MEMORY
https://doi.org/10.1080/09658211.2026.2652393

Routledge
Taylor & Francis Group

REVIEW ARTICLE

OPEN ACCESS Check for updates

Mending metacognitive illusions in JOLs: when neither cognitive nor metacognitive feedback is effective

Sofia Navarro-Báez[a], Arndt Bröder[b] and Monika Undorf[a]

[a]Department of Psychology, Technical University of Darmstadt, Darmstadt, Germany; [b]Department of Psychology, School of Social Sciences, University of Mannheim, Mannheim, Germany

ABSTRACT

Metacognitive monitoring of cognitive processes is not always accurate. Metacognitive illusions occur when metacognitive judgments rely on invalid information or fail to rely on valid information. This study tested the effectiveness of different forms of feedback in mending metacognitive illusions in judgments of learning (JOLs). Across four experiments, participants completed three study-test cycles with JOLs in which they studied different word lists. Participants received feedback or no feedback after each cycle. In Experiments 1 and 2, cognitive feedback about recall performance and JOL for each item was provided. In Experiments 3 and 4, additional metacognitive feedback about metacognitive illusions during the task was provided. Results showed that cognitive feedback was not effective for mending the font size illusion (Experiments 1 and 2), the stability bias (Experiment 1), or the font format illusion (Experiment 2). Additional metacognitive feedback partially remedied the stability bias in Experiment 3, but this effect did not replicate in Experiment 4. Regardless of whether participants received feedback and what type it was, the font size illusion decreased across cycles when manipulated orthogonally to a valid cue (Experiments 1, 3, and 4). In conclusion, this study shows that neither cognitive nor metacognitive feedback remedy metacognitive illusions.

ARTICLE HISTORY
Received 14 November 2025
Accepted 20 March 2026

KEYWORDS
Metacognitive illusions; metamemory; judgments of learning; cognitive feedback; metacognitive feedback

Metacognitive monitoring – the ongoing evaluation of cognitive processes – is important because it guides behaviour (Nelson & Narens, 1990). For example, a student with accurate metacognitive monitoring may identify which topics she has mastered sufficiently for the exam and continue to study those that she has not yet mastered. Experimental studies show that participants with higher monitoring accuracy can regulate their learning better by selecting material to restudy more appropriately (Dunlosky et al., 2021; Thiede et al., 2003; Tullis & Benjamin, 2012). This ultimately leads to better grades, as shown by a meta-analysis demonstrating the positive link between metacognition and academic performance, even when controlling for intelligence (Ohtani & Hisasaka, 2018). Unfortunately, metacognition is not always helpful because monitoring is sometimes inaccurate. Because metacognitive monitoring judgments are inferential and rely on cues (Koriat, 1997), their accuracy suffers when they rely on invalid cues or fail to rely on the valid ones, resulting in *metacognitive illusions* (Undorf et al., 2022a, 2022b).

Improving self-regulated learning thus requires metacognitive awareness of cue validity, meaning that metacognitive judgments should rely on valid cues and ignore invalid ones. *How to improve metacognitive awareness of cue validity* is therefore a practically relevant question. Research has demonstrated that correcting metacognitive illusions is very difficult, often ineffective, and yields only small improvements when successful (e.g., Dunlosky & Hertzog, 2000; Hertzog et al., 2009; Kornell & Bjork, 2009; Mueller et al., 2015; Pan & Rivers, 2023; Yan et al., 2016). In this study, we examined the effectiveness of various feedback forms in correcting metacognitive illusions in *judgments of learning* (JOLs) – predictions of future memory performance (Rhodes, 2016). In Experiments 1 and 2, we tested *cognitive feedback* (Balzer et al., 1989) – the presentation of JOL and recall status for each studied item – which is known to improve cue utilisation accuracy in judgments about the external world. Because the cognitive feedback alone was insufficient to correct metacognitive illusions, we included additional

CONTACT Sofia Navarro-Báez sofia.navarro@tu-darmstadt.de Department of Psychology, Technical University of Darmstadt, Alexanderstr. 10 (S1|15), 64823 Darmstadt, Germany

Supplemental data for this article can be accessed online at https://doi.org/10.1080/09658211.2026.2652393.

metacognitive feedback – written information about illusory metacognition – in Experiments 3 and 4. In the following, we review the literature on methods used to correct metacognitive illusions, highlighting related difficulties, and deriving the two types of feedback that may be effective in reducing such illusions. Consequently, we then discuss the cognitive feedback method used in Experiments 1 and 2, and the additional metacognitive feedback used in Experiments 3 and 4.

Methods used to foster metacognitive awareness

Task experience

Task experience across multiple learning-test cycles has been used as a method to improve metacognitive awareness of cues (Castel, 2008; Dunlosky & Hertzog, 2000; Hertzog et al., 2009; Mueller et al., 2015; Pan & Rivers, 2023; Sungkhasettee et al., 2011; Tauber & Rhodes, 2010). The idea behind it is that encoding and retrieval experiences from learning and test phases aid learners in becoming aware of cues that are helpful for their learning and memory. Because memory predictions for repeatedly studied and tested materials can be based on the memory-for-past-test heuristic (Ariel & Dunlosky, 2011; Finn & Metcalfe, 2008; Koriat & Bjork, 2006a; Tauber & Rhodes, 2012), task experience is especially relevant for cue discovery when novel materials are learned across cycles. Studies with novel item lists across cycles show that participants acquire correct knowledge about study strategies (e.g., imagery is more effective than repetition) as measured by global predictions and strategy effectiveness ratings. However, this knowledge is not reflected in item-by-item metacognitive judgments (Dunlosky & Hertzog, 2000; Hertzog et al., 2009; Mueller et al., 2015). Similarly, other studies have shown that task experience is ineffective in improving the sensitivity of item-by-item metacognitive judgments to the effects of word orientation (i.e., greater memory performance for inverted than upright words; Sungkhasettee et al., 2011), serial position effects in memory (Castel, 2008), and the lack of a font size effect in memory (Rhodes & Castel, 2008). An exception is a study in which task experience was effective at improving the accuracy of memory predictions for occupations and surnames (Tauber & Rhodes, 2010). Overall, despite being readily accessible, direct task experience appears to have limited effectiveness at best.

Warnings

Warnings that encourage the use of valid cues and the avoidance of invalid ones when metacognitive illusions occur are another method used to foster metacognition accuracy. For instance, Kornell and Bjork (2009) warned participants to keep in mind that their future memory would improve with the number of study opportunities. However, participants continued underestimating the impact of future study opportunities on memory performance. Similarly, Rhodes and Castel (2008) warned participants that the font size of words would not affect their future memory, but participants still based their predictions on font size. Yan et al. (2016) informed participants about the superiority of interleaving exemplars of to-be-learned categories over blocking exemplars by category, but participants still preferred blocking. At the same time, two studies indicate that warnings can be successful when individually tailored and detailed information is provided to participants (Koriat & Bjork, 2006b; Miller & Geraci, 2011). However, this approach is often difficult to implement in experimental studies. To sum up, metamemory studies have demonstrated the persistence of metacognitive illusions and a lack of reliance on valid cues even when experimental instructions provide very explicit warnings.

Test performance outcome feedback

Test performance feedback involves the presentation of remembered and non-remembered words in a metamemory task. This method allows participants to identify effective study strategies that they might otherwise fail to recognise due to an inferential deficit – that is, limited cognitive resources for monitoring test performance and drawing inferences about valid cues, known as the *inferential deficit hypothesis* (Dunlosky & Hertzog, 2000; Matvey et al., 2002). Thus, studies have used cue-related performance feedback as a support for participants to distinguish the validity of cues (Mueller et al., 2015; Pan & Rivers, 2023; Tullis et al., 2013). Tullis et al.'s (2013) study provided participants with the number of correctly recalled restudied and pretested items. This led to the accurate identification of pretesting as a more effective strategy than restudying. However, the study by Pan and Rivers (2023) found that test performance feedback did not increase awareness of the pretesting strategy, and positive results were obtained only when participants were prompted to recall their predictions. Importantly, both studies used global predictions, which primarily reflect metamemory beliefs, rather than item-by-item predictions, where learning experiences play a greater role. Overall, performance feedback seems to be effective for improving global judgments, but not for improving item-by-item judgments.

Main types of feedback tested in this study

Cognitive feedback

Given that warnings are ineffective or difficult to implement, and that task experience and test performance feedback only improve global judgments but not item-by-item judgments, we expected that a successful method for mending metacognitive illusions would need to teach how the cues affect both the memory criterion and judgment. A

seminal review by Balzer et al. (1989) on judgments about the external world indicates that so-called *cognitive feedback* is effective for improving the cue basis of judgments in contrast to mere outcome feedback. Cognitive feedback refers to (1) information about the relation between cues and criterion (i.e., cue validities), and (2) information about how such a relation between cues and criterion is perceived by the participant when making judgments (i.e., cue utilisation). For example, cognitive feedback in a metamemory task involves the presentation of the item-by-item actual memory performance and judgment, each item categorised by the cues manipulated in the study. This feedback is cognitive in nature because it provides information regarding how cues are utilised for making judgments, in contrast to mere outcome feedback that provides only information about the criterion.

Several studies have demonstrated that cognitive feedback improves the cue basis and accuracy of judgments about the external world (Karlsson et al., 2004; Little & Lewandowsky, 2009; Newell et al., 2009; Seong & Bisantz, 2008; Smithson et al., 2023). Since both judgments about external criteria and metacognitive judgments rely on probabilistic cues, cognitive feedback may also help learners to distinguish the different predictive validities of cues in metacognition. Most relevant for present purposes, the opportunity to relate JOLs to actual memory performance, for each item categorised by cues, should support an understanding of which cues being used for JOLs are valid for memory performance (e.g., "this helps versus impairs memory"), and which ones are not valid (e.g., "this does not affect memory").

Metacognitive feedback

At the same time, cognitive feedback may be insufficient to correct metacognitive illusions, as various biases and pre-existing beliefs can hinder learning from item-by-item feedback. For example, people often focus on positive information that confirms an incorrect hypothesis rather than on evidence that refutes it (Cooper & Vallée-Tourangeau, 2021). Given this potential limitation, we additionally developed a *metacognitive feedback* approach aimed at directly targeting such biases based on Fiedler et al.'s (2020) recommendations on effective debiasing treatment. Fiedler et al. state that an effective debiasing treatment should not only provide information about judgments that are "correct" versus "incorrect" but also relate to (a) the representation of stimuli, and (b) provide explicit instructions about how to make accurate judgments. Regarding the former, metacognitive feedback should consider the first-person perspective and explain which metacognitive stimuli representations likely occur during learning (e.g., *words in large font size seem easier*). Metacognitive feedback differs from cognitive feedback as described by Balzer et al. (1989) in that participants are presented not only with information about cue validities and cue utilisation (i.e., recall status and JOL for each studied item), but also with explicit references to illusory stimulus representations at the metacognitive level (e.g., fluency experiences arising during learning). Informing participants about the cognitions that may occur during learning from a first-person perspective makes the participants' experiences seem valid but also emphasises that these experiences are unreliable.

So far, the warnings provided in metamemory studies have not included metacognitive feedback, as they only indicated which cues to consider and/or ignore (i.e., the correct vs. incorrect aspect). For instance, in the study by Rhodes and Castel (2008), participants were warned that the font size of a word does not influence actual memory but were not informed about possible fluency experiences or automatic cognitions during learning. This type of warning carries the risk that participants may ignore it if their own experiences – such as feelings of ease when learning large-font words – suggest otherwise. Further, when their experiences are not reflected in the feedback, participants may prefer to rely on their own experiences because they view themselves as experts on their own cognition (see Yan et al., 2016).

This study

In this study, we tested the effectiveness of these various forms of feedback in mending metacognitive illusions on judgments of learning (JOLs). Since reliance on valid cues is a determinant factor for high relative accuracy or resolution (i.e., how well judgments discriminate between remembered and non-remembered items), we also expected that judgment resolution would be enhanced when the illusions are reduced. For exploratory reasons, we also considered calibration indexes.

In each of four experiments, participants completed three study-test cycles with JOLs and received either feedback or no feedback after each study-test cycle. In all four experiments, we orthogonally manipulated two cues in total, and one of these cues was font size (18 point vs. 48 point), a cue that is overweighted in JOLs – large-font words elicit higher JOLs than small-font words but font size has a very small or no effect on recall performance (Chang & Brainerd, 2022; Luna et al., 2018; Rhodes & Castel, 2008). In Experiment 1, 3, and 4, we additionally manipulated the announced number of study presentations (1 vs. 2), a cue that is underweighted in JOLs – JOLs made during the first presentation of a word often do not differ between words announced to be learned once and twice even though twice-learned words are better recalled. This illusion is known as stability bias (Kornell & Bjork, 2009). In Experiment 2, the additional cue manipulated was font format (standard vs. aLtErnAtiNg), a cue that is overweighted in JOLs – standard-format words elicit higher JOLs than alternating-format words but font format usually has no effect on recall performance (Rhodes & Castel, 2008; but see Mueller et al., 2013).

In each experiment, we compared experimental groups that received feedback against a control group that did not receive feedback. In Experiment 1, the experimental groups received either (1) "outcome feedback" consisting of item-by-item memory performance, (2) "cognitive feedback" consisting of item-by-item memory performance along with JOLs, or (3) "social-reference-feedback" consisting of average memory performance of previous participants in the experiment. Prior research has shown that social information can be integrated into confidence ratings (De Martino et al., 2017) and we examined whether social reference information would be sufficient for improvements or would even result in stronger improvements than the cognitive feedback. In Experiment 2, all previous forms of cognitive feedback were collapsed into one group only, "catch-all-cognitive-feedback" group. In Experiment 3, the cognitive feedback group was identical to the one in Experiment 2 but with additional metacognitive feedback, "cognitive-plus-metacognitive-feedback" group. In Experiment 4, one "cognitive-plus-metacognitive-feedback" group received cognitive and metacognitive feedback as in Experiment 3, and the other group received metacognitive feedback only, "metacognitive-feedback-only" group (see Table 1).

We hypothesised that feedback in comparison to no feedback would lead to JOLs relying increasingly on valid cues (i.e., announced number of study presentations) and people ignoring invalid cues (i.e., font size, font format). This means that, after participants receive feedback on a first study-test cycle, JOLs should be higher for words with two rather than one study opportunity announced. Further, JOLs should not differ between large-font and small-font words, and between standard-font and alternating-font words after feedback. Improvements in experimental and control groups would indicate that study-test experience fosters learning cue validities and implementing them in JOLs.

Experiment 1

Experiment 1 aimed to test the effectiveness of cognitive feedback (Balzer et al., 1989) for increasing the influence of announced number of study presentations (i.e., valid cue) and decreasing the influence of font size (i.e., invalid cue) on JOLs. Experiment 1 entailed four between-subjects groups. In the control group, participants received no feedback, so they only had their own memory of test performance as feedback on cue validity (Ariel & Dunlosky, 2011; Finn & Metcalfe, 2008; Tauber & Rhodes, 2012). In the outcome feedback group, referred to as the recall-feedback group, participants saw all words they had recalled and not recalled, organised by the two cues (see Figure 1). In the cognitive feedback group, referred to as the recall-and-JOL-feedback group, each word was accompanied by the JOL participants had given to it during study. Finally, the social-reference-feedback group was presented with a table showing the average performance of other participants doing the same task.

We hypothesised that the cognitive feedback provided to the recall-and-JOL-feedback group would improve cue weighting in JOLs (i.e., positive effect of announced number of study presentations and no effect of font size) and, in turn, increase relative accuracy. At the same time, it was an open question whether the outcome feedback in the recall-feedback group would lead to better cue weighting and accuracy as suggested by the inferential deficit hypothesis (Hertzog et al., 2009; Matvey et al., 2002) or whether social reference information would be sufficient for improvements or even go beyond the cognitive feedback (De Martino et al., 2017).

Method

Design

The design was a mixed design with font size (18, 48 point), number of study presentations (1, 2), and study-test cycle (1, 2, 3) as within-subjects factors and feedback group (no-feedback, recall-feedback, recall-and-JOL feedback, social-reference-feedback) as a between-subjects factor.

Materials

Stimuli were 120 German six-letter nouns. All words were of neutral valence ($M = 0.36$, $SD = 0.93$; rated on a 7-point scale, -3 = *very negative* to 3 = *very positive*), moderate arousal ($M = 2.59$, $SD = 0.80$; rated on a 5-point scale, 1 = *low arousal* to 5 = *high arousal*), and moderate concreteness ($M = 4.87$, $SD = 1.67$; rated on a 7-point scale, 1 = *low imageability* to 7 = *high imageability*). All normed values were taken from Võ et al. (2009). We constructed three study lists of 40 items that were parallel in all word characteristics. For each participant, study lists were randomly assigned to study-test cycles. For each participant, 20 randomly chosen words were presented once for study and the remaining 20 words were presented twice for study. One randomly selected half of the once- and twice presented words were displayed in small 18-point Arial font or in a large 48-point Arial font. The first four items represented each combination of number of study opportunities and font size and were used as buffer items that were not included in the analysis. Items were presented in a new random order for each participant.

Participants

Our primary analysis was the three-way interaction among group, cycle, and cue [number of future study presentations or font size] in a four-way ANOVA. We decided on a convenience sample of 40 participants per group. This sample size provides a statistical power of $(1-\beta) > .99$ for detecting medium-sized ($f = .25$, equivalent to $\eta_p^2 = .06$)

Table 1. Type of feedback, cues manipulated, and outcome in each of the four experiments.

Experiment	Types of feedback	Cues manipulated	Outcomes
Experiment 1	1. Outcome (recall only) 2. Cognitive (recall-and-JOL) 3. Social reference	Font size (invalid cue) Announced # of study presentations (valid cue)	No group differences in font size illusion No group differences in stability bias
Experiment 2	1. Catch-all-cognitive *(recall-and-JOL + social-reference)*	Font size (invalid cue) Font format (invalid cue)	No group differences in font size illusion No group differences in font format illusion
Experiment 3	1. Cognitive-plus-metacognitive (recall-and-JOL + social-reference + information about illusory metacognition)	Font size (invalid cue) Announced # of study presentations (valid cue)	No group differences in font size illusion Reduction of stability bias in cognitive-plus-metacognitive-feedback group
Experiment 4	1. Cognitive-plus-metacognitive *(recall-and-JOL + social-reference + information about illusory metacognition)* 2. Metacognitive-only *(social-reference + information about illusory metacognition)*	Font size (invalid cue) Announced # of study presentations (valid cue)	No group differences in font size illusion No group differences in stability bias

main effects of the within-subjects factors and interaction effects with $\alpha = .05$ in a mixed ANOVA and a statistical power of $(1\text{-}\beta) = .87$ for detecting medium-sized main effects of the between-subjects factor, when assuming a correlation of .50 between repeated measures (G*Power 3; Faul et al., 2007).

Participants were 160 University of Mannheim undergraduates. Participants were randomly allocated to the four feedback groups ($n = 40$ in each group[1]). We excluded participants who assigned the same JOL to all items in one or more study phases ($n = 2$) or who had zero recall performance in one or more test phases ($n = 2$). The final sample included 156 participants with a mean age of 23 years ($SD = 4.10$), $n = 37$ in the no-feedback group, $n = 41$ in the recall-feedback group, $n = 40$ in the recall-and-JOL-feedback group, $n = 38$ in the social-reference-feedback group.

Procedure

The experiment consisted of three study-test cycles, each of which included a study phase with JOLs, a filler task, a free recall test, and feedback for the experimental groups (see Figure 2). Instructions informed participants that in each cycle they would study 40 words and would be asked to recall as many words as they could remember in a memory test. Participants were also told that they would be asked to predict the chance of recalling each word immediately after studying it and that they would have an extra study opportunity for some words before the test. At study, each word appeared on the screen for 4 s. Immediately afterwards, the number of study presentations (1 vs 2) and the JOL prompt *Chance of recall (0-100)?* appeared on the screen, and participants pressed on one of 11 keys labelled 0, 10, … , 90, and 100 to make their JOL. Consequently, we obtained one JOL for once-presented words and two JOLs for twice-presented words. A 200-ms blank screen preceded the presentation of each word. Following a 1-min numerical filler task, participants had 4 min to write down as many studied words as they could remember. At the end of each study-test cycle, participants in the no-feedback group typed examples of one randomly chosen category (i.e., mountains, capitals, or rivers) for 3 min. Participants in the recall-feedback group saw an overview of the words they had remembered and had not remembered organised by font size and number of study presentations (see Figure 1). Recall feedback remained on the screen for as long as participants wished and for a minimum duration of 45 s. Participants in the recall-and-JOL-feedback group saw the same organised overview of remembered and not-remembered words for a minimum duration of 45 s as the recall-feedback group, however, complemented by their own JOLs from the study phase (see Figure 1). All participants in the social-reference-feedback group saw the same overview of the mean number of words remembered as a function of font size and number of study presentations, based on data from participants in a previous experiment. This overview was also presented for a minimum duration of 45 s and was followed by an explanation stating that many students have overestimated the influence of font size on their memory and underestimated the influence of additional study opportunity (see Figure 1).

Immediately prior to the next block, participants from all feedback groups then responded to the question: *What did you learn about your learning and memory from this feedback?* (see Supplementary Material 1).

Results

Effects are considered significant based on an alpha level of .05 and a Greenhouse-Geisser correction was applied when the sphericity assumption was violated.

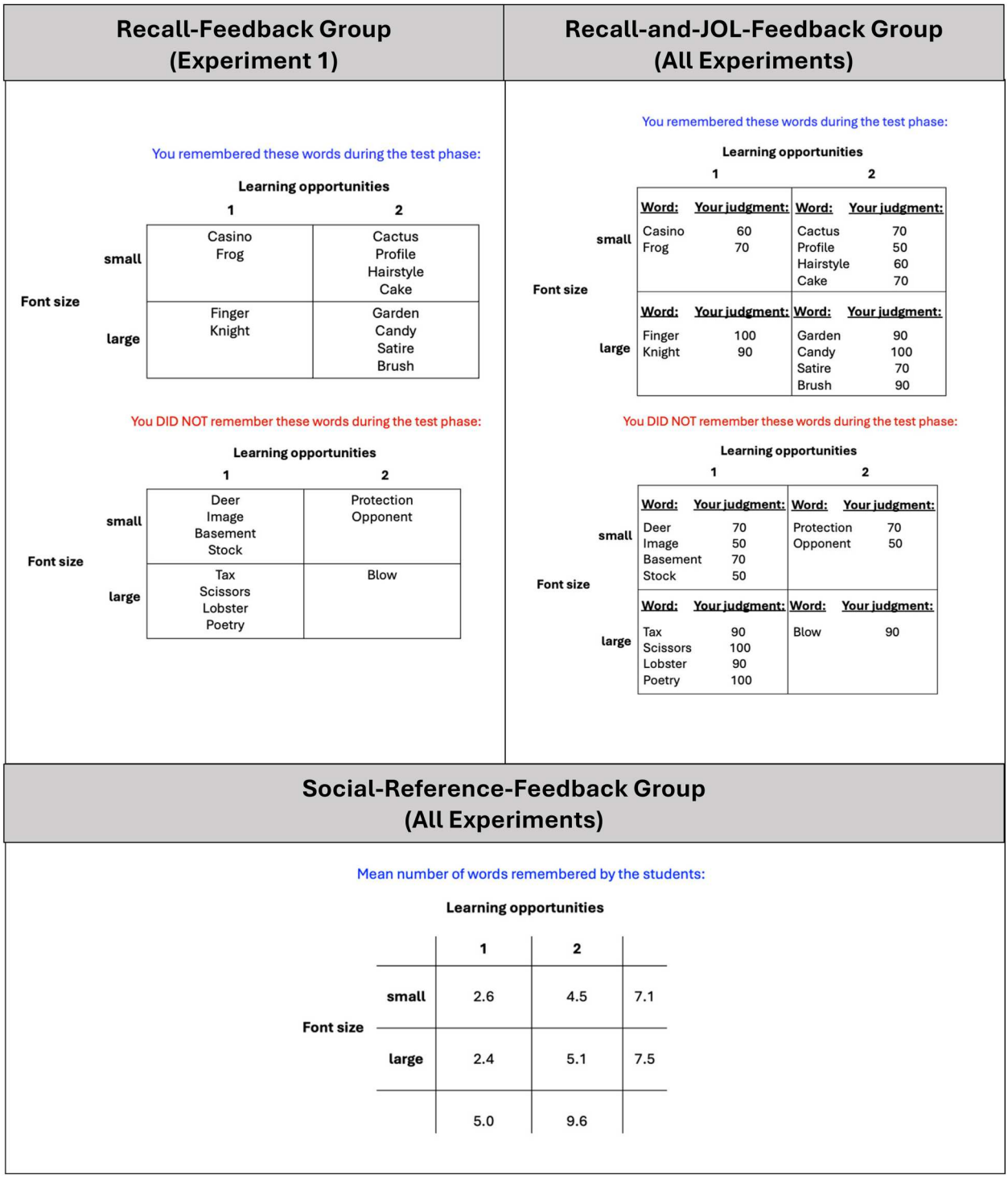

Figure 1. Example feedback presented to participants in the experimental groups. Note: The figure presents twelve or eleven exemplary items only. Participants in the experiment were presented with all the 40 items per study-test cycle.

JOLs

Figure 3 presents mean JOLs in each cycle by font size and number of study presentations in each group of Experiment 1. For words that were studied twice, the figure shows JOLs from the first study presentation. JOLs from the first study presentation were submitted to a mixed ANOVA with cycle (1, 2, 3), font size (18, 48 point), and number of study presentations (1, 2) as within-subjects factors and feedback group (no-feedback, recall-feedback, recall-and-JOL-feedback, social-reference-feedback) as a between-subjects factor.

A significant main effect of cycle revealed that JOLs decreased across cycles, $F(1.48, 225.22) = 10.33$, $p < .001$,

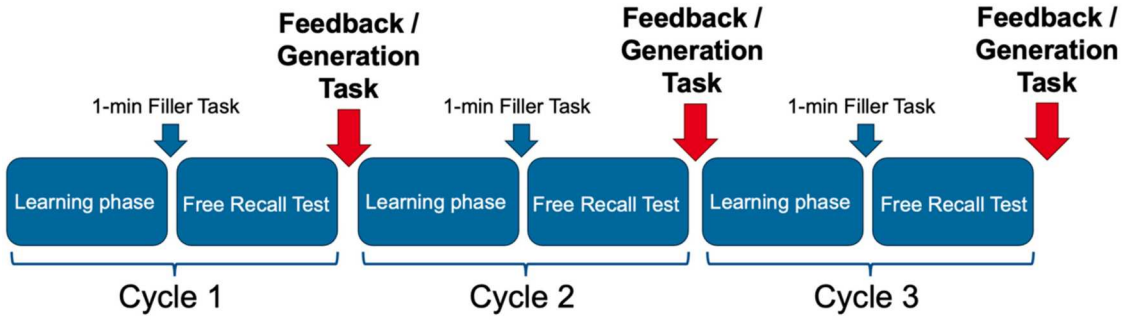

Figure 2. Representation of the procedure in all experiments. Note: Participants completed three study-test cycles with novel word lists in each cycle. Each cycle started with a learning phase in which study words were presented for 4 s each. Immediately after each word presentation, participants made a JOL on scale from 0% to 100% in increments of 10%. Afterwards, they completed a 1-min filler task where they worked on simple arithmetical equations. Then, a free recall test followed in which participants wrote as many recalled words as they could for 4 min. Finally, participants in the experimental groups received feedback (see Figure 1), whereas participants in the no-feedback group typed examples of one randomly chosen category (i.e., mountains, capitals, or rivers) for 3 min.

$\eta_p^2 = .06$. Pairwise follow-up t tests showed significant differences among all three cycles, $t(155) >= 2.07$, $p <= .04$, $d_z >= 0.17$. A significant main effect of font size revealed higher JOLs for words displayed in the large font than for words displayed in the small font, $F(1, 152) = 75.02$, $p < .001$, $\eta_p^2 = .33$. A significant main effect of number of study presentations revealed higher JOLs for twice-presented words than for once-presented words, $F(1, 152) = 3.96$, $p = .048$, $\eta_p^2 = .03$. The main effect of feedback group was not significant, $F(3, 152) = 1.31$, $p = .275$, $\eta_p^2 = .02$.

There were significant interactions between cycle and font size, $F(1.82, 276.24) = 6.25$, $p < .01$, $\eta_p^2 = .04$, and

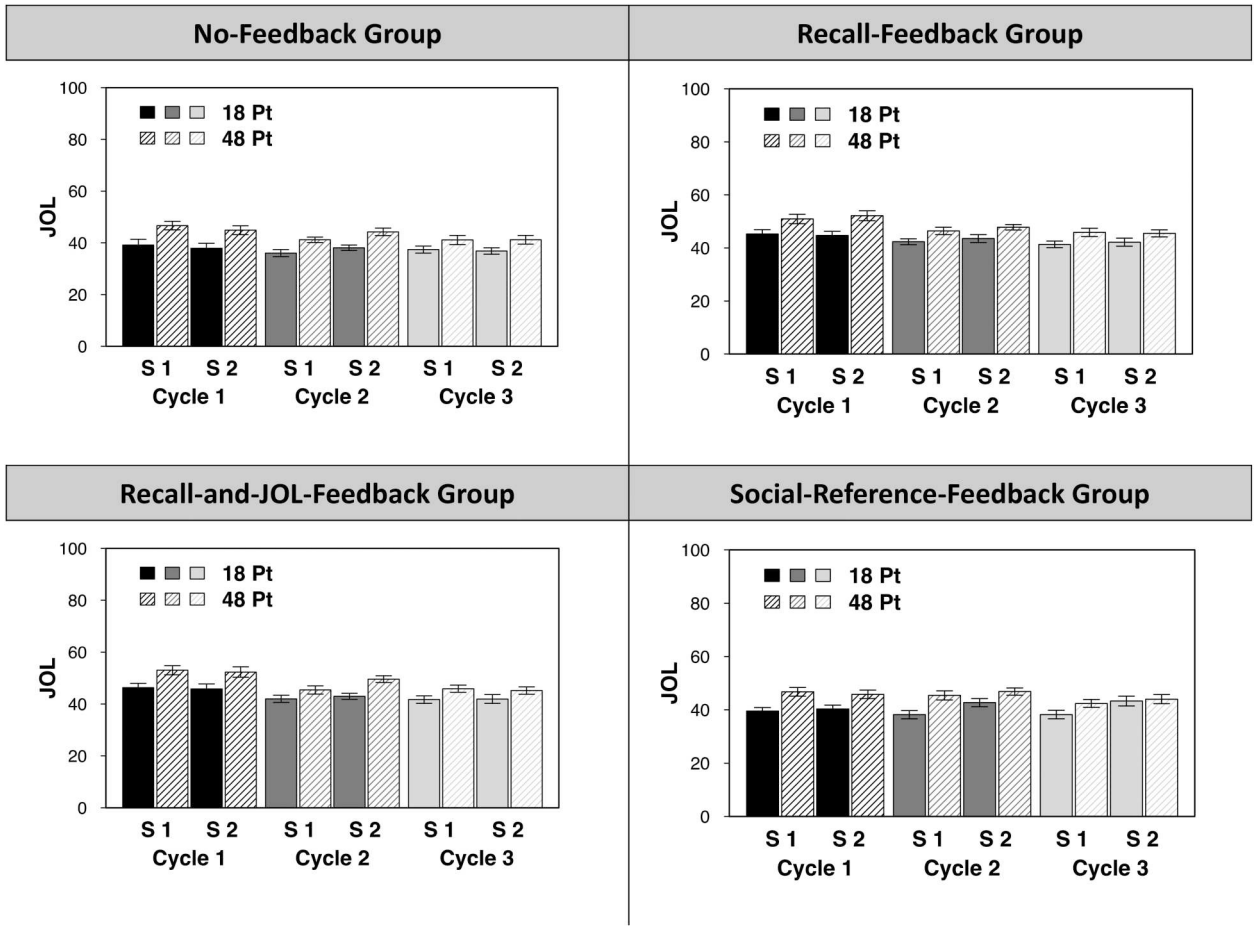

Figure 3. Mean judgments of learning (JOL) in each cycle for words presented once (S1) or twice (S2) in a small (18 pt) or a large (48 pt) font size in each group of Experiment 1. Note. Error bars represent one standard error of the mean. Note. Error bars represent one standard error of the mean.

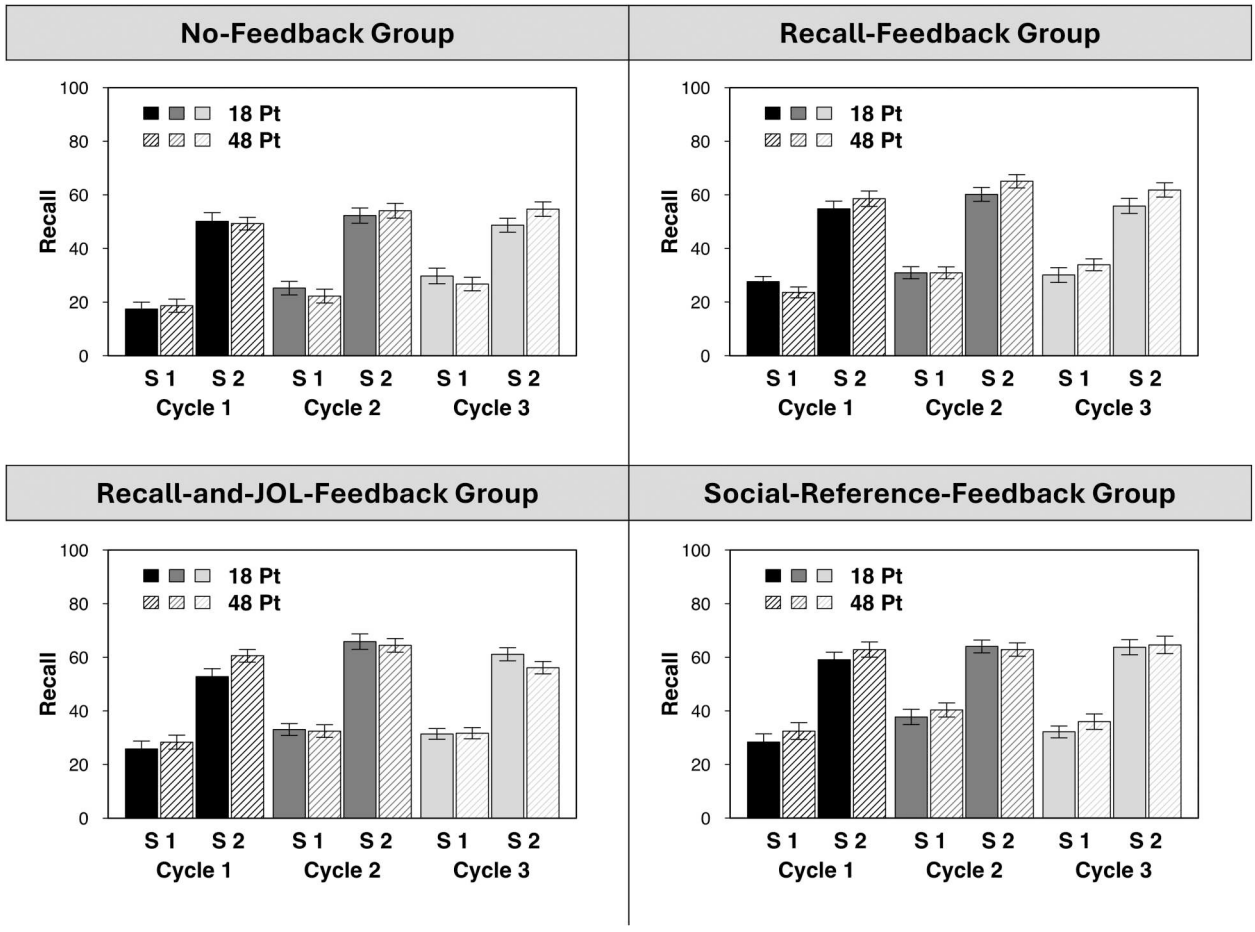

Figure 4. Mean percentage of correctly recalled words (recall) in each cycle for words presented once (S1) or twice (S2) in a small (18 point) or a large (48 point) font size in each group of Experiment 1. Note. Error bars represent one standard error of the mean.

between cycle and number of study presentations, $F(1.84, 280.12) = 5.88$, $p < .01$, $\eta_p^2 = .04$. Follow-up t tests indicated that the size of the font size effect on JOLs decreased across cycles, Cycle 1: $t(155) = 7.67$, $p < .001$, $d_z = 0.61$, Cycle 2: $t(155) = 6.78$, $p < .001$, $d_z = 0.54$, Cycle 3: $t(155) = 5.02$, $p < .001$, $d_z = 0.40$, and that JOLs were higher for twice-presented words than for once-presented words in Cycle 2, $t(155) = 3.69$, $p < .001$, $d_z = 0.30$, but not in Cycles 1 or 3, Cycle 1: $t < 1$, Cycle 3: $t(155) = 1.09$, $p = .276$, $d_z = 0.09$. None of the other interactions were significant, $F <= 1.71$, $p >= .168$.

Table 2. Means (SDs) of the Gamma Correlation between JOLs and Recall Performance in Each Cycle and Group of Experiments 1, 2, 3, and 4.

Experiment and group	Cycle 1	Cycle 2	Cycle 3
Experiment 1			
No-feedback	.33 (.23)	.30 (.34)	.26 (.37)
Recall-feedback	.21 (.27)	.36 (.26)	.29 (.28)
Recall-and JOL-feedback	.20 (.27)	.22 (.33)	.25 (.30)
Social-reference-feedback	.29 (.27)	.31 (.29)	.31 (.37)
Experiment 2			
No-feedback	.31 (.32)	.31 (.41)	.41 (.26)
Catch-all-cognitive-feedback	.26 (.26)	.30 (.29)	.23 (.37)
Experiment 3			
No-feedback	.21 (.25)	.36 (.23)	.30 (.28)
Cognitive-plus-metacognitive-feedback	.28 (.22)	.30 (.31)	.44 (.26)
Experiment 4			
No-feedback	.24 (.25)	.23 (.35)	.32 (.34)
Cognitive-plus-metacognitive-feedback	.24 (.24)	.29 (.28)	.35 (.39)
Metacognitive-only-feedback	.29 (.26)	.36 (.25)	.38 (.25)

Note: The "Recall-and-JOL-feedback" corresponds to cognitive feedback in Experiment 1.

Recall performance

Figure 4 presents the mean percentage of recalled words in each cycle by font size and number of study presentations in each group of Experiment 1. A 3 (cycle: 1, 2, 3) x 2 (font size: 18, 48 point) x 2 (number of study presentations: 1, 2) x 4 (feedback group: no-feedback, recall-feedback, recall-and-JOL-feedback, social-reference-feedback) mixed ANOVA on recall performance revealed that memory performance varied with cycle, $F(1.81, 275.36) = 12.85$, $p < .001$, $\eta_p^2 = .08$. Follow-up t tests showed worse memory performance in Cycle 1 than in Cycles 2 and 3, Cycle 1 vs. 2: $t(155) = 4.90$, $p < .001$, $d_z = 0.39$, Cycle 1 vs. 3: $t(155) = 3.23$, $p < .01$, $d_z = 0.26$, but no differences between Cycles 2 and 3, $t(155) = 1.53$, $p = .128$, $d_z = 0.12$. A small but significant main effect of font size revealed better memory

performance for words displayed in the large font than in the small font, $F(1, 152) = 5.51$, $p = .020$, $\eta_p^2 = .04$. A large significant main effect of number of study presentations revealed better memory performance for twice-presented words than for once-presented words, $F(1, 152) = 942.72$, $p < .001$, $\eta_p^2 = .86$. None of the other main effects or interactions were significant $F <= 2.88$, $p >= .058$.

Resolution

Gamma correlations

Gamma correlations could not be computed for three participants in one cycle due to perfect recall performance (n = 1 in the social-reference-feedback group in Cycle 2, n = 1 in the no-feedback group in Cycle 2, n = 1 in the recall-and-JOL-feedback group in Cycle 3).

Table 2 and Figure 5 present Gamma correlations in each cycle for each group of Experiment 1. Gamma correlations between JOLs and recall performance were significantly positive in all study-test cycles and groups, $t >= 4.21$, $p < .001$, $d >= 0.69$, indicating that participants from all feedback groups made JOLs with above chance resolution in all study-test cycles. A 3 (cycle) x 4 (feedback group) mixed ANOVA revealed no main effects of cycle, $F(1.99, 296.38) = 1.26$, $p = .286$, $\eta_p^2 = .01$, or group, $F(3, 149) = 0.90$, $p = .443$, $\eta_p^2 = .02$, and also no interaction, $F(5.97, 296.38) = 1.37$, $p = .228$, $\eta_p^2 = .03$.

Logistic mixed-effects models

To examine resolution in more depth, we conducted a logistic mixed-effects model analysis (Murayama et al., 2014). In the mixed-effects model analysis, we predicted

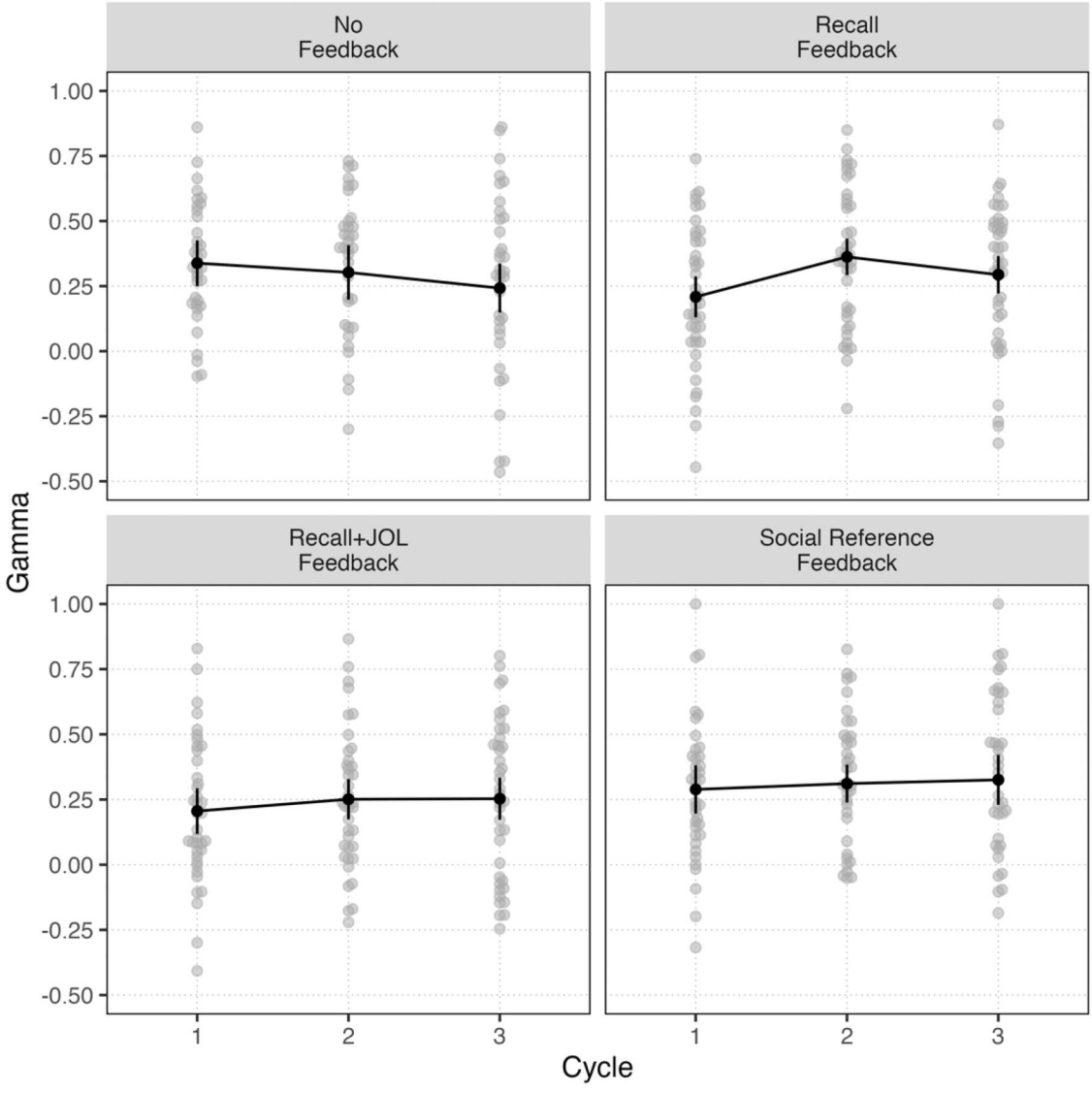

Figure 5. Mean gamma correlation between JOLs and recall performance in each cycle and group of Experiment 1. Note. Error bars represent one standard error of the mean.

recall performance (Recalled = 1, Not recalled = 0) from group, cycle, participant's mean-centered JOLs, and their interactions. To assess differences in accuracy between groups, we used a set of orthogonal contrasts. The first contrast tested the difference between the control group (No-feedback group; −3/4) and all three experimental groups (coded each as +1/4). The second contrast tested the difference between the social-reference-feedback group (+2/3) and the other two experimental groups (recall-feedback, recall-JOL-feedback; coded both as −1/3). The third contrast tested the difference between the recall-feedback group (−1/2) and the recall-JOL-feedback group (+1/2). We treated group, cycle, and JOLs as fixed effects predictors. We specified random intercepts for participants.

Results showed a significantly positive coefficient for JOLs, $b = 0.20$ ($SE = 0.01$), $z = 24.62$, $p < .001$, indicating that JOLs were predictive of recall performance. The interaction between JOLs and Cycle was also significant $b = 0.04$ ($SE = 0.01$), $z = 3.54$, $p < .001$, which indicates that the predictivity of JOLs increased across cycles. Furthermore, the significant JOLs, cycle, and first contrast interaction, $b = 0.06$ ($SE = 0.02$), $z = 2.48$, $p < .05$, indicated that this improvement was more pronounced in the experimental groups than in the control group. The main effect of cycle was also significant, $b = 0.14$ ($SE = 0.02$), $z = 6.54$, $p < .001$, corroborating the improvement in recall performance across cycles effect found in the mixed ANOVA analysis. Finally, there was a significantly positive coefficient for the first contrast, $b = 0.44$ ($SE = 0.14$), $z = 2.96$, $p < .01$, indicating better recall performance in the experimental groups than in the control group. No other effects were significant, $z <= 1.90$, $p >= .07$.

Calibration

A 3 (cycle) x 4 (feedback group) mixed ANOVA on calibration revealed that the difference between JOLs and recall performance (i.e., bias) varied with cycle, $F(1.77, 268.83) = 23.57$, $p < .001$, $\eta_p^2 = .13$, with pairwise comparisons indicating a switch from overconfidence in Cycle 1 ($M = 5.12$, $SD = 21.8$) to underconfidence in Cycle 2 ($M = -3.08$, $SD = 20.8$) and Cycle 3 ($M = -2.74$, $SD = 19.1$). Cycle 1 vs. 2: $t(155) = 5.93$, $p < .001$, $d = 0.47$, Cycle 1 vs. 3, $t(157) = 5.20$, $p < .001$, $d = 0.42$, Cycle 2 vs. 3: $t < 1$. Neither the main effect of feedback group, $F(3, 152) = 1.77$, $p = .155$, $\eta_p^2 = .03$, nor the interaction were significant, $F < 1$.

Discussion

Experiment 1 showed that neither type of feedback improved the cue basis of JOLs. Participants in all groups continued to overweight font size (main effect on JOLs across cycles; $\eta_p^2 = .33$; main effect on recall across cycles; $\eta_p^2 = .04$) and continued to underweight number of study presentations (main effect on JOLs across cycles; $\eta_p^2 = .03$; main effect on recall across cycles; $\eta_p^2 = .86$). Interestingly, we found that the effect of font size on JOLs decreased across cycles in all groups, indicating that participants may have learned from experience to rely less on font size.

Regarding resolution, we found no improvements in the Gamma correlation analysis. In the logistic mixed-effects model analysis, we found that the resolution of JOLs increased with cycle in all experimental groups compared to the control group. Inconsistent results in Gamma correlations and logistic models might arise because of the lower sensitivity of Gamma correlations due to discarding ties (Masson & Rotello, 2009; Spellman et al., 2014). While it will be important to replicate the effect found in the mixed-effect models analysis, it suggests that monitoring can improve from experience and feedback even when illusions persist. Furthermore, recall performance increased after the first cycle, probably because participants had a better idea of how to approach the task. In contrast, JOLs decreased after the first cycle. This produced a switch from overconfidence to underconfidence, a pattern that is well established for multiple study-test cycles using the same materials repeatedly and known as the underconfidence-with-practice effect (Koriat et al., 2002).

Experiment 2

Since feedback was not successful at improving the cue basis of JOLs in Experiment 1, Experiment 2 combined all forms of feedback in a "catch-all-cognitive-feedback" group to provide participants with maximum information. Participants in the catch-all-cognitive-feedback group saw a list of the words they had recalled and not recalled, accompanied by the JOL they had given to each word during the study, organised by the two cues (see recall-and-JOL-feedback group in Figure 1). This was followed by information about each cue's effectiveness and average performance of other participants in this same task (see social-reference-feedback group in Figure 1). If this extreme catch-all-cognitive-feedback does not improve cue use in JOLs and JOL accuracy, this will demonstrate the resistance of metacognitive judgments to change.

In Experiment 2, we manipulated font format (standard, aLtErnAtiNg) in addition to font size. Standard-font words typically elicit higher JOLs than alternating-font words, whereas there are typically no differences in recall performance between the two font formats (Rhodes & Castel, 2008; but see Mueller et al., 2013). We expected that manipulating font format instead of number of study presentations might facilitate the learning of predictive cue validities, because the two simultaneously manipulated cues are both perceptual in nature and are affecting JOLs despite having little or no impact on recall performance.

Method

Design

The design was a mixed design with font size (18, 48 point), font format (standard, aLtErnAtiNg), and study-

test cycle (1, 2, 3) as within-subjects factors and group (no-feedback, catch-all-cognitive-feedback) as between-subjects factor.

Materials

Stimuli were identical to Experiment 1. The only exception was that for each participant, one randomly selected half of the large- and small font words were displayed in standard or aLtErNaTiNg format instead of being presented once or twice for study. The first four items represented each combination of font size and font format and were used as buffer items that were not included in the analysis.

Participants

We aimed at recruiting at least N = 80 participants. The primary analysis was the three-way interaction among group, cycle, and cue [font size or font format] in a four-way ANOVA, so power calculations were identical to those reported in Experiment 1. We recruited 43 University of Mannheim undergraduates, 37 of which completed the study in the laboratory and 6 of which completed the study online. Due to the Covid-19 pandemic, we recruited 43 additional participants from the Prolific online pool (https://www.prolific.co). These participants were native-German speakers who were located in Germany, 18–35 years old, and mostly students (93.02%). Participants were randomly allocated to the control (n = 40), and feedback (n = 40) groups. We used the same exclusion criteria as in Experiment 1 and excluded participants who assigned the same JOL to all items in one or more study phases (n = 5) or who had zero recall performance in one or more test phases (n = 1). The final sample included 80 participants with a mean age of 23.3 years (SD = 5.19), n = 40 in the control group, and n = 40 in the feedback group.

Procedure

The procedure was identical to that of Experiment 1 with the following exceptions. All words were presented only once for study. Participants in the catch-all-cognitive-feedback group received first the same information as the recall-and-JOL-feedback group followed by the same information as the social-reference-feedback group in Experiment 1 (see Figure 1) with number of announced study presentations replaced by font format. The updated explanation stated that many students have overestimated the influences of font size and font format on their memory.

Results

JOLs

Figure 6 presents mean JOLs in each cycle by font format and size in each group of Experiment 2 in the upper row. JOLs were submitted to a mixed ANOVA with cycle (1 vs 2 vs 3), font size (18 vs 48 point), and font format (standard vs aLtErNaTiNg) as within-subjects factors and feedback group (no-feedback, catch-all-cognitive-feedback) as between-subjects factor.

A significant main effect of cycle revealed that JOLs decreased across cycles, F (1.42, 110.64) = 9.82, p < .001, η_p^2 = .11. Follow-up t tests showed larger JOLs in Cycle 1 than in Cycles 2 and 3, Cycle 1 vs. 2: t(79) = 2.95, p < .01, d_z = 0.33, Cycle 1 vs. 3: t(79) = 3.54, p < .001, d = 0.40, but no differences between Cycles 2 and 3, t(79) = 1.47, p = .146, d_z = 0.16. A significant main effect of font size revealed higher JOLs for words displayed in large font than for words displayed in small font, F (1, 78) = 13.78, p < .001, η_p^2 = .15, and a significant main effect of font format revealed higher JOLs for words displayed in standard format than for words displayed in alternating format, F (1, 78) = 23.60, p < .001, η_p^2 = .23. None of the other main effects or interactions were significant, F <= 3.17, p >= .08.

Recall performance

Figure 6 (lower row) presents the mean percentage of recalled words in each cycle by font format and font size in each group of Experiment 2. A 3 (cycle: 1, 2, 3) x 2 (font size: 18, 48 point) x 2 (font format: standard, aLtErNaTiNg) x 2 (feedback group: no-feedback, catch-all-cognitive-feedback) mixed ANOVA on recall performance revealed that memory performance varied with cycle, F (1.88, 146.85) = 19.35, p < .001, η_p^2 = .20. Follow-up t tests showed worse memory performance in Cycle 1 than in Cycles 2 and 3, Cycle 1 vs. 2: t(79) = 5.29, p < .001, d_z = 0.59, Cycle 1 vs. 3: t(79) = 4.84, p < .001, d_z = 0.54, but no differences between Cycles 2 and 3, t < 1. Neither the main effect of font size, F(1, 78) = 2.98, p = .09, η_p^2 = .04, nor the main effect of font format were significant, F(1, 78) = 1.84, p = .18, η_p^2 = .02. None of the other main effects or interactions were significant, F <= 2.18, p >= .12.

Resolution

Gamma correlations

Table 2 and Figure 7 present Gamma correlations in each cycle for each group of Experiment 2. As in Experiment 1, all Gamma correlations by group and cycle were significantly positive t >= 4.03, p < .001, d >= .64, indicating that participants from both groups made JOLs with above chance resolution in all study-test cycles. A 3 (cycle) x 2 (group) mixed ANOVA revealed no main effects of cycle, F < 1, or group, F(1,78) = 2.45, p = .121, η_p^2 = .03, and also no interaction, F(1.98, 154.76) = 1.85, p = .161, η_p^2 = .02.

Logistic mixed-effects models

In the mixed-effects model analysis, we predicted recall performance (Recalled = 1, Not recalled = 0) from group (No-Feedback Group = -1, Catch-All-Cognitive-Feedback Group = 1), cycle, participant's mean-centered JOLs, and

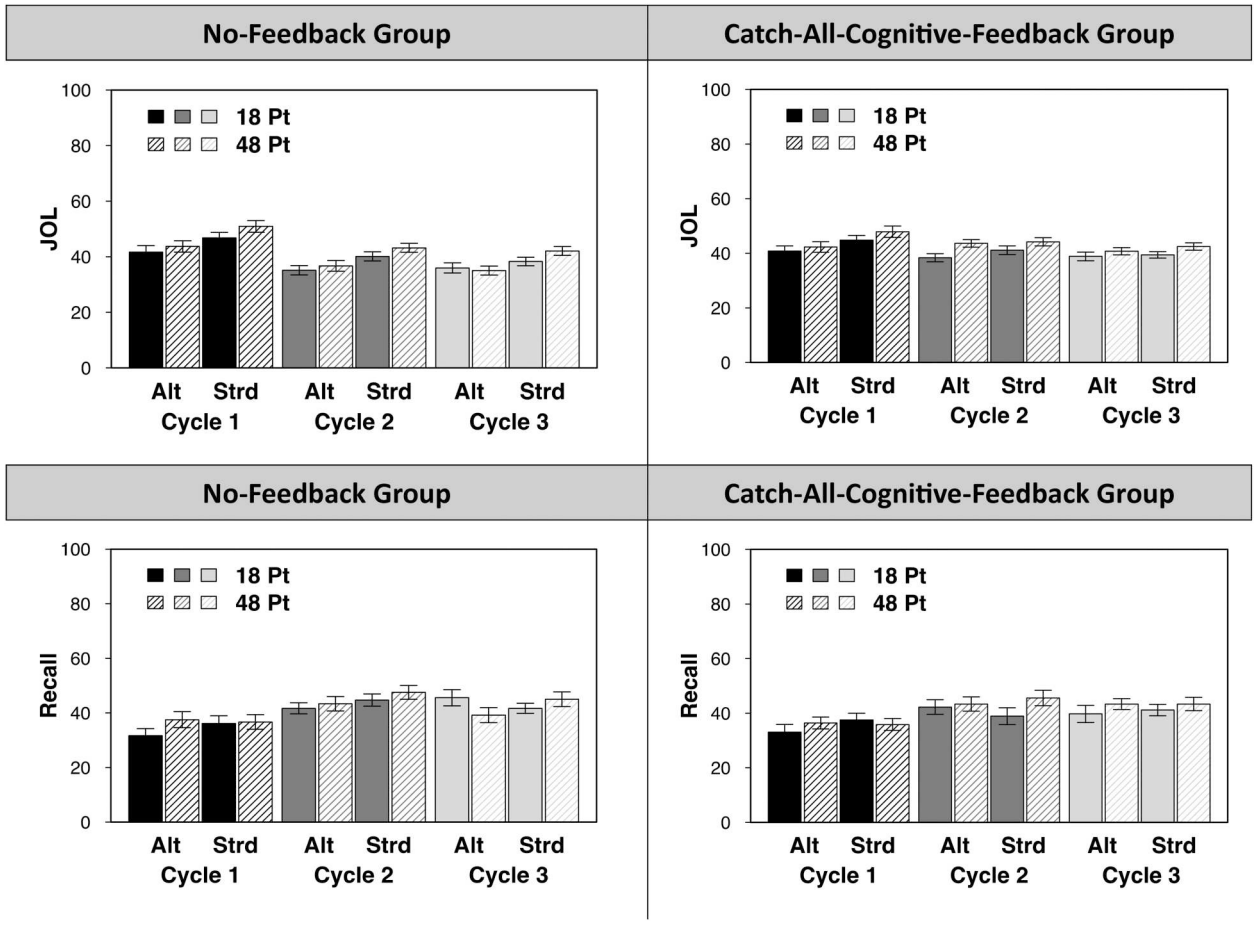

Figure 6. Mean judgments of learning (JOL) and percentage of correctly recalled words (Recall) in each cycle for words presented in alternating (Alt) or standard (Strd) font and in a small (18 pt) or a large (48 pt) font size in each group of Experiment 2. Note. Error bars represent one standard error of the mean.

their interactions. We treated group, cycle, and JOLs as fixed effects predictors and specified random intercepts for participants.

Results showed a significantly positive coefficient for JOLs, b = 0.23 (SE = 0.01), z = 19.48, p < .001, indicating that JOLs were predictive of recall performance. The interaction between JOLs and cycle was also significant, b = 0.05 (SE = 0.01), z = 3.61, p < .001, which indicates that the predictivity of JOLs increased across cycles. There was also a significantly negative coefficient for the interaction between group and JOLs, b = -0.03 (SE = 0.01), z = -2.85, p < .01, suggesting that the relation between JOLs and recall was stronger in the control than in the experimental group. The main effect of cycle was also significant, b = 0.23 (SE = 0.03), z = 7.61, p < .001, corroborating the improvement in recall performance across cycles found in the mixed ANOVA analysis. No other effects were significant, z <= 1.54, p >= .12.

Calibration

A 3 (cycle) x 2 (group) mixed ANOVA on calibration revealed that bias varied with cycle, F(1.64, 127.67) = 30.13, p < .001, η_p^2 = .28, with pairwise comparisons indicating a switch from overconfidence in Cycle 1 (M = 9.25, SD = 23.7) to underconfidence in Cycle 2 (M = -3.11, SD = 18.2) and Cycle 3 (M = -3.27, SD = 17.5). Cycle 1 vs. Cycle 2: t(79) = 5.97, p < .001, d_z = 0.67, Cycle 1 vs. 3: t(79) = 6.02, p < .001, d_z = 0.67, Cycle 2 vs. 3: t < 1. Neither the main effect of group, F < 1, nor the interaction were significant, F(1.64, 127.67) = 1.94, p = .156, η_p^2 = .02.

Discussion

Experiment 2 showed no improvements in cue use in the catch-all-cognitive-feedback group compared to the control group. Participants continued to overweight both font size (main effect on JOLs across cycles; η_p^2 = .15; main effect on recall across cycles; η_p^2 = .04) and font format (main effect on JOLs across cycles; η_p^2 = .23; main effect on recall across cycles; η_p^2 = .02) despite receiving maximum information (i.e., tables displaying JOLs and recall status for each word and average performance of other participants). Unlike Experiment 1, Experiment 2 showed that the illusory effect of font size on JOLs remained stable across cycles.

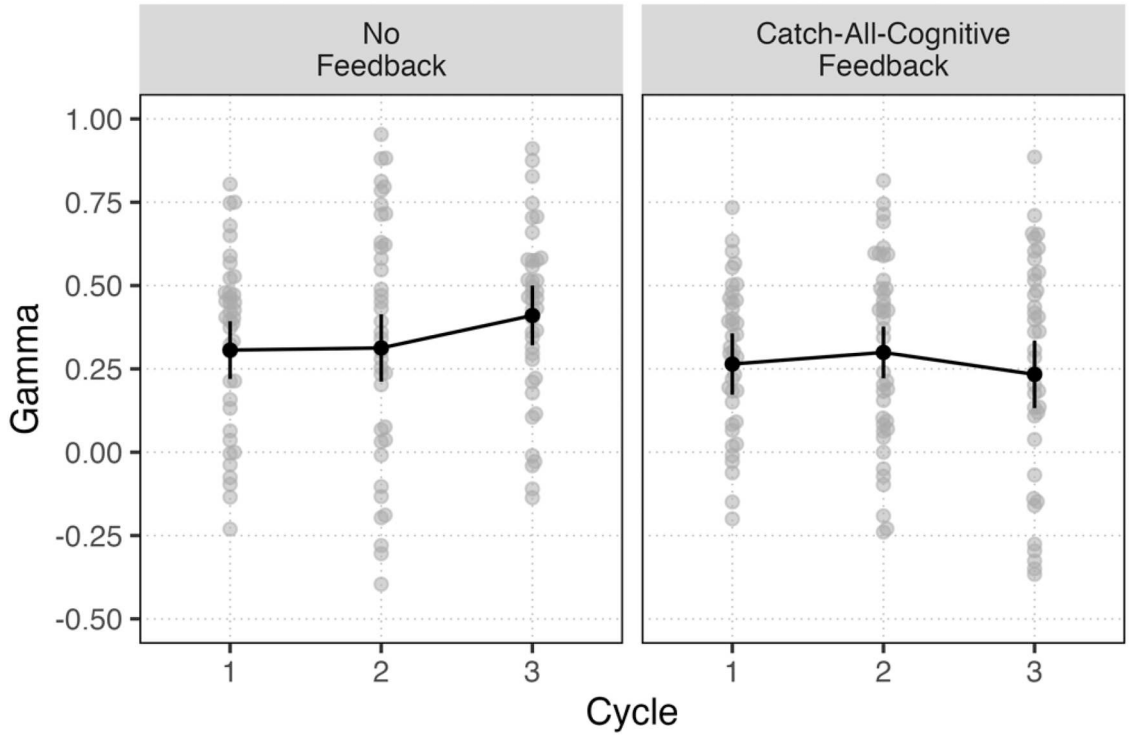

Figure 7. Mean gamma correlation between JOLs and recall performance in each cycle and group of Experiment 2. Note: Error bars represent one standard error of the mean.

Regarding resolution, the gamma correlation analysis showed no improvement across cycles. However, like in Experiment 1, the logistic mixed-effects model analysis showed that JOLs resolution increased with cycle, again suggesting that task experience was helpful for monitoring. Finally, like in Experiment 1, JOLs switched from overconfidence to underconfidence after Cycle 1.

One may wonder why participants did not change the cue basis of their JOLs with cognitive feedback (i.e., comparison of JOL and recall status for each word categorised by cues) and average performance of other students including a warning that many of them overestimated cue effects. One reason could be that JOLs were largely based on a single experiential cue (e.g., fluency) rather than on the two manipulated cues. However, this possibility is at odds with evidence showing that people base their JOLs on multiple cues (Undorf et al., 2018; Undorf & Bröder, 2020). Further, other studies have shown that beliefs play a larger role in the font size effect on JOLs than fluency (Luna, Nogueira, et al., 2019; Mueller et al., 2014; Undorf & Zimdahl, 2019). A more promising possibility is based on research by Yan et al. (2016), who argue that metacognitive judgments are hard to change because of (a) pre-existing beliefs about learning and memory, (b) experiences of fluency during learning, and (c) the belief of being unique as a learner. Thus, it might be that participants' pre-existing beliefs influence how they interpret and store the cognitive feedback provided (Cooper & Vallée-Tourangeau, 2021). In some situations, participants might even disregard the cognitive feedback because they consider themselves experts on their own cognitions or think that the average performance of other participants is irrelevant for them as unique learners (see Yan et al., 2016). Experiment 3 addressed this possibility.

Experiment 3

In Experiment 3, we designed a new form of feedback to remedy the possibility that participants might misperceive cognitive feedback due to pre-existing beliefs. For that purpose, we followed Fiedler et al.'s (2020) recommendations for effective forms of feedback. As explained in the introduction, these recommendations state that an effective debiasing treatment should not only provide information about judgments that are "correct" versus "incorrect", but also relate feedback to (a) the representation of the stimuli, and (b) provide explicit instructions about how to make accurate judgments. Accordingly, we designed an informative "metacognitive-feedback" that adopted a first-person perspective and explained the metacognitions likely occurring during learning (e.g., large-font words stand out more and therefore are more memorable), highlighted their biased nature, and provided instructions about which cues to consider when making JOLs. Our aim was to specifically tackle participants' metacognitions to dismantle erroneous pre-existing beliefs.

Further, Experiment 3 aimed to ensure that participants fully attended the cognitive feedback and achieved a deep understanding of cue validities and biased metacognitions during learning. To achieve this goal, we used an approach similar to Pan and Rivers (2023) and asked participants to describe how each of the cues affected their memory and their JOLs after receiving feedback.

In Experiment 3, we manipulated again the announced number of study presentations (1, 2) as in Experiment 1. This allowed us to test whether the reduction of the font size effect across cycles is related to the simultaneous manipulation of a valid cue in comparison to an invalid cue such as font format.

Method

Design

The design was a mixed design with font size (18, 48 point), number of study presentations (1,2), and study-test cycle (1, 2, 3) as within-subjects factors and group (no-feedback, cognitive-plus-metacognitive-feedback) as between-subjects factor.

Materials

Stimuli were identical to Experiment 1.

Participants

We aimed at recruiting at least N = 80 participants. The primary analysis was the three-way interaction among group, cycle, and cue [number of study presentations or font size] in a four-way ANOVA, so power calculations were identical to those reported in Experiment 1. We recruited 23 University of Mannheim undergraduates and 57 Technical University of Darmstadt undergraduates. Participants were randomly allocated to the no-feedback (n = 40), and cognitive-plus-metacognitive-feedback (n = 40) groups. We used the same exclusion criteria as in Experiment 1 and excluded participants who assigned the same JOL to all items in one or more study phases (n = 1) or who had zero recall performance in one or more test phases (n = 0). Additionally, we excluded incomplete data due to a technical error (n = 2). The final sample included 77 participants with a mean age of 21.92 years (SD = 2.92), n = 39 in the control group, and n = 38 in the feedback group.

Procedure

The procedure was identical to that of Experiment 2 with the following exceptions. All participants completed the experiment in the laboratory. All words were presented in a standard format at study. For each participant, one randomly selected half of the small-font (18-point Arial font) and large-font (48-point Arial font) words were presented once or twice for study. To verify that participants in the feedback group understood the feedback on the words they had remembered and had not remembered with JOLs (see recall-JOL-feedback group in Figure 1), they answered two questions about their recall performance and memory predictions for each cue (see Table 3) immediately after receiving the table with item-by-item JOL and recall performance in each feedback phase.

Table 3. Understanding Feedback Questions in Experiments 3 and 4.

	Recall Question	Prediction Question
	"Which statement best describes your actual memory performance in Part 1 [2, 3]? In the test, I was able to … "	"Which statement best describes your memory predictions in Part 1 [2, 3]?"
Answer Option 1	remember more words learned twice [in large font size] than words learned once [in small font size].	I underestimated the impact of an additional learning opportunity [of a large font size] on my memory.
Answer Option 2	remember fewer words learned twice [in large font size] than words learned once [in small font size].	I correctly assessed the influence of an additional learning opportunity [of a large font size] on my memory.
Answer Option 3	remember the same number of words learned twice [in large font size] and once [in small font size].	I overestimated the influence of an additional learning opportunity [of large font size] on my memory.

Then, participants were presented with the table displaying the average memory performance of other participants (see social-reference-feedback group in Figure 1). Finally, they read textual information about illusory metacognitions regarding font size and number of study presentations (see Appendix 1 for the exact wording).

Results

JOLs

Figure 8 presents mean JOLs in each cycle by font size and number of study presentations in each group of Experiment 3. For words that were studied twice, the figure shows JOLs from the first study presentation. JOLs from the first study presentation were submitted to a mixed ANOVA with cycle (1, 2, 3), font size (18, 48 point), and number of study presentations (1, 2) as within-subjects factors and feedback group (no-feedback, cognitive-plus-metacognitive-feedback) as a between-subjects factor.

A significant main effect of cycle revealed differences in JOLs across cycles, $F(1.57, 117.99) = 6.90$, $p < .01$, $\eta_p^2 = .08$. Specifically, JOLs were lower in Cycles 2 and 3 than in Cycle 1, Cycle 1 vs. Cycle 2: $t(76) = 2.51$, $p = .01$, $d_z = 0.29$, Cycle 1 vs. 3: $t(76) = 3.01$, $p < .01$, $d_z = 0.34$, but did not differ between Cycle 2 and 3, $t(76) = 1.66$, $p = .10$, $d_z = 0.19$. A significant main effect of font size revealed higher JOLs for words displayed in the large font than for words displayed in the small font, $F(1, 75) = 39.05$, p

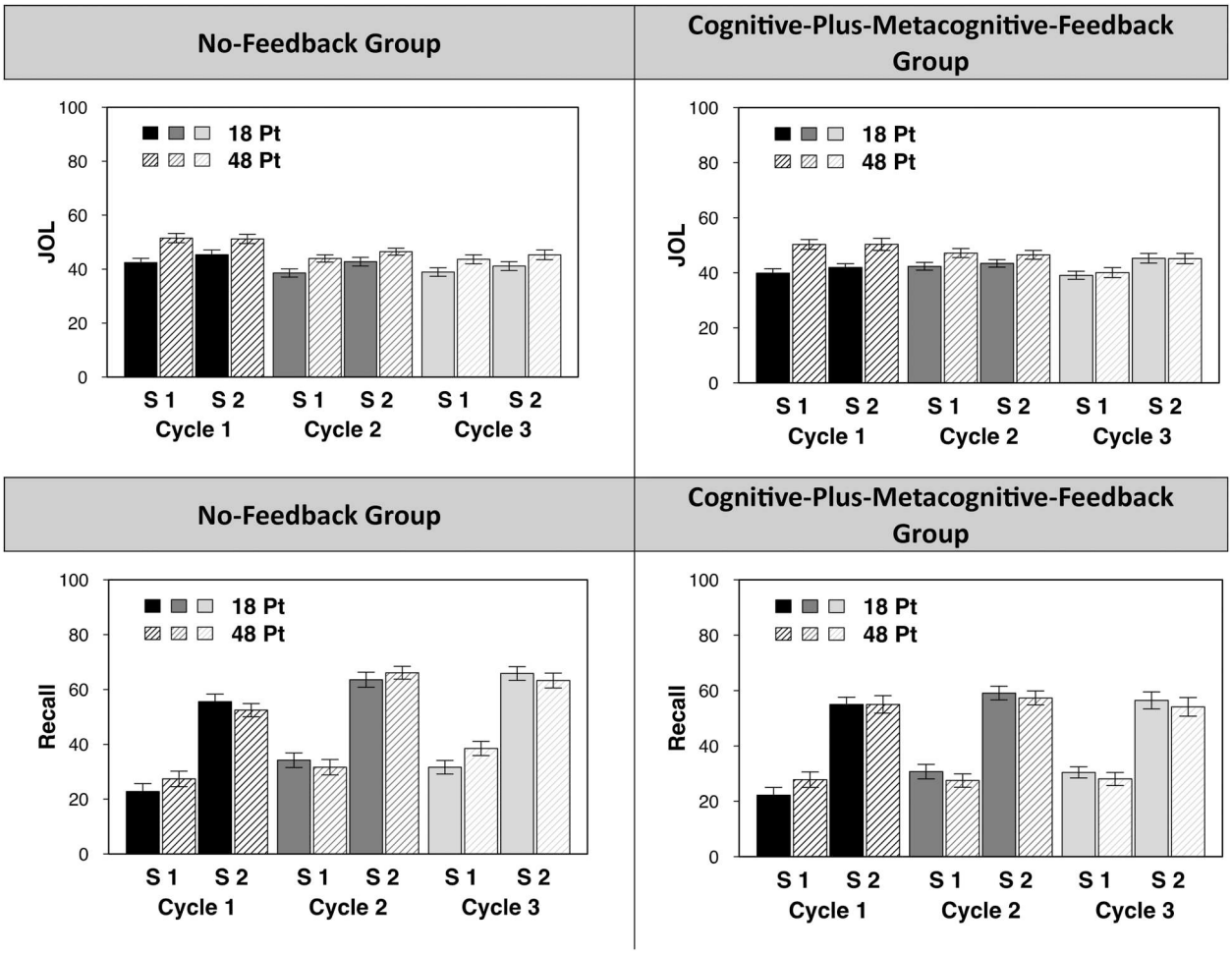

Figure 8. Mean judgments of learning (JOL) and percentage of correctly recalled words (Recall) in each cycle for words presented once (S1) or twice (S2) in a small (18 pt) or a large (48 pt) font size in each group of Experiment 3. Note. Error bars represent one standard error of the mean.

< .001, $\eta_p^2 = .34$. A significant main effect of number of study presentations revealed higher JOLs for twice-presented words than for once-presented words, F(1, 75) = 8.40, p < .01, $\eta_p^2 = .10$. The main effect of group was not significant, F < 1.

As in Experiment 1, there was a significant interaction between cycle and font size, F(1.68, 125.65) = 10.06, p < .001, $\eta_p^2 = .12$. Follow-up t tests indicated that the size of the font size effect on JOLs decreased across cycles, Cycle 1: t(76) = 6.47, p < .001, $d_z = 0.74$, Cycle 2: t(76) = 4.20, p < .001, $d_z = 0.48$, Cycle 3: t(76) = 2.29, p = .02, $d_z = 0.26$. Importantly, the three-way interaction between group, cycle, and number of study presentations was significant, F(1.99, 149.17) = 3.07, p = .0495, $\eta_p^2 = .04$. Follow-up analyses showed that the interaction between cycle and number of study presentations was not significant in the control group, F < 1, but was significant in the cognitive-plus-metacognitive-feedback group, F(1.93, 71.48) = 4.07, p = .02, $\eta_p^2 = .10$, with t tests showing that JOLs were higher for twice– than for once-studied words in Cycle 3, t(37) = 2.91, p < .01, $d_z = 0.47$, but not in Cycles 1 and 2, Cycle 1: t < 1, Cycle 2: t < 1. The three-way interaction between group, cycle and font size was not significant, F(1.68, 125.65) = 2.39, p = .10, $\eta_p^2 = .03$. None of the other interactions were significant, F <= 3.69, p >= .06.

Recall performance

Figure 8 presents the mean percentage of recalled words in each cycle by font size and number of study presentations in each group of Experiment 3. A 3 (cycle: 1, 2, 3) x 2 (font size: 18, 48 point) x 2 (number of study presentations: 1, 2) x 2 (group: no-feedback, cognitive-plus-metacognitive-feedback) mixed ANOVA on recall performance revealed that memory performance varied with cycle, $F(1.91, 143.50) = 8.84$, $p < .001$, $\eta_p^2 = .11$. Follow-up t tests showed worse memory performance in Cycle 1 than in Cycles 2 and 3, Cycle 1 vs. 2: $t(76) = 3.59$, $p < .001$, $d_z = 0.41$, Cycle 1 vs. 3: $t(76) = 3.29$, $p < .01$, $d_z = 0.38$, but no differences between Cycles 2 and 3, $t < 1$. A significant main effect of number of study presentations revealed better memory performance for twice-presented words than for once-presented words, $F(1, 75) = 666.88$, $p < .001$, $\eta_p^2 = .90$. None of the other main effects or interactions were significant $F <= 2.62$, $p >= .080$.

Resolution

Gamma correlations

Table 2 and Figure 9 present Gamma correlations in each cycle for each group of Experiment 3. As in Experiments 1 and 2, all Gamma correlations by group and cycle were significantly positive, $t >= 5.24$, $p < .001$, $d >= .84$, indicating that participants from both groups made JOLs with above chance resolution in all study-test cycles. A 3 (cycle) x 2 (group) mixed ANOVA revealed a main effect of cycle, $F(1.94, 145.76) = 5.51$, $p < .01$, $\eta_p^2 = .07$, no main effect of group, $F(1,75) = 1.46$, $p = .230$, $\eta_p^2 = .02$, and a significant interaction, $F(1.94, 145.76) = 3.57$, $p = .032$, $\eta_p^2 = .05$. In the cognitive-plus-metacognitive-feedback group, resolution was improved in Cycle 3, but did not differ between Cycles 1 and 2, Cycle 1 vs. 2: $t < 1$, Cycle 1 vs. 3: $t(37) = 2.84$, $p < .001$, $d_z = 0.46$, Cycle 2 vs. 3: $t(37) = 2.45$, $p = .019$, $d_z = 0.40$. In contrast, in the control group, resolution improved between Cycle 1 and 2, $t(38) = 2.90$, $p < .01$, $d_z = 0.46$, but it was similar in Cycles 1 and 3 and Cycles 2 and 3, Cycles 1 vs. 3: $t(38) = 1.51$, $p = .140$, $d_z = 0.24$, Cycles 2 vs. 3: $t(38) = 1.23$, $p = .224$, $d_z = 0.20$.

Logistic mixed-effects models

In the mixed-effects model analysis, we predicted recall performance (Recalled = 1, Not recalled = 0) from group (No-Feedback Group = −1, Cognitive-Plus-Metacognitive-Feedback group = 1), cycle, participant's mean-centered JOLs, and their interactions. We treated group, cycle, and JOLs as fixed effects predictors. We specified random intercepts for participants.

Results showed a significantly positive coefficient for JOLs, $b = 0.21$ ($SE = 0.01$), $z = 18.67$, $p < .001$, indicating that JOLs were predictive of recall performance. The interaction between JOLs and Cycle was also significant $b = 0.04$ ($SE = 0.01$), $z = 2.94$, $p < .01$, which indicates that the predictivity of JOLs increased across cycles. The main effect of cycle was also significant, $b = 0.19$ ($SE = 0.03$), $z = 6.37$, $p < .001$, corroborating the improvement in recall performance across cycles found in the mixed ANOVA analysis. There was also a significantly negative coefficient for the interaction between group and cycle, $b = -0.09$ ($SE = 0.03$), $z = -3.17$, $p < .01$, indicating less improvement in recall across cycles in the control group than in the cognitive-plus-metacognitive-feedback group. No other effects were significant, $z <= 1.20$, $p >= .23$.

Calibration

A 3(cycle) x 2(group) mixed ANOVA on calibration revealed a main effect of cycle, $F(1.82, 136.34) = 17.69$, $p < .001$, $\eta_p^2 = .19$, no main effect of group, $F(1, 75) = 1.13$, $p = .290$, $\eta_p^2 = .02$, and a significant interaction, $F(1.82, 136.34) = 4.38$, $p = .017$, $\eta_p^2 = .06$. In the control group, there was a switch from overconfidence in Cycle 1 ($M = 8.08$, $SD = 21.3$) to underconfidence in Cycle 2 ($M = -5.94$, $SD = 16.5$) and 3 ($M = -7.54$, $SD = 18.7$). Cycles 1 vs. 2: $t(38) = 4.82$, $p < .001$, $d_z = 0.77$, Cycles 1 vs. 3: $t(38) = 4.87$, $p < .001$, $d_z = 0.78$,

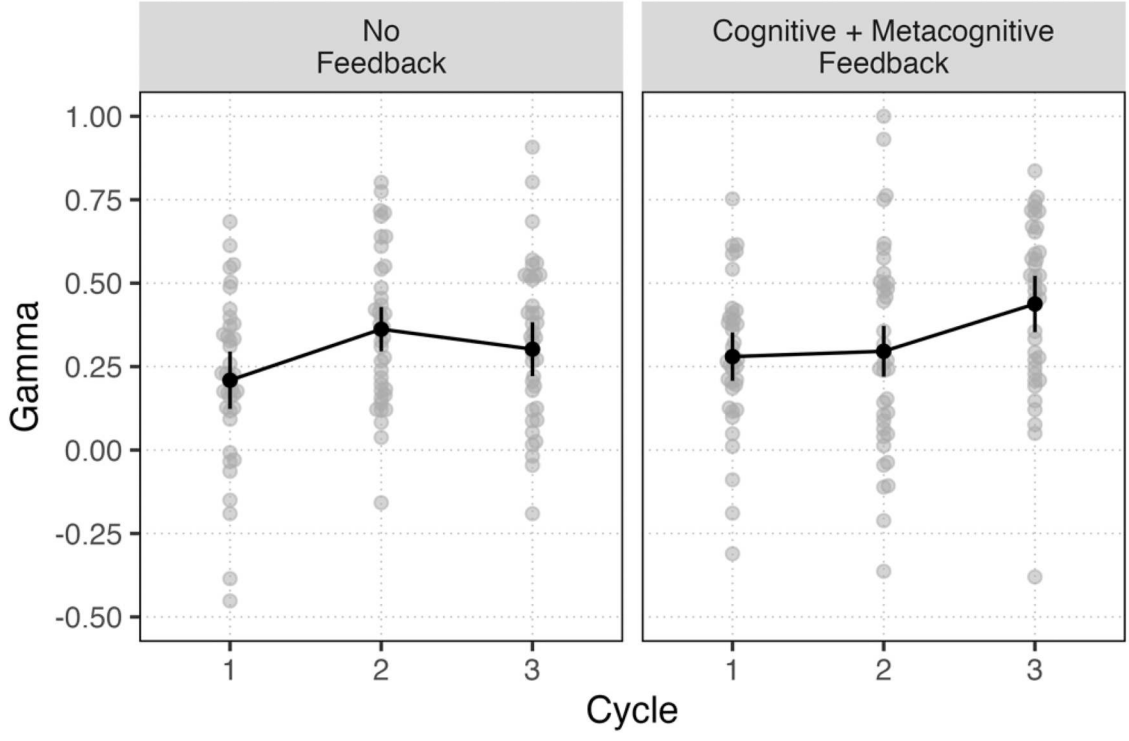

Figure 9. Mean gamma correlation between JOLs and recall performance in each cycle and group of Experiment 3. Note: Error bars represent one standard error of the mean

Cycles 2 vs. 3: $t < 1$. In the cognitive-plus-metacognitive-feedback group, calibration did not differ across cycles, but there was a non-significant trend towards a reduction in bias across cycles (Cycle 1: $M = 5.60$, $SD = 16.1$, Cycle 2: $M = 1.21$, $SD = 22.7$, Cycle 3: $M = 0.12$, $SD = 21.2$), $t(37) >= 1.99$, $p >= .053$, $d_z <= 0.32$.

Questionnaire data

Table 4 shows participants' responses to the questions after the item-by-item feedback in each cycle (see Table 3). It reveals three notable results. First, most participants correctly indicated that recall performance was greater for twice-presented than once-presented words in all cycles. Second, there seems to be a shift towards indicating that recall performance was equal for large and small font size words after Cycle 1. Third, participants' confidence in the correctness of their JOLs increased across cycles.

We separately analysed JOLs from each cycle for two subgroups of participants from the cognitive-plus-metacognitive-feedback group. The first subgroup consisted of participants who correctly indicated that recall performance was greater for twice-presented than once-presented words. The second subgroup consisted of participants who correctly indicated that recall performance was similar for large and small words. Comparing JOLs from each subgroup of participants with those from the control group revealed no interaction between group and the font size cue, indicating that the manipulation did not work for them. Further, the subgroup analyses did not reveal any evidence for the increased reliance on the number of study presentations in the cognitive-plus-metacognitive-feedback group that was observed in the main analysis. We suspect that this discrepancy may be due to reduced statistical power resulting from participant exclusions (for details, see Supplementary Material 2).

Discussion

Experiment 3 suggested that adding metacognitive feedback (i.e., written information about illusory metacognition) after the cognitive feedback (i.e., table displaying JOLs and recall status for each word) and social reference information effectively increased participants' reliance on the number of study presentations when making JOLs. Although participants continued to underestimate the influence of number-of-study-presentations cue, they learned that the announced number of study presentations is a valid predictor of memory performance and relied more on it in Cycle 3 than in Cycles 1 and 2. The questionnaire data show that most participants correctly recognised better memory performance for words studied twice compared to once. This positive effect of the number-of-study-presentations cue on JOLs is likely due to the provision of metacognitive feedback that informed participants about the illusory nature of their metacognitions. Alternatively, it might be possible that the questions aimed to evaluate whether the cognitive feedback was understood, included in Experiment 3 but not in Experiment 1 and 2, may have supported participants to make accurate inferences from the cognitive feedback.

Regarding the font size cue, there was still overestimation of its effect on memory (main effect on JOLs across cycles; $\eta_p^2 = .34$; main effect on recall across cycles; $\eta_p^2 < .001$). Although font size had a descriptively weaker influence on JOLs in Cycle 2 – and even more so in Cycle 3 – in the feedback group than in the no-feedback group, the three-way interaction among font size, group, and cycle was non-significant. Nevertheless, as in Experiment 1 and unlike in Experiment 2, we found that the font size effect on JOLs decreased across cycles in both groups.

Regarding resolution, Gamma correlations had different trajectories between groups. In the control group, gamma improved from Cycle 1 to Cycle 2 and remained stable in Cycle 3. In the feedback group, it was similar across

Table 4. Percentage of participants from the cognitive-plus-metacognitive-feedback group who reported (correctly) each of response options for the recall and prediction question in each cycle of experiment 3.

		Recall Question			Prediction Question		
		Response Options					
		twice [large] > once [small]	twice [large] < once [small]	twice [large] = once [small]	Under-estim. of twice [large]	Correct estimation	Over-estim. of twice [large]
Cycle 1	# Study Present.	78.95 (73.68)	13.16 (0)	7.89 (0)	42.11 (26.32)	31.58 (18.42)	26.31 (10.53)
	Font Size	47.37 (42.10)	15.79 (13.16)	36.84 (7.89)	34.21 (2.63)	42.11 (15.79)	23.68 (13.16)
Cycle 2	# Study Present.	81.58 (73.68)	7.89 (2.63)	10.53 (5.26)	23.68 (21.05)	55.26 (26.32)	21.05 (10.53)
	Font Size	21.05 (13.16)	31.58 (26.31)	47.37 (7.89)	13.16 (5.26)	63.16 (34.21)	23.68 (15.79)
Cycle 3	# Study Present.	76.31 (57.89)	13.16 (2.63)	10.53 (2.63)	31.58 (10.53)	52.63 (28.95)	15.79 (13.16)
	Font Size	28.95 (15.79)	26.31 (15.79)	44.74 (7.89)	15.79 (5.26)	57.89 (36.84)	26.32 (18.42)

Notes: Correct percentages for the recall questions were determined by comparing participants' responses with their actual recall performance. For example, if a participant stated that they remembered twice-presented words more often than once-presented words, and their recall performance confirmed this pattern, the response was scored as correct. For the prediction questions, correctness was assessed by comparing the mean JOL difference between the two levels of a factor with the mean recall performance difference. If the mean JOL difference exceeded the mean recall difference, the response was classified as overestimation; if it was smaller, as underestimation; and if the two values matched, as correct estimation.

Cycles 1 and 2, followed by an improvement in Cycle 3. In contrast, the logistic mixed-effects model analysis did not show differences in resolution among groups. Both analyses revealed an improvement in resolution across cycles indicating that task experience is beneficial for resolution. Finally, we found positive effects of feedback on calibration: Unlike in the control group, JOLs in the feedback group did not switch from overconfidence to underconfidence but approached zero, although the changes did not reach significance.

The improvements found on cue utilisation and resolution in Experiment 3's feedback group will be more closely examined in Experiment 4.

Experiment 4

Experiment 3 found that metacognitive feedback (i.e., written information about illusory metacognition) after the cognitive feedback (i.e., tables displaying JOL and recall status of each word) improved the cue basis of JOLs. Specifically, the influence of the valid cue number of study presentations on JOLs increased with feedback. This improvement in cue utilisation may result from participants incorporating the cognitive feedback more readily and, consequently, adjusting their JOLs. At the same time, it may be that written information about illusory metacognitions is sufficient for participants to learn the predictive validity of cues and update the cue basis of their JOLs. To test these possibilities, in Experiment 4, we had three groups: (1) no-feedback group, (2) cognitive-plus-metacognitive-feedback group, and (3) metacognitive-only-feedback group. If metacognitive feedback alone is responsible for improving the cue basis of JOLs, we should observe improvements in both experimental groups. In contrast, if metacognitive feedback is a support for an accurate comprehension and implementation of cognitive feedback, we should observe improvements only in the cognitive-plus-metacognitive feedback group and not in the metacognitive-only-feedback group. In any case, replicating findings from Experiment 3, we expected improvements in the JOL cue basis in the cognitive-plus-metacognitive feedback group compared to the no-feedback group.

Method

Design

The design was a mixed design with font size (18, 48 point), number of study presentations (1, 2), and study-test cycle (1, 2, 3) as within-subjects factors and group (no-feedback, cognitive-plus-metacognitive-feedback, metacognitive-only-feedback) as between-subjects factor.

Materials

Stimuli were identical to Experiment 1.

Participants

We aimed at recruiting at least $N = 120$ participants. The primary analysis was the three-way interaction among group, cycle, and cue [number of study presentations or font size] from the four-way-ANOVA, so power calculations were identical to those reported in Experiment 1. We recruited 121 Technical University of Darmstadt undergraduates. Participants were randomly allocated to the no-feedback ($n = 40$), cognitive-plus-metacognitive-feedback ($n = 40$), and metacognitive-only-feedback ($n = 41$) groups. No participant was excluded using the same exclusion criteria as in Experiment 1.

Procedure

The procedure was identical to that of Experiment 3 with the following exception: Participants in the metacognitive-only-feedback group were presented with the table displaying the average memory performance of other participants (see social-reference-feedback group in Figure 1) followed by the information about illusory metacognitions regarding font size and number of study presentations (see procedure of Experiment 3 and Appendix 1), but did not receive the table showing their item-by-item JOL and recall performance (see recall-and-JOL-feedback group in Figure 1) nor the understanding feedback questions (see Table 3).

Results

JOLs

Figure 10 presents mean JOLs in each cycle by font size and number of study presentations in each group of Experiment 3. For words that were studied twice, the figure shows JOLs from the first study presentation. JOLs from the first study presentation were submitted to a mixed ANOVA with cycle (1, 2, 3), font size (18, 48 point), and number of study presentations (1, 2) as within-subjects factors and feedback group (no-feedback, cognitive-plus-metacognitive-feedback, metacognitive-only-feedback) as a between-subjects factor.

A significant main effect of cycle revealed differences in JOLs across cycles, $F(1.63, 191.94) = 7.28$, $p < .01$, $\eta_p^2 = .06$. Specifically, JOLs were lower in Cycles 2 and 3 than in Cycle 1, Cycles 1 vs. 2: $t(120) = 2.75$, $p < .01$, $d_z = 0.25$, Cycles 1 vs. 3: $t(120) = 3.11$, $p < .01$, $d_z = 0.28$, but did not differ between Cycles 2 and 3, $t(120) = 1.43$, $p = .16$, $d_z = 0.13$. A significant main effect of font size revealed higher JOLs for words displayed in the large font than for words displayed in the small font, $F(1, 118) = 57.09$, $p < .001$, $\eta_p^2 = .33$. A significant main effect of number of study presentations revealed higher JOLs for twice-presented words than for once-presented words, $F(1, 118) = 12.71$, $p < .001$, $\eta_p^2 = .10$. The main effect of group was not significant, $F < 1$.

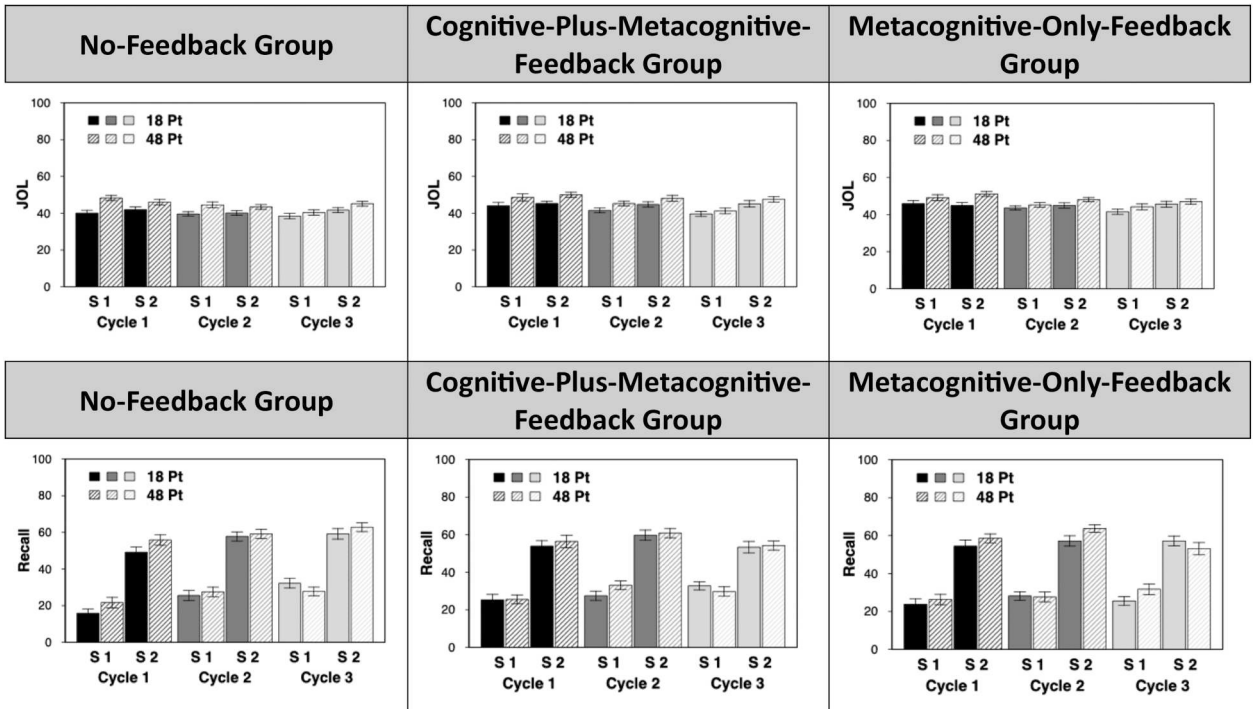

Figure 10. Mean judgments of learning (JOL) and percentage of correctly recalled words (recall) in each cycle for words presented once (S1) or twice (S2) in a small (18 pt) or a large (48 pt) font size in each group of Experiment 4. Note. Error bars represent one standard error of the mean.

As in Experiment 1 and 3, there was a significant interaction between cycle and font size, $F(1.81, 213.35) = 4.78$, $p < .05$, $\eta_p^2 = .04$. Follow-up t tests indicated that the size of the font size effect on JOLs decreased across cycles, Cycle 1: $t(120) = 5.94$, $p < .001$, $d_z = 0.54$, Cycle 2: $t(120) = 5.18$, $p < .001$, $d_z = 0.47$, Cycle 3: $t(120) = 3.59$, $p < .001$, $d_z = 0.33$. The interaction between cycle and number of study presentations was also significant, $F(2, 236) = 6.96$, $p = .001$, $\eta_p^2 = .06$. Follow-up t tests indicated that the effect of the number of study presentations on JOLs was not significant in Cycles 1 and 2, Cycle 1: $t(120) = 0.85$, $p = .40$, Cycle 2: $t(120) = 1.62$, $p = .11$, but was significant in Cycle 3, $t(120) = 4.64$, $p < .001$, $d_z = 0.42$. Neither the three-way interaction between group, cycle, and font size nor the three-way interaction between group, cycle, and number of study presentations were significant, $F < 1$. None of the other interactions were significant, $F <= 1.44$, $p >= .23$.

Recall performance

Figure 10 presents the mean percentage of recalled words in each cycle by font size and number of study presentations in each group of Experiment 3. A 3 (cycle: 1, 2, 3) x 2 (font size: 18, 48 point) x 2 (number of study presentations: 1, 2) x 3 (no-feedback, cognitive-plus-metacognitive-feedback group, metacognitive-only-feedback) mixed ANOVA on recall performance revealed that memory performance varied with cycle, $F(1.66, 195.43) = 7.44$, $p < .01$, $\eta_p^2 = .06$. Follow-up t tests showed worse memory performance in Cycle 1 than in Cycles 2 and 3, Cycles 1 vs. 2: $t(120) = 3.95$, $p < .001$, $d_z = 0.36$, Cycles 1 vs. 3: $t(120) = 2.50$, $p < .05$, $d_z = 0.23$, but no differences between Cycles 2 and 3, $t < 1$. A significant main effect of font size revealed better memory performance for words displayed in the large font than in the small font, $F(1, 118) = 6.39$, $p = .013$, $\eta_p^2 = .05$. A significant main effect of number of study presentations revealed better memory performance for twice-presented words than for once-presented words, $F(1, 118) = 915.02$, $p < .001$, $\eta_p^2 = .89$.

There was also a significant interaction between number of study presentations and cycle, $F(2, 236) = 3.36$, $p = .037$, $\eta_p^2 = .03$. Follow up t tests showed that the effect of number of study presentations on memory decreased across cycles, Cycle 1: $t(120) = 20.45$, $p < .001$, $d_z = 1.86$, Cycle 2: $t(120) = 19.25$, $p < .001$, $d_z = 1.75$, Cycle 3: $t(120) = 16.49$, $p < .001$, $d_z = 1.50$. Finally, there was an unexpected four-way significant interaction, $F(4, 236) = 3.35$, $p = .011$, $\eta_p^2 = .05$. Separate three-way ANOVAs for each group showed that the three-way interaction between cycle, font size, and number of study presentations was significant in the metacognitive-only-feedback group, $F(2, 80) = 4.88$, $p < .01$, $\eta_p^2 = .11$, but not in the other groups, $F <= 1.22$, $p >= .30$. Separate two-way ANOVAs for each cycle in the metacognitive-only-feedback group showed an interaction between font size and number of study presentations in Cycle 3, $F(1, 40) = 6.00$, $p = .019$, $\eta_p^2 = .13$, but not in Cycles 1 and 2, some states please. In Cycle 3, large-font words were remembered better than small-font words for once-presented words, $t(40) = 2.32$, $p = .025$, $d_z = 0.36$, but not for twice-presented words, $t(40) = 1.24$, $p > .05$.

Resolution

Gamma correlations

Table 2 and Figure 11 present Gamma correlations in each cycle for each group of Experiment 4. As in all previous experiments, all Gamma correlations by group and cycle were significantly positive, $t >= 4.20$, $p < .001$, $d >= .66$, indicating that participants from all groups made JOLs with above chance resolution in all study-test cycles. A 3(cycle) x 3(group) mixed ANOVA revealed a main effect of cycle, $F(2, 236) = 4.21$, $p = .015$, $\eta_p^2 = .03$, no main effect of group, $F(2, 188) = 1.54$, $p = .219$, $\eta_p^2 = .03$, and no significant interaction, $F < 1$. Follow-up t tests showed that resolution improved between Cycle 1 and Cycle 3, but not between Cycles 2 and 3 or Cycles 1 and 2, Cycles 1 vs. 2: $t(120) = 1.37$, $p = .17$, $d_z = 0.12$, Cycles 2 vs. 3: $t(120) = 1.58$, $p = .117$, $d_z = 0.14$, Cycles 1 vs. 3: $t(120) = 2.84$, $p < .01$, $d_z = 0.26$.

Logistic mixed-effects models

In the mixed-effects model analysis, we predicted recall performance (Recalled = 1, Not recalled = 0) from group, cycle, participant's mean-centered JOLs, and their interactions. To assess differences between groups, we used two orthogonal contrasts. The first contrast tested the difference between the control group (-1) and both groups with feedback (cognitive-plus-metacognitive-feedback, metacognitive-only-feedback; coded both as +1/2). The second contrast tested the difference between the cognitive-plus-metacognitive-feedback group (+1/2) and the metacognitive-only-feedback group (-1/2). We treated group, cycle, and JOLs as fixed effects predictors and specified random intercepts for participants.

Results showed a significantly positive coefficient for JOLs, $b = 0.24$ ($SE = 0.01$), $z = 24.44$, $p < .001$, indicating that JOLs were predictive of recall performance. The interaction between JOLs and cycle was also significant, $b = 0.05$ ($SE = 0.01$), $z = 4.56$, $p < .001$, which indicates that the predictivity of JOLs increased across cycles. A significant main effect of cycle, $b = 0.13$ ($SE = 0.02$), $z = 5.61$, $p < .001$, corroborating the improvement in recall performance across cycles was found in the mixed ANOVA. There was also a significantly negative coefficient for the interaction between the first group contrast and cycle, $b = -0.12$ ($SE = 0.03$), $z = -3.53$, $p < .001$, indicating less improvement in recall across cycles in the control group than in the two feedback groups. Finally, there was a significantly negative coefficient for the interaction between the second group contrast and JOLs, $b = -0.02$ ($SE = 0.01$), $z = -2.01$, $p = .045$, indicating that JOLs from the metacognitive-only-feedback group were more predictive than those from the cognitive-plus-metacognitive-feedback group. The three-way interaction among the first group contrast and cycle and JOLs was non-significant, $b = 0.01$ ($SE = 0.02$), $z = 0.88$, $p = 38$, the same for the three-way interaction of the second group contrast and cycle and JOLs, $b = -0.00$ ($SE = 0.01$), $z = -0.02$, $p = 98$, indicating that we did not replicate the improvement in resolution in the cognitive-plus-metacognitive-feedback group. No other effects were significant, $z <= 1.78$, $p >= .07$.

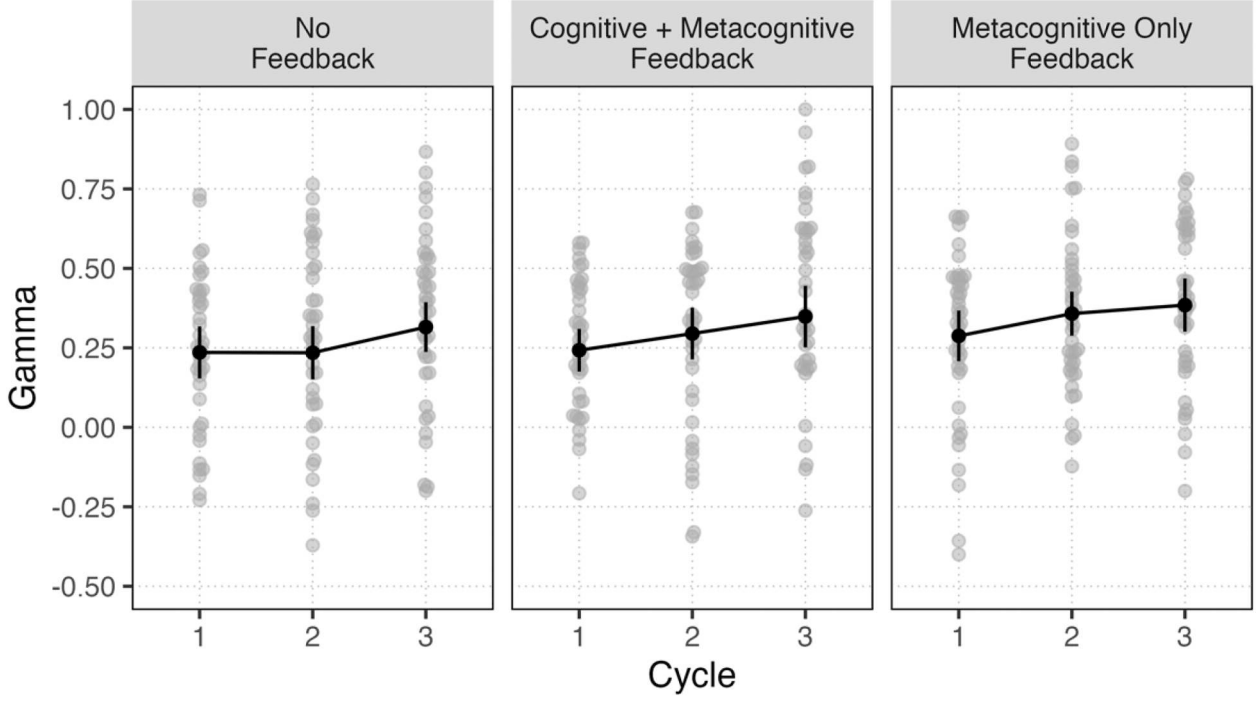

Figure 11. Mean gamma correlation between JOLs and recall performance in each cycle and group of Experiment 4. Note. Error bars represent one standard error of the mean.

Calibration

A 3(cycle) x 3(group) mixed ANOVA on calibration revealed that bias varied with cycle, $F(1.66, 195.84) = 16.19$, $p < .001$, $\eta_p^2 = .12$, with pairwise comparisons indicating a switch from overconfidence in Cycle 1 ($M = 7.39$, $SD = 21.14$) to good calibration in Cycle 2 ($M = 0.14$, $SD = 20.90$) and Cycle 3 ($M = -0.11$, $SD = 23.79$). Cycles 1 vs. 2: $t(120) = 5.14$, $p < .001$, $d_z = 0.47$, Cycles 1 vs. 3: $t(79) = 4.14$, $p < .001$, $d_z = 0.38$, Cycles 2 vs. 3: $t < 1$. Neither the main effect of group, $F < 1$, nor the interaction were significant, $F(3.32, 195.84) = 1.47$, $p = .221$, $\eta_p^2 = .02$.

Questionnaire data

Table 5 shows participants' responses to the questions after item-by-item feedback in each cycle (see Table 3). As in Experiment 3, most participants correctly indicated that recall performance was greater for twice- than once-presented words across all cycles, and there was a shift after Cycle 1 toward judging recall as equal for large- and small-font words. Also, participants' confidence in their JOLs again increased across cycles.

As in Experiment 3, we separately analysed JOLs from each cycle for two subgroups of participants from the *cognitive-plus-metacognitive-feedback group*. The first subgroup consisted of participants who correctly indicated that recall performance was greater for twice-presented than once-presented words. The second subgroup consisted of participants who correctly indicated that recall performance was similar for large and small words. Comparing JOLs from each subgroup of participants with those from the control group revealed no interactions between group and cue effects, indicating that the feedback manipulation also did not work for them (for details, see Supplementary Material 2).

Discussion

In contrast to Experiment 3, results of Experiment 4 showed that providing metacognitive feedback in addition to cognitive feedback did not improve the cue basis of JOLs. This was found even though, as in Experiment 3, most participants accurately reported better memory performance for words learned twice than once in the questionnaire after the item-by-item feedback. We will discuss this discrepancy in results in the General Discussion. Experiment 4 also showed that the metacognitive feedback alone was not effective at improving the cue basis of JOLs. Overall, participants in all groups continued to overweight font size (main effect on JOLs across cycles; $\eta_p^2 = .33$; main effect on recall across cycles; $\eta_p^2 = .05$) and number of study presentations (main effect on JOLs across cycles; $\eta_p^2 = .10$; main effect on recall across cycles; $\eta_p^2 = .89$). As in Experiments 1 and 3, the font size effect on JOLs decreased with experience across cycles in all groups. Further, there was a significant effect of number of study presentations on JOLs in Cycle 3. This suggests that experience enabled participants to improve the cue basis of their JOLs.

Regarding resolution, Gamma correlations showed improved resolution from experience with the task. This result was replicated in a logistic mixed-effects model. The logistic-mixed-effects model also showed higher resolution of JOLs in the metacognitive-only-feedback group than in the cognitive-plus-metacognitive-feedback group. However, this effect did not interact with cycle and was likely not due to the provision of feedback but from pre-experimental differences between groups.

Table 5. Percentage of participants from the cognitive-plus-metacognitive-feedback group who reported (correctly) each of the three response options for the recall and prediction question in each cycle of experiment 4.

		Response Options					
		Recall Question			Prediction Question		
		twice [large] > once [small]	twice [large] < once [small]	twice [large] = once [small]	Under-estim. of twice [large]	Correct estim.	Over-estim. of twice [large]
Cycle 1	# Study Present.	75.00 (72.50)	15.00 (10.00)	10.00 (5.00)	40.00 (37.50)	37.50 (2.50)	22.50 (5.00)
	Font Size	47.50 (30.00)	22.50 (17.50)	30.00 (5.00)	35.00 (20.00)	37.50 (2.50)	27.50 (20.00)
Cycle 2	# Study Present.	90.00 (87.50)	2.50 (0.00)	7.50 (2.50)	22.50 (22.50)	47.50 (2.50)	30.00 (7.50)
	Font Size	40.00 (32.50)	17.50 (10.00)	42.50 (5.00)	22.50 (17.50)	57.50 (0.00)	20.00 (12.50)
Cycle 3	# Study Present.	72.50 (72.50)	10.00 (5.00)	17.50 (0.00)	25.00 (25.00)	37.50 (0.00)	37.50 (15.00)
	Font Size	30.00 (25.00)	25.00 (20.00)	45.00 (15.00)	15.00 (15.00)	62.50 (10.00)	22.50 (17.50)

Notes: Correct percentages for the recall questions were determined by comparing participants' responses with their actual recall performance. For example, if a participant stated that twice-presented words were remembered more often than once-presented words, and their recall performance confirmed this pattern, the response was scored as correct. For the prediction questions, correctness was assessed by comparing the mean JOL difference between the two levels of a factor with the mean recall performance difference. If the mean JOL difference exceeded the mean recall difference, the response was classified as correct overestimation; if it was smaller, as correct underestimation; and if the two values matched, as correct estimation.

Further, the logistic mixed-effects model showed better recall performance in the experimental groups than in the control group. Although the feedback was not intended to increase recall performance, it might have increased motivation, or it might have supported performance monitoring and strategy acquisition. However, this finding should be considered with caution since it was not found in the other experiments. Finally, calibration also improved with task experience: JOLs were overconfident in Cycle 1 but well-calibrated in Cycles 2 and 3.

General discussion

This study systematically investigated the effectiveness of different forms of feedback for improving the cue basis and accuracy of JOLs in the context of metacognitive illusions. Each of the four experiments aimed to remedy two illusions by providing multiple study-test-cycles with novel study lists, and feedback or no feedback after each cycle. In all experiments, one of the illusions was the font size illusion (Rhodes & Castel, 2008). In Experiments 1, 3, and 4, we additionally focused on the stability bias – the assumption that memory will remain stable over time and will not benefit from future learning (Kornell & Bjork, 2009). In Experiment 2, we additionally focused on the font format illusion (Rhodes & Castel, 2008).

We found that cognitive feedback – presented in a table showing individual task performance (recall and JOL) for each item – was ineffective in correcting either illusion (Experiments 1 and 2). While Experiment 3 showed that additional metacognitive feedback – written information about the biased nature of illusory metacognition – partially remedied the stability bias, we could not replicate this result in Experiment 4. Although we cannot exclude the possibility that a reduction in the stability bias through feedback replicates in other experiments, the present findings suggest that it is either not robust or considerably smaller than would be expected for a meaningful effect. Additionally, it is important to note that the reduction in stability bias found in Experiment 3 was only partial, with participants still showing substantial underestimation of the number-of-study-presentations cue. In summary, this study demonstrates that feedback at the cognitive and metacognitive level did not mend metacognitive illusions in JOLs.

One may wonder why metacognitive illusions in JOLs were not eliminated by the feedback provided in this study. One possibility is that participants did not understand the feedback. While we cannot rule out this possibility in Experiment 1 and 2, the questionnaire data from Experiment 3 and 4 clearly show that participants understood the feedback: Most participants correctly recognised that words studied twice were remembered better than those studied once, and about half correctly reported the lack of influence of font size on memory. Although we did not directly assess the participants' beliefs about memory, the questionnaire data suggest that they likely corrected their prior faulty beliefs when acquiring knowledge about cue validities. Despite this, conditional JOL analyses for these subgroups of participants suggest that they did not adjust the cue basis of their JOLs accordingly. One possible reason for this is that they failed to use corrected beliefs about cues at the general level when making JOLs at the trial level. As other metamemory studies have indicated, beliefs must be activated to impact JOLs (e.g., Ariel et al., 2014; Schaper & Bayen, 2025; Undorf & Erdfelder, 2015). Thus, remedying metacognitive illusions in item-by-item JOLs is not only a matter of acquiring correct beliefs but also of finding ways to make learners use the corrected beliefs when making the judgments. This is consistent with prior research showing that participants can hold correct beliefs about how cues impact their memory as measured by global predictions and study strategy effectiveness ratings, but still do not use the cues or use them inaccurately when making item-by-item judgments (Dunlosky & Hertzog, 2000; Hertzog et al., 2009; Mueller et al., 2015; Tullis et al., 2013).

A failure to implement a corrected belief adequately may be due to other prominent experiential cues during the learning process, such as fluency or idiosyncratic cues (e.g., Koriat & Ackerman, 2010; Undorf & Erdfelder, 2011; Undorf et al., 2022a, 2022b). In comparison to judgments about external criteria (Karelaia & Hogarth, 2008), there is a large systematic covariation between metacognitive judgments and the memory criterion beyond what can be explained by cues known to the researcher (Bröder & Undorf, 2019; Undorf et al., 2022). Further research is needed to investigate how participants can apply their corrected knowledge effectively in the presence of other cues when making item-by-item metacognitive judgments. One idea in this direction is to enhance participants' awareness of their item-wise judgment formation. Our questionnaire data suggest that participants struggle with evaluating how they considered cues in their JOLs, indicating that their use of this knowledge for item-by-item judgments may be flawed.

Despite metacognitive illusions not being remedied, we still found improvements from task experience in the cue basis of JOLs and in resolution. Specifically, the font size illusion reliably reduced across cycles in all three experiments that additionally manipulated the number of study presentations, which is a valid cue. In contrast, when font size was manipulated simultaneously with font format (Experiment 2), another invalid cue, the font size illusion did not reduce. This suggests that the weight with which individual cues affect JOLs may depend on whether other valid or invalid cues are manipulated simultaneously. For example, other studies have shown that effects of font size on JOLs can be moderated or eliminated when other cues are manipulated simultaneously (Chang & Brainerd, 2023; Luna, Albuquerque, et al., 2019; Experiment 3 and 6; Rhodes & Castel, 2008).

Furthermore, in all experiments, logistic mixed-effects model analysis consistently showed that JOLs become more predictive of test performance with experience across cycles. This is interesting because participants studied novel materials across cycles and could not base their JOLs on memory for past test performance, which is a reliable predictor of future performance (Ariel & Dunlosky, 2011; Finn & Metcalfe, 2008; Tauber & Rhodes, 2012). Based on the assumption that reliance on valid cues determines resolution (Koriat, 1997), the resolution improvement across cycles in our study shows that task experience is effective in discovering valid cues predictive of memory performance. Those cues are probably ones under the participant's control, such as the detection of effective learning strategies. Future studies could investigate the cue basis underlying resolution improvements from task experience. Taken together, the experience-related effects on the cue basis and resolution of JOLs observed in this study suggest that monitoring improvements from experience can occur independently of feedback. This challenges the *inferential deficit hypothesis,* which posits that JOLs do not improve across cycles due to a failure to monitor test performance and make correct inferences from cues (Dunlosky & Hertzog, 2000; Matvey et al., 2002).

In this study, we investigated whether two types of feedback could reduce metacognitive illusions in judgments of learning. Consistent with prior work (e.g., Dunlosky & Hertzog, 2000; Hertzog et al., 2009; Kornell & Bjork, 2009; Mueller et al., 2015; Pan & Rivers, 2023; Yan et al., 2016), our findings highlight the difficulty of altering the biased cue basis underlying item-level judgments. Thus, we cannot yet recommend effective strategies for alleviating such metacognitive illusions in applied settings. Nevertheless, both cognitive and metacognitive feedback – though task-specific – may be adapted to contexts where valid and invalid cues are known or can be identified. Implementing metacognitive feedback further requires identifying the relevant metacognitive experiences and their influence on judgments, either directly or indirectly via beliefs. While figuring out cues and metacognitive experiences is demanding, it is feasible and warranted if such feedback proves effective. Importantly, alternative approaches to debiasing, such as warnings, entail similar prerequisites, as they also depend on the identification of cues likely to mislead participants.

One limitation is that the highly detailed feedback provided in this study may have imposed substantial working memory demands on participants, thereby reducing its effectiveness. Future studies could examine simpler versions of cognitive and metacognitive feedback. At the same time, determining the optimal balance between overly detailed and overly simplistic feedback will likely remain a challenge.

In conclusion, this study shows that feedback at the cognitive level (JOL and recall status of each studied item) and at the metacognitive level (information about illusory metacognition) do not remedy metacognitive illusions in item-wise JOLs. This is likely because of a failure to implement corrected beliefs about memory when making item-wise judgments. We recommend that further studies do not only focus on debiasing metacognitive beliefs but also focus on strategies for an adequate belief implementation in metacognitive judgment formation. Importantly, this study also shows that despite persisting illusions, modest monitoring improvements from task experience can occur.

Note

1. Due to an error, 39 participants were assigned to the no-feedback group and 41 to the recall-feedback group.

Open Scholarship

This article has earned the Center for Open Science badges for Open Data and Open Materials through Open Practices Disclosure. The data and materials are openly accessible at https://osf.io/vgy7d and link: https://osf.io/vgy7d.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by grant 404409389 (Monika Undorf & Arndt Bröder) from the Deutsche Forschungsgemeinschaft and by the University of Mannheim's Graduate School of Economic and Social Sciences (Sofia Navarro-Báez). We thank Elena Huber, Maike Klingemann, Linn Körner, Marina Meißner, Leila Plohmann, Franziska Ingendahl, Charlotte Smid, Anna Fee Wefelmeier, and Clara Wiebel for help with data collection.

Authors' contributions

MU and AB developed the study concept. All authors contributed to the study methodology and design. MU and SNB programmed the experiments. SNB performed the data analysis under the supervision of MU and AB. SNB wrote the original manuscript draft. MU and AB edited parts of the manuscript and provided critical revisions. All authors approved the final version of the manuscript for submission.

Availability of data and materials

The data, materials, and analysis code for all experiments are available at https://osf.io/vgy7d/?view_only=5d02381f12754484ad1c422cd1d4f44b. Analysis code (R script) available on OSF (see above).

Ethics approval

As no deception was used, data were collected anonymously and no psychological or physical risk was involved, the study did not require approval under the statutes of the IRB in Mannheim or Darmstadt. Participation was voluntary, and participants were free to withdraw from the study at any time without penalty.

Open practices statement

The data, materials, and analysis code for all experiments are available at *https://osf.io/vgy7d/?view_only=5d02381f12754484ad1c422cd1d4f44b*.

Consent to participate

Before the experiment began, participants read an informed consent form, which included a description of the study tasks and information about data handling. Participants could then declare that they were at least 18 years old and that they have understood the information in the consent form. They could either agree or refuse to participate and were told that they could withdraw their consent at any time without negative consequences.

Consent for publication

Participants were informed about publication of the data in anonymous form in the informed consent form.

References

Ariel, R., & Dunlosky, J. (2011). The sensitivity of judgment-of-learning resolution to past test performance, new learning, and forgetting. *Memory & Cognition*, *39*(1), 171–184. https://doi.org/10.3758/s13421-010-0002-y

Ariel, R., Hines, J. C., & Hertzog, C. (2014). Test framing generates a stability bias for predictions of learning by causing people to discount their learning beliefs. *Journal of Memory and Language*, *75*, 181–198. https://doi.org/10.1016/j.jml.2014.06.003

Balzer, W. K., Doherty, M. E., & O'Connor, R. (1989). Effects of cognitive feedback on performance. *Psychological Bulletin*, *106*(3), 410–433. https://doi.org/10.1037/0033-2909.106.3.410

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

Castel, A. D. (2008). Metacognition and learning about primacy and recency effects in free recall: The utilization of intrinsic and extrinsic cues when making judgments of learning. *Memory & Cognition*, *36*(2), 429–437. https://doi.org/10.3758/MC.36.2.429

Chang, M., & Brainerd, C. J. (2022). Association and dissociation between judgments of learning and memory: A meta-analysis of the font size effect. *Metacognition and Learning*, *17*(2), 443–476. https://doi.org/10.1007/s11409-021-09287-3

Chang, M., & Brainerd, C. J. (2023). The font size effect depends on inter-item relation. *Memory & Cognition*, *51*(7), 1702–1713. https://doi.org/10.3758/s13421-023-01419-1

Cooper, S., & Vallée-Tourangeau, F. (2021). The effects of numeracy and presentation format on judgments of contingency. *Memory & Cognition*, *49*(2), 389–399. https://doi.org/10.3758/s13421-020-01084-8

De Martino, B., Bobadilla-Suarez, B., Nouguchi, S., Sharot, T., Love, T., & C, B. (2017). Social information is integrated into value and confidence judgments according to its reliability. *The Journal of Neuroscience*, *37*(25), 6066–6074. https://doi.org/10.1523/JNEUROSCI.3880-16.2017

Dunlosky, J., & Hertzog, C. (2000). Updating knowledge about encoding strategies: A componential analysis of learning about strategy effectiveness from task experience. *Psychology and Aging*, *15*(3), 462–474. https://doi.org/10.1037/0882-7974.15.3.462

Dunlosky, J., Mueller, M. L., Morehead, K., Tauber, S. K., Thiede, K. W., & Metcalfe, J. (2021). Why does excellent monitoring accuracy not always produce gains in memory performance? *Zeitschrift Für Psychologie*, *229*(2), 104–119. https://doi.org/10.1027/2151-2604/a000441

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

Fiedler, K., Schott, M., Kareev, Y., Avrahami, J., Ackerman, R., Goldsmith, M., Mata, A., Ferreira, M. B., Newell, B. R., & Pantazi, M. (2020). Metacognitive myopia in change detection: A collective approach to overcome a persistent anomaly. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(4), 649–668. https://doi.org/10.1037/xlm0000751

Finn, B., & Metcalfe, J. (2008). Judgments of learning are influenced by memory for past test. *Journal of Memory and Language*, *58*(1), 19–34. https://doi.org/10.1016/j.jml.2007.03.006

Hertzog, C., Price, J., Burpee, A., Frentzel, W. J., Feldstein, S., & Dunlosky, J. (2009). Why do people show minimal knowledge updating with task experience: Inferential deficit or experimental artifact? *Quarterly Journal of Experimental Psychology*, *62*(1), 155–173. https://doi.org/10.1080/17470210701855520

Karelaia, N., & Hogarth, R. M. (2008). Determinants of linear judgment: A meta-analysis of lens model studies. *Psychological Bulletin*, *134*(3), 404–426. https://doi.org/10.1037/0033-2909.134.3.404

Karlsson, L., Juslin, P., & Olsson, H. (2004). Representational shifts in a multiple-cue judgment task with continuous cues. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *26*(26), 648–653.

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

Koriat, A., & Ackerman, R. (2010). Metacognition and mindreading: Judgments of learning for self and other during self-paced study. *Consciousness and Cognition*, *19*(1), 251–264. https://doi.org/10.1016/j.concog.2009.12.010

Koriat, A., & Bjork, R. A. (2006a). Illusions of competence during study can be remedied by manipulations that enhance learners' sensitivity to retrieval conditions at test. *Memory & Cognition*, *34*(5), 959–972. https://doi.org/10.3758/BF03193244

Koriat, A., & Bjork, R. A. (2006b). Mending metacognitive illusions: A comparison of mnemonic-based and theory-based procedures. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(5), 1133–1145. https://doi.org/10.1037/0278-7393.32.5.1133

Koriat, A., Sheffer, L., & Ma'ayan, H. (2002). Comparing objective and subjective learning curves: Judgments of learning exhibit increased underconfidence with practice. *Journal of Experimental Psychology: General*, *131*(2), 147–162. https://doi.org/10.1037/0096-3445.131.2.147

Kornell, N., & Bjork, R. A. (2009). A stability bias in human memory: Overestimating remembering and underestimating learning. *Journal of Experimental Psychology: General*, *138*(4), 449–468. https://doi.org/10.1037/a0017350

Little, D. R., & Lewandowsky, S. (2009). Better learning with more error: Probabilistic feedback increases sensitivity to correlated cues in categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(4), 1041–1061. https://doi.org/10.1037/a0015902

Luna, K., Albuquerque, P. B., & Martín-Luengo, B. (2019). Cognitive load eliminates the effect of perceptual information on judgments of learning with sentences. *Memory & Cognition*, *47*(1), 106–116. https://doi.org/10.3758/s13421-018-0853-1

Luna, K., Martín-Luengo, B., & Albuquerque, P. B. (2018). Do delayed judgements of learning reduce metamemory illusions? A meta-analysis. *Quarterly Journal of Experimental Psychology*, *71*(7), 1626–1636. https://doi.org/10.1080/17470218.2017.1343362

Luna, K., Nogueira, M., & Albuquerque, P. B. (2019). Words in larger font are perceived as more important: Explaining the belief that font size affects memory. *Memory (Hove, England)*, *27*(4), 555–560. https://doi.org/10.1080/09658211.2018.1529797

Masson, M. E. J., & Rotello, C. M. (2009). Sources of bias in the Goodman – Kruskal gamma coefficient measure of association: Implications for studies of metacognitive processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(2), 509–527. https://doi.org/10.1037/a0014876

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

Miller, T. M., & Geraci, L. (2011). Training metacognition in the classroom: The influence of incentives and feedback on exam predictions. *Metacognition and Learning*, *6*(3), 303–314. https://doi.org/10.1007/s11409-011-9083-7

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

Mueller, M. L., Dunlosky, J., Tauber, S. K., & Rhodes, M. G. (2014). The font-size effect on judgments of learning: Does it exemplify fluency effects or reflect people's beliefs about memory? *Journal of Memory and Language*, *70*, 1–12. https://doi.org/10.1016/j.jml.2013.09.007

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

Murayama, K., Sakaki, M., Yan, V. X., & Smith, G. M. (2014). Type I error inflation in the traditional by-participant analysis to metamemory accuracy: A generalized mixed-effects model perspective. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(5), 1287–1306. https://doi.org/10.1037/a0036914

Nelson, T. O., & Narens. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *Psychology of learning and motivation* (Vol. 26, pp. 125–173). Elsevier. https://doi.org/10.1016/S0079-7421(08)60053-5.

Newell, B. R., Weston, N. J., Tunney, R. J., & Shanks, D. R. (2009). The effectiveness of feedback in multiple-cue probability learning. *Quarterly Journal of Experimental Psychology*, *62*(5), 890–908. https://doi.org/10.1080/17470210802351411

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

Pan, S. C., & Rivers, M. L. (2023). Metacognitive awareness of the pretesting effect improves with self-regulation support. *Memory & Cognition*, *51*(6), 1461–1480. https://doi.org/10.3758/s13421-022-01392-1

Rhodes, M. G. (2016). *Judgments of learning: Methods, data, and theory* (J. Dunlosky & S. (Uma) K. Tauber, Eds.; Vol. 1). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199336746.013.4

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

Schaper, M. L., & Bayen, U. J. (2025). Manipulating belief partially remedies the metamemory expectancy illusion in schema-based source monitoring. *Memory & Cognition*, https://doi.org/10.3758/s13421-025-01757-2

Seong, Y., & Bisantz, A. M. (2008). The impact of cognitive feedback on judgment performance and trust with decision aids. *International Journal of Industrial Ergonomics*, *38*(7–8), 608–625. https://doi.org/10.1016/j.ergon.2008.01.007

Smithson, C. J. R., Eichbaum, Q. G., & Gauthier, I. (2023). Object recognition ability predicts category learning with medical images. *Cognitive Research: Principles and Implications*, *8*(1), 9. https://doi.org/10.1186/s41235-022-00456-9

Spellman, B. A., Bloomfield, A., & Bjork, R. A. (2014). Measuring memory and metamemory. In *Handbook of metamemory and memory*. Routledge, https://doi.org/10.4324/9780203805503.ch6

Sungkhasettee, V. W., Friedman, M. C., & Castel, A. D. (2011). Memory and metamemory for inverted words: Illusions of competency and desirable difficulties. *Psychonomic Bulletin & Review*, *18*(5), 973–978. https://doi.org/10.3758/s13423-011-0114-9

Tauber, S. K., & Rhodes, M. G. (2010). Metacognitive errors contribute to the difficulty in remembering proper names. *Memory (Hove, England)*, *18*(5), 522–532. https://doi.org/10.1080/09658211.2010.481818

Tauber, S. K., & Rhodes, M. G. (2012). Multiple bases for young and older adults' judgments of learning in multitrial learning. *Psychology and Aging*, *27*(2), 474–483. https://doi.org/10.1037/a0025246

Thiede, K. W., Anderson, M. C. M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology*, *95*(1), 66–73. https://doi.org/10.1037/0022-0663.95.1.66

Tullis, J. G., & Benjamin, A. S. (2012). Consequences of restudy choices in younger and older learners. *Psychonomic Bulletin & Review*, *19*(4), 743–749. https://doi.org/10.3758/s13423-012-0266-2

Tullis, J. G., Finley, J. R., & Benjamin, A. S. (2013). Metacognition of the testing effect: Guiding learners to predict the benefits of retrieval. *Memory & Cognition*, *41*(3), 429–442. https://doi.org/10.3758/s13421-012-0274-5

Undorf, M., & Bröder, A. (2020). Cue integration in metamemory judgements is strategic. *Quarterly Journal of Experimental Psychology*, *73*(4), 629–642. https://doi.org/10.1177/1747021819882308

Undorf, M., & Erdfelder, E. (2011). Judgments of learning reflect encoding fluency: Conclusive evidence for the ease-of-processing hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(5), 1264–1269. https://doi.org/10.1037/a0023719

Undorf, M., & Erdfelder, E. (2015). The relatedness effect on judgments of learning: A closer look at the contribution of processing fluency. *Memory & Cognition*, *43*(4), 647–658. https://doi.org/10.3758/s13421-014-0479-x

Undorf, M., Navarro-Báez, S., & Bröder, A. (2022a). "You don't know what this means to me" – Uncovering idiosyncratic influences on metamemory judgments. *Cognition*, *222*(105011), 1–9. https://doi.org/10.1016/j.cognition.2021.105011

Undorf, M., Navarro-Báez, S., & Zimdahl, M. F. (2022b). Metacognitive illusions. In R. F. Pohl, *Cognitive illusions (3rd ed.*, pp. 307–323). Routledge. https://doi.org/10.4324/9781003154730-22

Undorf, M., Söllner, A., & Bröder, A. (2018). Simultaneous utilization of multiple cues in judgments of learning. *Memory & Cognition*, *46*(4), 507–519. https://doi.org/10.3758/s13421-017-0780-6

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, and Cognition*, *45*(1), 97–109. https://doi.org/10.1037/xlm0000571

Võ, M. L. H., Conrad, M., Kuchinke, L., Urton, K., Hofmann, M. J., & Jacobs, A. M. (2009). The Berlin affective word list reloaded (BAWL-R). *Behavior Research Methods*, *41*(*2*), 534–538. https://doi.org/10.3758/BRM.41.2.534

Yan, V. X., Bjork, E. L., & Bjork, R. A. (2016). On the difficulty of mending metacognitive illusions: A priori theories, fluency effects, and misattributions of the interleaving benefit. *Journal of Experimental Psychology: General*, *145*(7), 918–933. https://doi.org/10.1037/xge0000177

Appendix

As you have seen, many students overestimated the influence of font size on their memory in this part of the experiment.

This is because words written in a large font size are particularly conspicuous during learning and are perceived as particularly easy to read and learn. However, these perceptions during learning say little about actual test performance: The font size of words does not usually affect memory performance.

You have also seen that many students underestimated the influence of an additional learning opportunity on their memory.

This is because the sensations during learning are very similar for words with an additional learning opportunity and for words without an additional learning opportunity. Despite the similarity in the sensations during learning, the following applies: An additional learning opportunity greatly improves memory performance in the test.

The following therefore applies in this experiment:

The perceptions about the ease of learning triggered by the font size say little about how well the words can actually be learnt. These perceptions should therefore not play a role in your assessments. An additional learning opportunity, on the other hand, has a stronger influence on memory.