paradigm of JOL reactivity research introduced by Soderstrom et al. (2015) and manipulated judgment group and pair type. Judgment group was manipulated between subjects: We solicited JOLs in the JOL group but not in the no-JOL group. Word pair type was manipulated within subjects: Each participant studied and recalled related and unrelated word pairs. In some experiments, the original design included an additional pair type (e.g., backward related pairs) or additional conditions (e.g., conditions in which participants self-paced their study) that were omitted from the current analyses. We omitted these pair types and conditions because they deviated from the standard paradigm of JOL reactivity research and were irrelevant for the current investigation. For completeness, we report results for all omitted pair types in Appendix A. We do not report any data from tasks that took place after the cued-recall test (e.g., personality tests, a second study-test cycle).

Participants

Samples consisted of participants from the University of Mannheim community who received course credit, monetary compensation, or were entered into a draw for gift cards in exchange for their participation.

Table 1 presents the composition of the sample and the number of participants in each experiment. Samples sizes ranged from 64 (Experiments 1 and 3) to 168 (Experiment 6), M = 101.83, SD = 42.93. Sensitivity power analyses showed that, on average, sample sizes were adequate to detect medium-sized effects (Cohen's d: M = 0.60, SD = 0.13; range: 0.43 – 0.71) in two-tailed t tests for independent samples with a statistical power of $(1 - \beta) = .80$ and $\alpha = .05$ (all power analyses conducted via G*Power 3; Faul et al., 2007). In mixed ANOVAs, small-to-medium-sized interactions (Cohen's f: M = 0.15, SD = 0.03, range: .11 – .18) between judgment group and word pair type could be detected with $(1 - \beta) = .80$, $\alpha = .05$, and a correlation of .50 between repeated measures.

In these and all subsequent experiments, we excluded participants who copied words at study according to experimenter reports in supervised experiments and according to reports of the participants themselves in unsupervised online experiments. The number of excluded participants per experiment and condition is reported in Table 1.

Materials

Materials were German word pairs. Each experiment included equal numbers of related and unrelated pairs. Associative strengths for related pairs were taken from Melinger and Weber (2006) and ranged from .04 to .98. All unrelated pairs had an associative strength of zero. Table 1 presents the exact number of pairs and the associative strength of related pairs for each experiment. Materials in Experiment 2 also comprised a third pair type (see Table 1) that was excluded from the reported analyses (see Appendix A for detailed information and results).

Procedure

Experiments 1 to 3 took place in the lab and Experiments 4 to 6 were unsupervised online experiments (see Table 1).

Unless otherwise specified, all experiments used the following procedure. Participants were asked to study word pairs for a later memory test. At study, each pair appeared on a computer screen for a fixed amount of time. Exact study times for each experiment can be found in Table 1. Study times were identical for all participants in each experiment. In the JOL group of all experiments, each pair was presented on the screen for a fixed portion of this time before the JOL prompt (*Chance to recall?*) was added to the screen (see Table 1 for exact times). Participants had to make their JOL on a 0%-100% scale in the remaining portion of the time. The programs advanced to the next pair when the study time was up regardless of whether participants had provided a JOL or not.[1] Participants in the no-JOL group were not prompted to make JOLs. After the study phase, participants completed a distractor task in which they typed in words or solved a Sudoku for a couple of minutes (see Table 1 for details). After each experiment's retention interval, participants took a self-paced cued-recall test in which cue words were presented one-by-one and participants typed in targets. In all experiments and for each participant, items were presented in new random orders at study and test.

Experiments 1 and 3 deviated from the standard procedure in that study time varied across participants but was identical across the JOL and no-JOL groups. In these experiments, the study time of one participant from the JOL

1 We repeated all analyses excluding participants from the JOL groups who failed to provide a JOL in 10% or more of the trials. This slightly changed the number of experiments with significant reactivity for related pairs (German materials: 1 instead of 0 experiments) and for unrelated pairs (German materials: 3 instead of 4 experiments; English materials: 1 instead of 3 experiments) but did not change any other results. Notably, since this procedure involves omitting trials only from the JOL group, it might bias the results and should be considered with caution (see Halamish & Undorf, 2023).

Table 1. Effect of Making JOLs on Recall Performance

Exp.	Setting	Sample population	Language	Retention interval	No. items	Mean association related pairs (range)	Condition	Study time	No. participants (excluded)	Related pairs		Unrelated pairs		Judgment X Item type interaction
										Mean recall (*SD*)	Effect of judgment	Mean recall (*SD*)	Effect of judgment	
1	Lab	UoM community	German	3 min	30 related, 30 unrelated	.56 (.41-.74)	no JOL	7.88s[a]	32	.76 (.17)	$t(62) = 1.21$, $p = .229$, $d = 0.30$	.33 (.23)	$t(62) = -1.72$, $p = .091$, $d = -0.43$	$F(1, 62) = 10.87$, $p = .002$, $\eta_p^2 = .15$
							JOL	3.88s + 4s[a]	32	.82 (.20)		.24 (.21)		
2	Lab	UoM community	German	10 min	20 related, 20 unrelated, 20 backward related	.20 (.04-.53)	no JOL	6s	36	.72 (.21)	$t(70) = 1.10$, $p = .276$, $d = 0.26$	.47 (.27)	$t(70) = -3.57$, $p < .001$, $d = -0.84$	$F(1, 70) = 35.61$, $p < .001$, $\eta_p^2 = .34$
							JOL	3s + 3s	36	.77 (.18)		.26 (.22)		
3	Lab	UoM community	German	24 hr	30 related, 30 unrelated	.56 (.41-.74)	no JOL	10.77s[a]	32	.67 (.19)	$t(62) = 1.28$, $p = .204$, $d = 0.32$	.12 (.13)	$t(62) = -2.68$, $p = .010$, $d = -0.67$	$F(1, 62) = 13.77$, $p < .001$, $\eta_p^2 = .18$
							JOL	6.77s + 4s[a]	32	.72 (.15)		.05 (.08)		
4	Unsuperv. online	UoM community	German	3 min	30 related, 30 unrelated	.69 (.54-.92)	no JOL	8s	67 (3)	.81 (.12)	$t(128) = 0.35$, $p = .726$, $d = 0.06$	.52 (.27)	$t(128) = -3.78$, $p < .001$, $d = -0.67$	$F(1, 128) = 21.88$, $p < .001$, $\eta_p^2 = .15$
							JOL	4s + 4s	67 (1)	.82 (.10)		.35 (.25)		
5	Unsuperv. online	UoM community	German	3 min	30 related, 30 unrelated	.52 (.22-.92)	no JOL	8s	54 (2)	.74 (.25)	$t(96) = 0.74$, $p = .462$, $d = 0.15$	.42 (.26)	$t(96) = -1.48$, $p = .143$, $d = -0.30$	$F(1, 96) = 7.41$, $p = .008$, $\eta_p^2 = .07$
							JOL	4s + 4s	55 (9)	.77 (.21)		.34 (.24)		
6	Unsuperv. online	UoM community	German	3 min	30 related, 30 unrelated	.63 (.44-.98)	no JOL	8s	84 (4)	.79 (.17)	$t(161) = 1.36$, $p = .175$, $d = 0.21$	.51 (.29)	$t(161) = -2.29$, $p = .023$, $d = -0.36$	$F(1, 161) = 20.09$, $p < .001$, $\eta_p^2 = .11$
							JOL	4s + 4s	84 (1)	.83 (.21)		.41 (.25)		
7	Lab	BIU students	Hebrew	5 min	20 related, 20 unrelated	.33 (.30-.40)	no JOL	8s	25	.82 (.16)	$t(43) = 0.78$, $p = .438$, $d = 0.23$	.52 (.33)	$t(43) = -1.28$, $p = .207$, $d = -0.38$	$F(1, 43) = 5.87$, $p = .020$, $\eta_p^2 = .12$
							JOL	4s + 4s	20	.86 (.16)		.42 (.21)		
8	Lab/ superv. online	BIU students	Hebrew	5 min	30 related, 30 unrelated	.32 (.30-.36)	no JOL	8s	34	.76 (.18)	$t(67) = -0.35$, $p = .726$, $d = -0.09$	.40 (.29)	$t(67) = -3.55$, $p = .001$, $d = -0.86$	$F(1, 67) = 15.72$, $p < .001$, $\eta_p^2 = .19$
							JOL	4s + 4s	35	.75 (.20)		.20 (.16)		
9	Lab/ superv. online	BIU students	Hebrew	5 min	30 related, 30 unrelated	.32 (.30-.36)	no JOL	8s	35 (1)	.75 (.19)	$t(65) = 1.55$, $p = .125$, $d = 0.38$	.39 (.28)	$t(65) = -1.48$, $p = .144$, $d = -0.36$	$F(1, 65) = 9.16$, $p = .004$, $\eta_p^2 = .12$
							JOL	4s + 4s	33	.81 (.12)		.30 (.19)		
10	Unsuperv. online	Midgam panel	Hebrew	5 min	20 related, 20 unrelated	.33 (.30-.40)	no JOL	8s	66 (6)	.71 (.23)	$t(122) = 0.71$, $p = .480$, $d = 0.13$	.32 (.29)	$t(122) = -1.21$, $p = .230$, $d = -0.22$	$F(1, 122) = 5.17$, $p = .025$, $\eta_p^2 = .04$
							JOL	4s + 4s	65 (1)	.73 (.23)		.26 (.25)		
11[b]	Unsuperv. online	Students on Prolific	English	3 min	20 related, 20 unrelated, 20 identical	.52 (.34-.75)	no JOL	8s	65 (4)	.62 (.26)	$t(126) = 2.43$, $p = .016$, $d = 0.43$	.26 (.30)	$t(126) = -1.95$, $p = .053$, $d = -0.35$	$F(1, 126) = 27.15$, $p < .001$, $\eta_p^2 = .18$
							JOL	4s + 4s	67	.72 (.21)		.17 (.20)		
12	Unsuperv. online	Students on Prolific	English	3 min	18 related, 18 unrelated	.54 (.41-.75)	no JOL	6s	47	.71 (.22)	$t(97) = 2.27$, $p = .025$, $d = 0.46$	.27 (.27)	$t(97) = -0.71$, $p = .480$, $d = -0.14$	$F(1, 97) = 8.94$, $p = .004$, $\eta_p^2 = .08$
							JOL	2s + 4s	52	.80 (.15)		.23 (.22)		
13	Unsuperv. online	Students on Prolific	English	9 min	18 related, 18 unrelated	.54 (.41-.75)	no JOL	6s	83 (2)	.69 (.22)	$t(155) = 4.49$, $p < .001$, $d = 0.72$	.24 (.22)	$t(155) = -0.54$, $p = .589$, $d = -0.09$	$F(1, 155) = 26.12$, $p < .001$, $\eta_p^2 = .14$
							JOL	2s + 4s	80 (4)	.83 (.16)		.22 (.20)		
14	Unsuperv. online	Students on Prolific	English	3 min	20 related, 20 unrelated, 20 related aLtErNaTiNg font	.52 (.34-.75)	no JOL	8s	64	.69 (.23)	$t(127) = 2.05$, $p = .042$, $d = 0.36$	.27 (.23)	$t(127) = -3.14$, $p = .002$, $d = -0.55$	$F(1, 127) = 33.91$, $p < .001$, $\eta_p^2 = .21$
							JOL	4s + 4s	65	.76 (.17)		.16 (.17)		
15	Unsuperv. online	Students on Prolific	English	3 min	30 related, 30 unrelated	.52 (.34-.75)	no JOL	8s	51 (3)	.75 (.20)	$t(94) = 0.62$, $p = .536$, $d = 0.13$	.33 (.29)	$t(94) = -2.10$, $p = .039$, $d = -0.43$	$F(1, 94) = 11.53$, $p = .001$, $\eta_p^2 = .11$
							JOL	4s + 4s	49 (1)	.77 (.16)		.23 (.20)		
16[b]	Unsuperv. online	Students on Prolific	English	3 min	20 related, 20 unrelated, 20 identical	.52 (.34-.75)	no JOL	8s	68 (2)	.66 (.23)	$t(129) = 2.58$, $p = .011$, $d = 0.45$	.28 (.25)	$t(129) = -2.33$, $p = .021$, $d = -0.41$	$F(1, 129) = 29.56$, $p < .001$, $\eta_p^2 = .19$
							JOL	4s + 4s	65	.75 (.18)		.19 (.20)		

17[b]	Unsuperv. online	Students on Prolific	English	3 min	20 related, 20 unrelated, 20 identical	.52 (.34-.75)	no JOL JOL	8s 4s + 4s	136 (6) 135 (3)	.66 (.22) .79 (.16)	$t(260) = 5.67$, $p < .001$, $d = 0.70$	.24 (.23) .25 (.23)	$t(260) = 0.42$, $p = .673$, $d = 0.05$	$F(1, 260) = 26.97$, $p < .001$, $\eta_p^2 = .09$

Note. BIU = Bar-Ilan University; superv. = supervised; unsuperv. = unsupervised; UoM = University of Mannheim.

[a] Mean study times across all participants (also see Footnote 3).

[b] Published in Halamish and Undorf (2023).

Table 2. JOLs by Word Pair Relatedness

Experiment	Mean JOL (*SD*)		Effect of relatedness on JOLs			
	Related	Unrelated	*t*	*df*	*p*	*d*
1	67.20 (20.38)	23.12 (14.61)	9.94	62	< .001	2.40
2	56.09 (15.83)	20.20 (13.94)	10.21	70	< .001	2.40
3	68.56 (10.83)	29.74 (15.01)	11.87	62	< .001	2.99
4[a]	74.00 (16.29)	28.91 (16.09)	15.66	126	< .001	2.79
5[b]	67.88 (19.50)	34.35 (16.91)	8.32	80	< .001	1.83
6	72.21 (17.55)	31.24 (14.26)	15.81	158	< .001	2.50
7	81.16 (13.10)	35.57 (17.03)	9.49	38	< .001	2.94
8	77.15 (13.78)	25.56 (14.43)	15.60	68	< .001	3.73
9	78.60 (18.21)	37.91 (22.17)	8.15	64	< .001	1.98
10[a]	74.82 (17.86)	31.40 (18.77)	13.09	120	< .001	2.37
11[c]	58.15 (19.68)	24.03 (16.27)	10.89	131	< .001	1.95
12[c]	69.42 (12.84)	22.06 (10.86)	10.89	100	< .001	3.97
13[d]	73.45 (16.77)	28.39 (13.03)	18.25	146	< .001	2.98
14	67.09 (16.44)	26.56 (15.72)	14.37	128	< .001	2.52
15[d]	65.07 (13.75)	29.13 (17.44)	9.64	92	< .001	2.38
16[c]	66.18 (14.65)	32.10 (16.98)	12.16	126	< .001	2.14
17[d]	69.09 (14.98)	30.89 (18.53)	18.29	259	< .001	2.25

Note. [a] 3 participants were excluded due to missing values. [b] 4 participants were excluded due to missing values. [c] 1 participant was excluded due to missing values. [d] 2 participants were excluded due to missing values.

group and one participant from the no-JOL group was identical to the mean study time of a participant from a self-paced study condition not reported here. Mean study time was 7.88 s (*SD* = 1.16, 6.50 – 10.03) in Experiment 1 and 10.77 s (*SD* = 1.92, 8.60 – 16.07) in Experiment 3. Also, in Experiment 3, participants took the test one day after the study phase (*M* = 22.84 hr, *SD* = 1.58, 19.72 – 26.53).

Analysis code and all data are available at https://osf.io/q3jgm.

Results

Mean JOLs (see Table 2) were higher for related than for unrelated pairs in all experiments, all $t >= 8.32$, all $p < .001$. Table 1 shows cued-recall performance for related and unrelated pairs in the JOL and no-JOL groups. As can be seen in the table, each of the six experiments revealed a significant Judgment group X Item type interaction in a 2 (Item type) X 2 (Judgment group) mixed ANOVA, all $F >= 7.41$, all $p <= .008$, indicating that the effect of making JOLs on cued-recall performance differed between related and unrelated pairs. Recall for related pairs was numerically better in the JOL group than in the no-JOL group in all experiments, but not reliably so, all $t <= 1.36$, all $p >= .175$. Recall for unrelated pairs was numerically worse in the JOL group than in the no-JOL group in all experiments, which was reliable in four experiments (66.67%).

Discussion

Making JOLs increased the effect of word pair relatedness on cued-recall performance in each of experiment. All six experiments showed numerically positive JOL reactivity for related pairs and numerically negative JOL reactivity for unrelated pairs. JOL reactivity for related pairs was reliable in neither experiment, whereas JOL reactivity for unrelated pairs was reliable in two thirds of the experiments.

Experiments 7 to 10

Participants in Experiments 7 to 10 were people from Israel who studied and recalled word pairs in their everyday language Hebrew.

Method

Design

The design was identical to that in Experiments 1 to 6.

Participants

Samples consisted of students from Bar-Ilan University (Experiments 7 to 9) and people active on the Israeli Midgam project web panel (Experiment 10, https://www.midgampanel.com). All participants were pre-screened to be 18-30 years old with Hebrew as a first language. Participants received course credit or monetary compensation for their participation. In Experiment 9, participants additionally received a performance-based monetary bonus for each correctly remembered target (0.25 ILS, about 0.07 USD).

Samples sizes ranged from 45 (Experiment 7) to 131 (Experiment 10), *M* = 78.25, *SD* = 36.87. Sensitivity power analyses showed that, on average, sample sizes were ade-

quate to detect medium-sized effects (Cohen's *d*: *M* = 0.68, *SD* = 0.15; range: 0.49 – 0.86) in two-tailed *t* tests for independent samples with a statistical power of (1 - β) = .80 and α = .05. In mixed ANOVAs, small-to-medium-sized interactions (Cohen's *f*: *M* = 0.17, *SD* = 0.04, range: .12 – .21) between judgment group and word pair type could be detected with (1 - β) = .80, α = .05, and a correlation of .50 between repeated measures. Table 1 presents the composition of the sample and the number of participants in each experiment.

Materials

Materials were Hebrew word pairs. Associative strengths for related pairs were taken from Henik et al. (2005) and ranged from .30 to .40. All unrelated pairs had an associative strength of zero. Table 1 presents the exact number of pairs and the associative strength of related pairs for each experiment.

Procedure

Experiments 7 to 9 took place in the lab or while participants were supervised by an experimenter using the videoconferencing software Zoom (Zoom Video Communications Inc., 2016). Experiment 10 was an unsupervised online experiment. All experiments followed the standard procedures described for Experiments 1 to 6. Analysis code and all data are available at https://osf.io/q3jgm.

Results

Mean JOLs (see Table 2) were higher for related than for unrelated pairs in all experiments, all $t >= 8.15$, all $p < .001$. As can be seen in Table 1, all experiments revealed a significant Judgment X Item type interaction, indicating that the effect of making JOLs on cued-recall performance differed between related and unrelated pairs, all $F >= 5.17$, all $p <= .025$. Recall for related pairs was numerically better in the JOL group than in the no-JOL group in three of the four experiments, but not reliably so, all $t <= 1.55$, all $p >= .125$. Recall for unrelated pairs was numerically worse in the JOL group than in the no-JOL group in all experiments, with *t* tests for independent samples revealing significant differences in one experiment (25.00%).

Discussion

As in the experiments with German samples, making JOLs always increased the effect of word pair relatedness on cued-recall performance and consistently produced numerically positive JOL reactivity for related pairs and numerically negative JOL reactivity for unrelated pairs. Also as in the experiments with German samples, JOL reactivity for related pairs was unreliable throughout, whereas JOL reactivity for unrelated pairs was reliable in one fourth of the experiments.

Experiments 11 to 17

Participants in Experiments 11 to 17 were students from the UK and the USA who studied and recalled word pairs in their everyday language English.

Method

Design

The design was identical to that in Experiments 1 to 10.

Participants

Samples consisted of people active on Prolific (http://www.prolific.ac) who were prescreened to be 18-30-year-old students from the UK or the USA with English as a first language and prior approval rates of 95% or higher.[2] All participants received monetary compensation in exchange for their participation.

Samples sizes ranged from 99 (Experiment 12) to 264 (Experiment 17), *M* = 145.43, *SD* = 55.86 (see Table 1 for the number of participants in each experiment). Sensitivity power analyses showed that, on average, sample sizes were adequate to detect small-to-medium-sized effects (Cohen's *d*: *M* = 0.49, *SD* = 0.08; range: 0.35 – 0.57) in two-tailed *t* tests for independent samples with a statistical power of (1 – β) = .80 and α = .05 (all power analyses conducted via G*Power 3; Faul et al., 2007). In mixed ANOVAs, small-to-medium-sized interactions (Cohen's *f*: *M* = 0.12, *SD* = 0.02, range: .09 – .14) between judgment group and word pair type could be detected with (1 – β) = .80, α = .05, and a correlation of .50 between repeated measures.

Materials

Materials were English word pairs. Associative strengths for related pairs were taken from Nelson et al. (1998) and ranged from .34 to .75. All unrelated pairs had an associative strength of zero. Some experiments also included a third pair type (see Table 1) that was excluded from the reported analyses (see Appendix A for detailed information and results). Table 1 presents the exact number of pairs and the associative strength of related pairs for each experiment.

Procedure

All experiments were unsupervised online experiments. Procedures were identical to the standard procedures of Experiments 1 to 10 with the following exceptions. In Experiment 16, participants indicated whether each cue word was part of an identical pair, a related pair, or an unrelated pair

2 Experiments 11-17 were all unsupervised online experiments, because neither author is based in an English-speaking country, which is why we could not run lab-based experiments with participants whose everyday language is English.

by clicking one of three labeled buttons (self-paced) before they tried to type in the target at test. In Experiment 17, half of the participants from the JOL and the no-JOL group made such relatedness judgments at test. We collapsed across participants who did and did not make relatedness judgments at test in Experiment 17, because neither the main effect of relatedness judgment group nor any interactions involving relatedness judgment group were significant (for more details, see Halamish & Undorf, 2023). Analysis code and all data are available at https://osf.io/q3jgm.

Results

Mean JOLs (see Table 2) were higher for related than for unrelated pairs in all experiments, all $t >= 9.64$, all $p < .001$. In all experiments, significant Judgment X Item type interactions indicated that the effect of making JOLs on cued-recall performance differed between related and unrelated pairs, all $F >= 8.94$, all $p <= .004$. Recall for related pairs was numerically better in the JOL group than in the no-JOL group in all experiments, with t tests revealing significant differences in all but one experiment (85.71%). Recall for unrelated pairs was numerically worse in the JOL group than in the no-JOL group in all but one experiment, and t tests revealed significant differences in three experiments (42.86%). In the remaining experiment (Experiment 17), recall for unrelated pairs was numerically but not reliably better in the JOL group than in the no-JOL group, $t < 1$.

Discussion

As in the experiments with German- and Hebrew-speaking participants, making JOLs always increased the effect of word pair relatedness on cued-recall performance. Also as in the previous experiments, making JOLs consistently produced numerically positive JOL reactivity for related pairs and numerically negative JOL reactivity for unrelated pairs with the exception of Experiment 17, which showed numerically positive reactivity for unrelated pairs ($t < 1$). Unlike in the experiments with German- and Hebrew-speaking participants, JOL reactivity for related pairs was reliable in almost all experiments. In contrast, the finding that JOL reactivity for unrelated pairs was reliable in about half of the experiments was similar to the experiments with German- and Hebrew-speaking participants.

Small-scale meta-analyses

To synthesize the data obtained from German-, Hebrew-, and English-speaking samples, and to test for potential moderators of JOL reactivity, we performed a series of small-scale meta-analyses and moderator analyses.

All meta-analyses were conducted with the *meta* package in *R* (Balduzzi et al., 2019) using Hedges's g as the measure of effect size. Positive effect sizes indicate a benefit for the JOL group over the no-JOL group (positive JOL reactivity) and negative effect sizes indicate a disadvantage for the JOL group relative to the no-JOL group (negative JOL reactivity). Because a prior meta-analysis (Double et al., 2018) found differences in JOL reactivity between related and unrelated word pairs, we conducted separate meta-analyses for related and unrelated word pairs. For each pair type, we first pooled the effect sizes of all experiments with a random-effects model and using the restricted maximum likelihood estimator to calculate the heterogeneity variance τ^2 (Viechtbauer, 2005). We then examined between-experiment heterogeneity and conducted univariate meta-regressions to assess the role of the potential moderators language, additional pair type, experimental setting, study time, and number of word pairs.

Related pairs. The meta-analysis revealed small-to-moderate positive reactivity, Hedges's g =.33, 95% CI = [.21, .45], $p < .001$. As can be seen in Figure 1, recall performance was better in the JOL groups than in the no-JOL groups. Heterogeneity between experiments was nonsignificant, $Q(16) = 23.93$, $p = .091$. As can be seen in Table 2, the only variable that significantly moderated positive reactivity was the language of word pairs, $Q(2) = 12.58$, $p = .002$. Positive reactivity was larger for English word pairs ($g = 0.51$) than for German word pairs ($g = 0.19$) or Hebrew word pairs ($g = 0.15$) and was even nonsignificant in the last group of experiments. This might, however, be related to the small number of experiments with Hebrew word pairs (4 experiments). In contrast, neither the presence of an additional pair type, nor experimental setting, study time, or the total number of word pairs had an impact on positive reactivity, all $Q <= 1.92$, all $p >= .166$. Positive reactivity was of low-to-moderate size in every subgroup of experiments and reliable for all subgroups except for the tightly controlled experiments, all of which involved German or Hebrew word pairs.

Unrelated pairs. The meta-analysis revealed small-to-moderate negative reactivity, Hedges's g = -.37, 95% CI = [-.50, -.25], $p < .001$. As can be seen in Figure 1, recall performance was worse in the JOL groups than in the no-JOL groups. Heterogeneity between experiments was significant, $Q(16) = 28.64$, $p = .027$. A moderator analysis (see Table 3) revealed that the experimental setting significantly moderated negative reactivity, $Q(1) = 4.77$, $p = .029$. Contrary to expectations, negative reactivity was larger in tightly controlled experiments ($g = -0.59$) than in unsupervised online experiments ($g = -0.29$), but significant in both experimental settings. In contrast, neither of the other potential moderators had an impact on negative reactivity, all $Q <= 2.93$, all $p >= .087$. Negative reactivity was of low-to-moderate size in every subgroup of experiments and reliable for all subgroups except for the very few experiments with study times below 8 s (3 experiments).

Relationship between positive and negative reactivity

Figure 2 presents the relationship between reactivity for related and unrelated pairs across experiments (reactivity for unrelated pairs multiplied by -1 for readability). Spearman's rank correlation between reactivity for related and unrelated pairs per experiment was -0.56, $p = .022$. As can be seen in Figure 2, this correlation indicates that experiments showing strong negative reactivity for unrelated

Table 3. Moderator Analyses

Moderator	k	Related pairs				Unrelated pairs			
		Hedges's *g*	SE	95% CI	*Q*	Hedges's *g*	SE	95% CI	*Q*
Language					12.58**				4.32
German	6	0.19*	0.08	[0.03; 0.36]		-0.52***	0.10	[-0.72; -0.31]	
Hebrew	4	0.15	0.11	[-0.08; 0.37]		-0.42**	0.14	[-0.69; -0.15]	
English	7	0.51***	0.06	[0.38; 0.63]		-0.25**	0.08	[-0.41; -0.08]	
Pair types					1.14				1.16
Only related and unrelated	12	0.29***	0.07	[0.15; 0.43]		-0.42***	0.07	[-0.57; -0.27]	
Additional type	5	0.43***	0.11	[0.21; 0.64]		-0.27*	0.12	[-0.50; -0.04]	
Setting					0.92				4.77*
Tightly controlled	6	0.23	0.12	[0.00; 0.46]		-0.59***	0.12	[-0.83; -0.36]	
Unsupervised	11	0.36***	0.07	[0.23; 0.50]		-0.29***	0.07	[-0.42; -0.16]	
Study time[a]					1.92				0.19
< 8 s	3	0.51***	0.14	[0.23; 0.79]		-0.30	0.16	[-0.61; 0.01]	
8 s	12	0.29***	0.07	[0.15; 0.42]		-0.37**	0.08	[-0.53; -0.22]	
Total no. of pairs					0.55				2.93
36-40	4	0.41**	0.13	[0.16; 0.66]		-0.18	0.13	[-0.43; 0.07]	
60	13	0.30***	0.07	[0.17; 0.44]		-0.43***	0.07	[-0.57; -0.29]	

Note. [a] Experiments 1 and 3, where study time varied across participants but was identical across the JOL and no-JOL condition, were omitted from this analysis.
* $p < .05$. ** $p < .01$. *** $p < .001$.

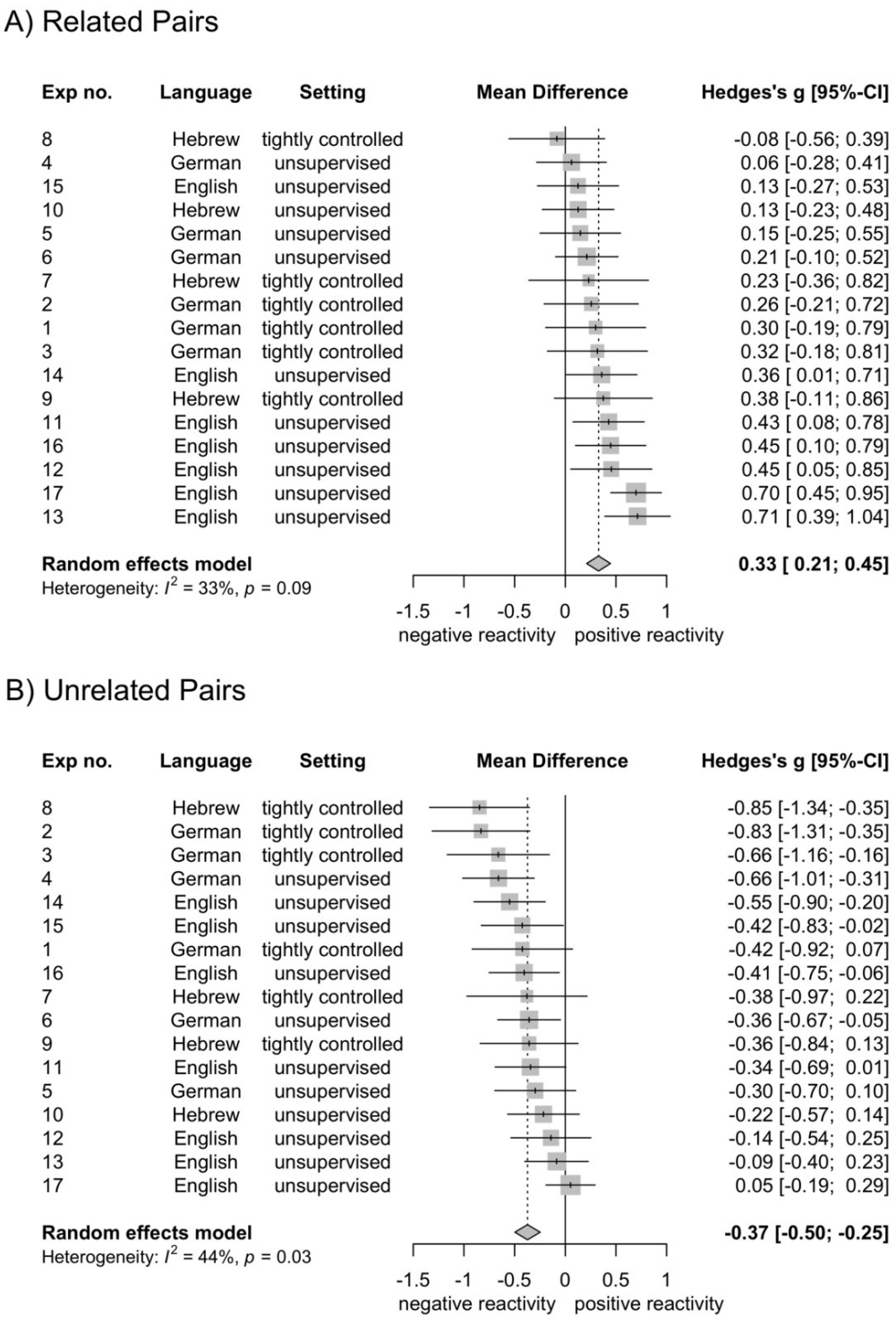

Figure 1. Forest Plot Comparing Recall Performance Between JOL and No-JOL Groups

Note. The sizes of the squares in the forest plot are proportional to the weights of the experiments, which are calculated as the inverse sampling variances.

pairs tended to show weak positive reactivity for related pairs, whereas experiments showing strong positive reactivity for related pairs tended to show weak negative reactivity for unrelated pairs.

Exploratory analyses

Because the language of word pairs moderated positive reactivity and the experimental setting moderated negative reactivity, additional small-scale meta-analyses examined the relationship between recall performance in the no-JOL groups and (1) the language of word pairs and (2) the exper-

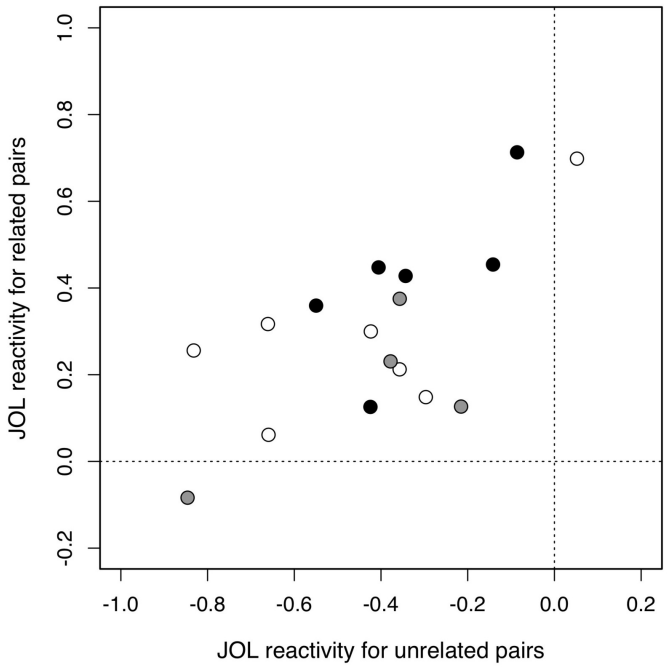

Figure 2. Hedges's g for JOL Reactivity for Unrelated and Related pairs by Experiment

Note. White circles represent experiments with German pairs, grey circles represent experiments with Hebrew pairs, and black circles represent experiments with English pairs.

imental setting. The recall differences between related and unrelated pairs were independent of the language of word pairs, $Q(2) = 2.62$, $p = .270$. However, the language of word pairs reliably moderated the level of recall performance for both related and unrelated pairs, related pairs: $Q(2) = 10.46$, $p = .005$, unrelated pairs: $Q(2) = 6.18$, $p = .046$. Specifically, recall performance was lower in experiments with English materials, related pairs: $g = 0.69$, unrelated pairs: $g = 0.27$, than in experiments with German materials, related pairs: $g = 0.77$, unrelated pairs: $g = 0.39$, or in experiments with Hebrew materials, related pairs: $g = 0.77$, unrelated pairs: $g = 0.40$. The experimental setting moderated neither the recall difference between related and unrelated pairs, $Q(1) < 0.01$, $p = .981$, nor the level of recall performance, related pairs: $Q(1) = 1.26$, $p = .262$, unrelated pairs: $Q(1) = 0.27$, $p = .602$.

Discussion

This study reports 17 experiments that obtained cued-recall performance for related and unrelated word pairs from participants who made JOLs (JOL groups) and from participants who did not make JOLs (no-JOL groups). Samples consisted of participants with three different everyday languages (German, Hebrew, and English) and data were collected in unsupervised online experiments and in tightly controlled experiments.

Consistent with previous research (Double et al., 2018; Myers et al., 2020; Rivers et al., 2021; Soderstrom et al., 2015), making JOLs always increased the effect of word pair relatedness on cued-recall performance. Also consistent with prior work, almost all experiments showed numerical positive reactivity (16 out of 17) that was, however, significant in less than half the experiments (6 out of 17). A novel finding was that almost all experiments showed numerical negative reactivity for unrelated pairs (16 out of 17) that was, however, significant in just about half the experiments (8 out of 17). Importantly, an explanation in terms of low statistical power cannot fully account for these findings because several high-powered experiments did not yield reliable positive or negative reactivity.

Small-scale meta-analyses of the current experiments revealed small-to-moderate positive reactivity of making JOLs for related word pairs. The size of positive reactivity in this study is virtually identical to that obtained in a meta-analysis based on a different set of studies: Hedges's g was .33 in the current study and .32 in the meta-analysis by Double et al. (2018). This is remarkable, because the studies included in Double et al. (2018)'s meta-analysis were all done with English-speaking participants and English word pairs, included both self- and experimenter-paced study times, included both within- and between-experiment manipulations of making JOLs, included pure lists of related or unrelated pairs as well as mixed lists, and also varied more than the experiments reported here in terms of retention interval (no retention interval to 2 days vs. 3 min to 1 day) and study time (5 to 12 s vs. 6 to 8 s).

A moderator analysis revealed that, in the current study, positive reactivity was larger when Prolific participants located in the US or UK studied and recalled English word pairs than when German or Israeli participants studied and recalled German or Hebrew word pairs. In contrast, positive reactivity was independent of the presence of an additional pair type, experimental setting, study time, or the number of word pairs.

An important new finding was moderately sized negative reactivity of making JOLs for unrelated word pairs. This

negative reactivity was stronger in tightly controlled experiments than in unsupervised online experiments, but independent of the language of word pairs, the presence of an additional pair type, study time, or the number of word pairs.

Another novel finding was that experiments revealing strong positive reactivity for related pairs tended to show little or no negative reactivity for unrelated pairs, whereas experiments revealing strong negative reactivity for unrelated pairs tended to show little or no positive reactivity for related pairs.

Exploratory analyses revealed lower recall performance for related and unrelated pairs in experiments with English materials and participants from the US or UK than in experiments with German or Hebrew materials and German or Israeli participants. Even though further research will be needed to assess whether this unexpected finding is robust, we consider it as a potential reason for the observed moderator effects on positive and negative reactivity (see below).

Robustness of negative JOL reactivity for unrelated pairs

What do these results tell us about the size and the robustness of negative JOL reactivity for unrelated pairs? The numbers of numerical and reliable effects in individual experiments and the effect sizes from the reported small-scale meta-analyses all show that negative JOL reactivity for unrelated pairs was similar in size and robustness to positive JOL reactivity for related pairs. Also important, the reported moderator analyses demonstrate that negative JOL reactivity for unrelated pairs is not limited to experiments with German or Hebrew materials and German or Israeli participants or to experiments that included an additional pair type over and above related and unrelated pairs. The results of the moderator analyses are, however, consistent with the possibility that experiments with English materials and participants with English as their first language yield particularly strong positive and particularly weak negative reactivity. This conclusion is based not only on the finding that the language of word pairs moderated positive reactivity but also on the finding that the experimental setting moderated negative reactivity. This is due to the current study being not fully balanced in that all experiments with English materials were unsupervised online experiments, whereas more than half of the experiments with German or Hebrew materials were more tightly controlled. Consequently, the moderating effects of language and experimental setting might both suggest that the balance between positive and negative reactivity depends on the language of word pairs.

At present, we can only speculate about the reasons for why experiments with English materials yielded stronger positive and weaker negative reactivity than experiments with German or Hebrew materials. One possibility is that the observed differences in positive and negative reactivity stem directly from differences in the construction or language of word pairs or from differences in the language of instructions (see Double & Birney, 2019a). Another possibility is that cultural differences directly affect positive and negative reactivity. It is important to note, however, that all samples in the current study were drawn from educated, industrialized, rich, and democratic societies (Henrich et al., 2010) and that prior studies on memory and metamemory yielded similar findings with samples from all three populations (e.g., Halamish, 2018; Rhodes & Castel, 2008; Undorf & Zimdahl, 2019). Yet another possibility is that differences in the language of word pairs are correlated with individual differences that in turn impact positive and negative reactivity. Potentially relevant individual differences might include cognitive abilities, motivation, or self-confidence, among others (see, e.g., Double & Birney, 2019a; Kleitman & Stankov, 2007; Ohtani & Hisasaka, 2018; Tauber & Witherby, 2019). Finally, it is possible that differences in the language of word pairs affected positive and negative reactivity indirectly via producing differences in recall performance. Further research will be needed to distinguish between these possibilities and to determine why word pair language affected the strength of positive and negative reactivity. Importantly, future work should aim to disentangle word pair language from any other factor that might impact reactivity.

Overall, the current study indicates that negative JOL reactivity for unrelated pairs is similarly robust and reliable as positive JOL reactivity for unrelated pairs. This finding raises the question: Why was positive JOL reactivity for related pairs found in many published studies, whereas negative JOL reactivity for unrelated pairs was found in very few published studies (see Mitchum et al., 2016; Rivers et al., 2021)? The current study suggests that this might be indeed related to most published studies on JOL reactivity involving samples from the US, because the reported moderator analyses showed particularly strong positive reactivity for related pairs and particularly weak negative reactivity for unrelated pairs in experiments with English materials and native-English-speaking participants.

It is also possible that published studies rarely reported negative reactivity for unrelated pairs because of publication bias. Specifically, because conditions conforming to the standard paradigm of JOL reactivity often serve as control conditions, researchers may consider experiments as flawed or unreliable when negative reactivity for unrelated pairs dominates over positive reactivity for related pairs. They may therefore be less inclined to submit these experiments for publication (see Cooper et al., 1997). Of course, it is also possible that studies revealing negative JOL reactivity are more often rejected during the peer-review process than studies revealing positive JOL reactivity. Publication bias might also explain why Double et al.'s (2018) meta-analysis of published reactivity studies found a null effect of making JOLs on cued-recall performance for unrelated pairs. A comprehensive meta-analysis that includes both published and unpublished reactivity studies from various researchers will be needed to directly test whether publication bias might have contributed to the low number of published reports of negative reactivity.

Theoretical implications

Robust negative JOL reactivity for unrelated pairs has several important implications. An obvious implication is that datasets in which the interaction between judgment (JOL vs. no-JOL group) and item type (unrelated vs. related word pairs) is driven by negative JOL reactivity for unrelated pairs are no aberration and should not be considered unreliable or flawed. At a theoretical level, robust negative reactivity for unrelated pairs indicates that theoretical accounts of JOL reactivity need to address this phenomenon.

As explained in the Introduction, negative JOL reactivity is inconsistent with the attentional reorienting account (Rivers et al., 2021; Tauber & Witherby, 2019; Zhao et al., 2021) according to which reactivity is due to JOL prompts reorienting participants' attention to the study pair and, consequently, increased engagement with the pairs. Robust negative reactivity for unrelated pairs therefore indicates that attentional reorienting cannot account for JOL reactivity for unrelated pairs. Moreover, robust negative JOL reactivity indicates that the cue-strengthening hypothesis (Soderstrom et al., 2015), according to which making JOLs improves memory performance because it strengthens the cues underlying JOLs, needs to be supplemented with the assumption that making JOLs interferes with learning (dual-task costs; Janes et al., 2018; also see Mitchum et al., 2016). Finally, robust negative JOL reactivity is consistent with the changed-goal hypothesis (Mitchum et al., 2016) assuming that making JOLs shifts learners' goals towards selectively focusing on related pairs at the expense of unrelated pairs.

The current finding that neither short study times nor large numbers of to-be-studied word pairs—both of which could plausibly increase dual-task costs—moderated negative JOL reactivity for unrelated pairs and that unsupervised experimental settings reduced rather than increased negative reactivity might be taken to suggest that dual-task costs of making JOLs play a minor role in negative reactivity. However, this conclusion would be premature for several reasons. First, the reported moderator analyses were based on a relatively small number of experiments in each subgroup and therefore had limited statistical power. Moreover, it is possible that larger variability in study time and in the number of study pairs than implemented in the current experiments would be required for appreciable differences in negative JOL reactivity to emerge. Also, it might be possible that greater involvement and higher motivation in tightly controlled experiments increase dual-task costs, because participants take the JOL task more seriously. Finally, results from prior studies are consistent with dual-task costs contributing to JOL reactivity (Mitchum et al., 2016; Rivers et al., 2021; Tauber & Witherby, 2019).

Whereas both the cue-strengthening hypothesis with dual-task costs (Janes et al., 2018; Soderstrom et al., 2015) and the changed-goal hypothesis (Mitchum et al., 2016) are consistent with the existence of robust negative reactivity for unrelated pairs, only the former hypothesis is compatible with the observed inverse relationship between JOL reactivity for unrelated and related pairs across experiments. According to the cue-strengthening hypothesis with dual-task costs, high dual-task costs should not only impair memory for unrelated pairs but also reduce the benefits of cue-strengthening for memory for related pairs. Consequently, experiments that yield strong positive JOL reactivity for related pairs should yield strong negative JOL reactivity for unrelated pairs. In contrast, the changed-goal hypothesis (Mitchum et al., 2016) predicts positive relations between JOL reactivity for unrelated and related pairs, because strong strategy shifts towards related pairs should impair memory for unrelated pairs but improve memory for related pairs. Consequently, experiments that yield strong positive JOL reactivity for related pairs should yield little or no negative JOL reactivity for unrelated pairs and experiments that yield strong negative reactivity for unrelated pairs should yield reduced or no positive reactivity for related pairs. The inverse relationship obtained in this study therefore supports the cue-strengthening hypothesis with dual-task costs (Janes et al., 2018; Soderstrom et al., 2015) and argues against the changed-goal hypothesis (Mitchum et al., 2016).

Practical implications

Based on the published research on JOL reactivity suggesting that positive JOL reactivity for related pairs is a robust phenomenon, whereas negative JOL reactivity for unrelated pairs is not, various researchers have suggested that eliciting JOLs might enhance learning in educational settings and academic achievement (e.g., Double & Birney, 2019b; Janes et al., 2018; Tekin & Roediger, 2020; Witherby & Tauber, 2017). For instance, Soderstrom et al. (2015) and Witherby and Tauber (2017) argued that making JOLs during learning could improve exam performance and Tekin and Roediger (2020) proposed that encouraging students to monitor their learning could be an impactful learning strategy. Indeed, soliciting JOLs for the purpose of fostering real-world learning and academic achievement has the obvious advantage that JOL prompts can be easily integrated into a wide variety of learning situations and that predicting one's own performance does not require continuous support by others. JOLs thus appeared to be a potentially effective low-cost method for enhancing self-regulated learning and academic achievement.

In contrast, the robust negative JOL reactivity for unrelated pairs reported in this study raises the possibility that soliciting metacognitive judgments or encouraging learners to monitor their learning could have detrimental effects, at least for specific materials. Even though we fully acknowledge that examining to what extent JOL reactivity transfers beyond word-pair experiments is a research question it its own right (e.g., Ariel et al., 2021; Schäfer & Undorf, 2024), the current findings clearly suggest that a more nuanced view on the potential effects of making JOLs on learning and achievement in educational settings is in order.

Limitations

As with any moderator analysis, the moderator analyses reported here are observational and do not allow for causal

conclusions (Borenstein & Higgins, 2013). They therefore do not make targeted experimental work on selected moderator variables redundant. Even though only experimental setting moderated the negative effects of making JOLs on memory for unrelated pairs (and did so in the unpredicted direction), it is plausible that additional factors modulate negative reactivity. As was mentioned above, the moderating effects of additional moderator variables examined here might become apparent in a comprehensive meta-analysis based on a larger number of individual studies or when manipulated experimentally. Another possibility is that procedural details or individual differences that were not considered or assessed in this study moderate negative JOL reactivity. While we ensured that all examined moderator variables are relevant in view of prior theoretical or empirical work, we could not include all potentially relevant moderator variables. Thus, further research will be needed to examine these possibilities.

Even though the 17 experiments reported here clearly demonstrate that negative reactivity for unrelated word pairs can be as large and robust as positive reactivity for related word pairs, it is an open question whether this conclusion also holds for published and unpublished research by others. A comprehensive meta-analysis would be well-suited for addressing this question.

Conclusions

The current study demonstrates that making JOLs does not only improve cued-recall memory for related word pairs but also impairs cued-recall memory for unrelated word pairs, with positive and negative JOL reactivity being similar in effect size and robustness. Until we have a more complete understanding of when making JOLs is more likely to result in positive reactivity for related pairs or, alternatively, in negative reactivity for unrelated pairs, the best indicator of JOL reactivity is an increased effect of word pair relatedness on cued-recall performance in the JOL group (also see Janes et al., 2018). We therefore recommend that researchers regard all Judgment X Item type interactions that indicate a larger effect of word pair relatedness on memory in JOL than in no-JOL conditions as reliable evidence for JOL reactivity and call for further investigations of the factors that affect the nature of this interaction.

Funding

This research was supported by grants No. 350/15 and 1082/21 from the Israel Science Foundation to Vered Halamish.

Competing Interests

None to disclose.

Data Accessibility Statement

All data and analysis code are available at https://osf.io/q3jgm. None of the experiments was preregistered.

Submitted: March 26, 2024 PDT, Accepted: April 05, 2024 PDT

References

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*(2), 693–712. https://doi.org/10.1007/s10648-020-09556-8

Balduzzi, S., Rücker, G., & Schwarzer, G. (2019). How to perform a meta-analysis with R: A practical tutorial. *Evidence-Based Mental Health*, *22*(4), 153–160. https://doi.org/10.1136/ebmental-2019-300117

Borenstein, M., & Higgins, J. P. T. (2013). Meta-Analysis and Subgroups. *March*, 134–143. https://doi.org/10.1007/s11121-013-0377-7

Bröder, A., & Undorf, M. (2019). Metamemory viewed through the judgment lens. *Acta Psychologica*, *197*, 153–165. https://doi.org/10.1016/j.actpsy.2019.04.011

Clifford, S., & Jerit, J. (2014). Is there a cost to convenience? An experimental comparison of data quality in laboratory and online studies. *Journal of Experimental Political Science*, *1*(2), 120–131. https://doi.org/10.1017/xps.2014.5

Cooper, H., DeNeve, K., & Charlton, K. (1997). Finding the missing science: The fate of studies submitted for review by a human subjects committee. *Psychological Methods*, *2*(4), 447–452. https://doi.org/10.1037/1082-989X.2.4.447

Double, K. S., & Birney, D. P. (2019a). Do confidence ratings prime confidence? *Psychonomic Bulletin and Review*, *26*(3), 1035–1042. https://doi.org/10.3758/s13423-018-1553-3

Double, K. S., & Birney, D. P. (2019b). 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. (2018). 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, & 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

Gould, S. J. J., Cox, A. L., Brumby, D. P., & Wiseman, S. (2015). Home is where the lab is: A comparison of online and lab data from a time-sensitive study of interruption. *Human Computation*, *2*(1), 45–67. https://doi.org/10.15346/hc.v2i1.4

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. (2020). Do learners spontaneously monitor their memory? *Zeitschrift Für Psychologie*, *228*(4), 304–305. https://doi.org/10.1027/2151-2604/a000429

Halamish, V., & Undorf, M. (2023). Why do judgments of learning modify memory? Evidence from identical pairs and relatedness judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *49*(4), 547–556. https://doi.org/10.1037/xlm0001174

Henik, A., Rubinstein, O., & Anaki, D. (2005). *Norms for Hebrew words*. Ben-Gurion University of the Negev.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

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

Kleitman, S., & Stankov, L. (2007). Self-confidence and metacognitive processes. *Learning and Individual Differences*, *17*(2), 161–173. https://doi.org/10.1016/j.lindif.2007.03.004

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

Maxwell, N. P., & Huff, M. J. (2024). Judgment of learning reactivity reflects enhanced relational encoding on cued-recall but not recognition tests. *Metacognition Learning*, *19*(1), 189–213. https://doi.org/10.1007/s11409-023-09369-4

Melinger, A., & Weber, A. (2006). *Database of noun associations for German*. http://www.coli.uni-saarland.de/projects/nag/

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

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*(5), 519–533. https://doi.org/10.1016/S0022-5371(77)80016-9

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

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). *The University of South Florida word association, rhyme, and word fragment norms*. http://w3.usf.edu/FreeAssociation/

Ohtani, K., & Hisasaka, T. (2018). Beyond intelligence: A meta-analytic review of the relationship among metacognition, intelligence, and academic performance. *Metacognition and Learning*, *13*(2), 179–212. https://doi.org/10.1007/s11409-018-9183-8

Peer, E., Rothschild, D., Gordon, A., Evernden, Z., & Damer, E. (2021). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*. https://doi.org/10.3758/s13428-021-01694-3

Rhodes, M. G., & Castel, A. D. (2008). Memory predictions are influenced by perceptual information: Evidence for metacognitive illusions. *Journal of Experimental Psychology: General*, *137*(4), 615–625. https://doi.org/10.1037/a0013684

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

Schäfer, F., & Undorf, M. (2024). On the educational relevance of immediate judgment of learning reactivity: No effects of predicting one's memory for general knowledge facts. *Journal of Applied Research in Memory and Cognition*, *13*(1), 113–123. https://doi.org/10.1037/mac0000113

Sicken, L. (2019). *Learning cue validities in metamemory* [Unpublished master's thesis]. University of Mannheim.

Soderstrom, N. C., Clark, C. T., Halamish, V., & Bjork, E. L. (2015). Judgments of learning as memory modifiers. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *41*(2), 553–558. https://doi.org/10.1037/a0038388

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

Undorf, M., & Zimdahl, M. F. (2019). Metamemory and memory for a wide range of font sizes: What is the contribution of perceptual fluency? *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *45*(1), 97–109. https://doi.org/10.1037/xlm0000571

Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics*, *30*(3), 261–293. https://doi.org/10.3102/10769986030003261

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

Zhao, W., Li, B., Shanks, D. R., Zhao, W., Zheng, J., Hu, X., Su, N., Fan, T., Yin, Y., Luo, L., & Yang, C. (2021). When judging what you know changes what you really know: Soliciting metamemory judgments reactively enhances children's learning. *Child Development*, 1–13. https://doi.org/10.1111/cdev.13689

Appendix A

The 20 identical pairs used in Experiments 11, 16, and 17 had cues that were identical to the targets (e.g., *KISS* – *KISS*).

The 20 backward related pairs used in Experiment 2 had a mean forward associative strength of .00 (.00-.07) and a mean backward associative strength of .20 (.04-.53, Melinger & Weber, 2006; Sicken, 2019).

The 20 related pairs printed in alternating font in Experiment 14 were identical in associative strength to the related pairs printed in regular font (*M* = .52, .34-.75, Nelson et al., 1998).

Table A1. Descriptive and Inferential Statistics for the Additional Item Types.

Experiment	Item type	Condition	Mean recall (*SD*)	Effect of judgment	Mean JOL (*SD*)
2	Backward related	JOL	.49 (.23)	$t(70) = 0.37$, $p = .714$, $d = 0.09$	54.63 (16.71)
		no JOL	.48 (.22)		
11	Identical	JOL	.78 (.22)	$t(126) = 2.62$, $p = .010$, $d = 0.47$	69.06 (22.74)
		no JOL	.66 (.30)		
14	Related in aLtErNaTiNg font	JOL	.74 (.18)	$t(127) = 0.59$, $p = .554$, $d = 0.10$	61.09 (18.36)
		no JOL	.72 (.21)		
16	Identical	JOL	.74 (.18)	$t(129) = 0.29$, $p = .773$, $d = 0.05$	72.39 (17.00)
		no JOL	.73 (.23)		
17	Identical	JOL	.77 (.22)	$t(260) = 2.08$, $p = .039$, $d = 0.26$	75.35 (19.81)
		no JOL	.71 (.21)		

Supplementary Materials

Peer Review Communication

Download: https://collabra.scholasticahq.com/article/117108-making-judgments-of-learning-either-enhances-or-impairs-memory-evidence-from-17-experiments-with-related-and-unrelated-word-pairs/attachment/225581.docx?auth_token=fNxCGiftt88X3I_tWGE4