Metacognitive Unawareness of the Errorful Generation Benefit and its Effects on Self-Regulated Learning

Chunliang Yang, Rosalind Potts, and David R. Shanks

University College London

Author Note

This research was supported by the China Scholarship Council (CSC).

Correspondence concerning this article should be addressed to David R. Shanks, Division of Psychology and Language Sciences, University College London, 26 Bedford Way, London WC1H 0AP. Email: d.shanks@ucl.ac.uk.

All data have been made publicly available via the Open Science Framework (OSF) at https://osf.io/9bk8y/.

Abstract

Generating errors followed by corrective feedback enhances retention more effectively than does reading – the benefit of errorful generation – but people tend to be unaware of this benefit. The current research explored this metacognitive unawareness, its effect on self-regulated learning, and how to alleviate or reverse it. People's beliefs about the relative learning efficacy of generating errors followed by corrective feedback compared to reading, and the effects of generation fluency, are also explored. In Experiments 1 and 2, lower judgements of learning (JOLs) were consistently given to incorrectly generated word pairs than to studied (read) pairs and led participants to distribute more study resources to incorrectly generated pairs, even though superior recall of these pairs was exhibited in the final test. In Experiment 3, a survey revealed that people believe that generating errors followed by corrective feedback is inferior to reading. Experiment 4 was designed to alter participants' metacognition by informing them of the errorful generation benefit prior to study. Although metacognitive misalignment was partly countered, participants still tended to be unaware of this benefit when making item-by-item JOLs. In Experiment 5, in a delayed JOL condition, higher JOLs were given to incorrectly generated pairs and read pairs were more likely to be selected for restudy. The current research reveals that people tend to underestimate the learning efficiency of generating errors followed by corrective feedback relative to reading when making immediate item-by-item JOLs. Informing people of the errorful generation benefit prior to study and asking them to make delayed JOLs are effective ways to alleviate this metacognitive miscalibration.

Keywords: metacognitive unawareness; errorful generation benefit; self-regulated learning; delayed JOLs

Increasing importance is being attached to self-regulated learning with the development of technologies such as the internet, smart phones and web-based courses, that can support learning (Bjork, Dunlosky, & Kornell, 2013; Kornell & Bjork, 2007; Soderstrom & Bjork, 2014; Yan, Thai, & Bjork, 2014). The complexity and rapidly changing pace of technology create many situations for self-regulated learning, outside formal class and without explicit guidance from educators. Many self-regulated learning strategies have been investigated (Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013). Previous research has found that some effective techniques are not as well appreciated by learners as less effective ones. For instance, the merits of self-testing, reviewing of studied materials, and spacing study tend to be underestimated (Kornell & Bjork, 2007, 2008a; Yan, Thai, et al., 2014). Hence, how people manage their study to foster effective and enduring learning is a core challenge in cognitive and behavioural studies.

Individuals' decisions about selecting what information to study and how to allocate study time are two important aspects of self-regulated learning. The current research investigates how people manage their learning according to their metamemory monitoring, especially focusing on restudy decision making and restudy time allocation in the context of errors committed during learning.

Metamemory monitoring illusions and their effect on metamemory control

Metamemory has been intensively studied because of its importance in metamemory monitoring (assessing one's on-going learning) and metamemory control (managing one's learning). Previous research has found that metamemory control is related to metamemory monitoring. People manage their learning to decrease the gap between their perceived on-going learning state and their expected mastery of studied materials (Dunlosky & Ariel, 2011; Dunlosky et al., 2013). Son and Metcalfe (2000) undertook a comprehensive literature review on the relationship between metacognitive judgements and subsequent study time allocation. Thirty-five out of 46 studies showed a positive relationship between judged difficulty and study time allocation. More study time is allocated to materials which are judged less likely to be remembered (T. O. Nelson & Leonesio, 1988; Soderstrom & Bjork, 2014; Soderstrom, Clark, Halamish, & Bjork, 2015; Son & Metcalfe, 2000).

Learners are able to regulate their learning optimally when their assessments of learning are accurate (Kornell & Metcalfe, 2006). For instance, Kornell and Metcalfe (2006) allowed participants to choose which half of a set of word pairs to restudy. In a final test, participants in the honouring condition, who restudied the pairs they selected, significantly outperformed those in the dishonouring condition who restudied the pairs they did not select. These findings reveal that people have the ability to make reasonable decisions about selecting which pairs to restudy and that self-regulated restudy choices can be made rationally to enhance memory outcomes. However, giving learners control over their learning processes does not always lead to better learning. For example, Kornell and Bjork (2008b) asked participants to study Swahili-English word pairs. Some participants were allowed to drop some pairs which they thought they knew well during studying. Others had no opportunity to do so and were asked to restudy all pairs. Participants who were allowed to drop some pairs during learning stopped learning prematurely. Being allowed to remove pairs from study impaired participants' learning, slightly but consistently. Therefore, Kornell and Bjork (2008b) emphasized that the efficacy of self-regulated learning is highly dependent on the accuracy of metacognitive monitoring.

Many studies have been conducted to examine the consequences of metacognitive illusions on metacognitive control (Finn, 2008; Metcalfe & Finn, 2008; Rhodes & Castel, 2009). Bias in judgments of learning (JOLs) can affect people's subsequent study strategies (Finn, 2008; Mazzoni & Cesare, 1993; Metcalfe, 2002; Metcalfe & Finn, 2008; Rhodes & Castel, 2009). For instance, Rhodes and Castel (2009) found that JOLs can be influenced by auditory perceptual information. In their study, lower JOLs were made to quiet words, and participants preferred to restudy these words, even though auditory volume did not affect memory. Other research has similarly demonstrated that people are more likely to restudy items to which they give lower JOLs, even though no difference in actual memory is found in a later test (Finn, 2008; Metcalfe & Finn, 2008; Rhodes & Castel, 2009). Are there some circumstances in which people attach lower JOLs to better remembered information? And if there are, will these metamemory illusions affect people's restudy choices and restudy time allocation? We set out to explore a situation under which people give lower JOLs to better remembered information and asked whether they choose more of these items to restudy and/or spend more time restudying them. This research

provides novel and striking evidence to support the important theoretical claim that metamemory control is more strongly related to metamemory monitoring than to actual retention. In a departure from previous research investigating metamemory illusions and control (Finn, 2008; Metcalfe & Finn, 2008; Rhodes & Castel, 2009), the errorful generation paradigm was employed in the current study.

The errorful generation benefit

Previous studies have revealed that, in some situations, generating errors followed by corrective feedback enhances learning more effectively than spending the same amount of time on studying/reading (Kornell, Hays, & Bjork, 2009; Potts & Shanks, 2014). Several terms and phrases have been used to refer to this effect. Potts and Shanks (2014) used the term "errorful generation" to refer to the learning of novel associations and "unsuccessful retrieval", following Kornell et al. (2009), for the situation where responses are generated to cues which have pre-existing semantic associations. For simplicity, in this article we use the term "errorful generation benefit" to refer to the memorial advantage of generating errors compared with reading in either scenario.

In Kornell et al.'s (2009) Experiment 3, participants were asked to study 60 weakly associated English word pairs (e.g., *pond-frog*), 30 in a Read condition and 30 in a Generate condition. In the Read condition, a cue word and target were presented alongside each other and studied for 5 sec. In the Generate condition, a cue word was presented for 8 sec and participants were asked to guess the target; corrective feedback was then provided (the cue and target were presented together for 5 sec). In a later test, incorrectly generated pairs were better recalled than Read pairs. More strikingly, in their Experiment 4, even though the exposure time of Read pairs was extended to 13 sec, the same total duration as for Generate pairs, incorrectly generated pairs were still better recalled than Read pairs in a later test. Three theoretical explanations were proposed by Kornell et al. (2009) to account for this errorful generation benefit. Grimaldi and Karpicke (2012) termed these three theories the *search set theory*, the *error correction theory* and the *additional cue theory*.

According to the *search set theory*, attempts to retrieve information from memory activate related candidates, which potentiate subsequent learning of the correct answer. Although people tend

to retrieve a strongly related candidate (e.g., *water*) and produce it as a response when shown a cue (e.g., *pond-?*), other possible candidates (less strongly associated ones, e.g., *frog, fish, duck, swim,* etc.) are activated synchronously and facilitate subsequent encoding of the correct answer (e.g., *frog*). The *error correction theory* proposes in contrast that unsuccessful attempts facilitate learning by drawing deeper attention to the error (S. H. Kang et al., 2011). When people generate an incorrect response and then are shown corrective feedback, they will realize the gap between the correct answer and their generation, and then the error-correction process works to strengthen the associative link between the cue and target. The amount of learning is highly dependent on the magnitude of the perceived gap, on this account. These two theories attribute the errorful generation benefit to enhanced elaborative encoding processes. The *additional cue theory* mainly concerns the retrieval process. During memory retrieval, incorrect guesses may function as mediators between the cue and target, which assist retrieval of the correct answer (Soraci et al., 1994; Yan, Yu, Garcia, & Bjork, 2014). For example, when people are shown *pond-?* and asked to produce a response, they may generate a strongly related candidate *water* before they are shown the correct target, *frog*. At a later test, when the cue word, *pond-?* is presented, people may recall *water* first and then recover *frog* as the correct target.

In addition to these three theories, Potts and Shanks (2014) proposed an *attention capturing theory*, which proposes that a greater degree of active engagement, attention and effort is aroused in the Generate condition than in the Read condition. M. J. Kang et al. (2009) used brain imaging to scan participants while they studied trivia questions. Their findings revealed that the levels of activation of memory-related brain regions, as well as recall itself, were positively correlated with ratings of curiosity when participants guessed incorrectly.

Unawareness of the errorful generation benefit

Despite the fact that generating errors followed by corrective feedback enhances memory more effectively than reading, people tend to underestimate the efficacy of generating errors followed by corrective feedback. For instance, Potts and Shanks (2014) asked participants to study foreign word translations (e.g., *igel-frog*) and make item-by-item JOLs after studying each word pair. Across all studies, an advantage of generating errors followed by corrective feedback over reading was observed.

However, lower JOLs were given to incorrectly generated pairs than to Read pairs. Thus participants seemed to be misaligned in the sense that they did not appreciate the errorful generation benefit. Similar findings were observed by Huelser and Metcalfe (2012). In Huelser and Metcalfe's (2012) Experiment 2, participants were asked to study related and unrelated word pairs. Once again, the errorful generation benefit was replicated in the related word pair condition. However, when participants were asked to make retrospective efficacy rankings of the two study methods, lower rankings were given to the Generate method.

Potts and Shanks (2014) proposed that processing fluency plays an important role in this metacognitive misalignment. They hypothesized that participants based their JOLs on *ease of processing* (Koriat & Ma'ayan, 2005). In the Generate condition, the generation process and the encoding of corrective feedback are assumed to be more effortful and less fluent than the encoding process in the Read condition. Although more effort in the Generate condition enhances memory (a "desirable difficulty": Bjork, 1994), people do not appreciate this benefit because the encoding process in the Generate condition is dysfluent. Huelser and Metcalfe (2012), in contrast, proposed that people may hold a bias against believing that errors are beneficial. People's beliefs about the relative efficacy of generating errors followed by corrective feedback and reading may contribute to this unawareness.

Many studies have been conducted to investigate the conditions under which generating errors followed by corrective feedback is more effective than reading and the possible mechanisms underlying the errorful generation benefit (Grimaldi & Karpicke, 2012; Hays, Kornell, & Bjork, 2013; Huelser & Metcalfe, 2012; Knight, Hunter Ball, Brewer, DeWitt, & Marsh, 2012). The mechanisms underlying the metacognitive illusion have not been explored yet. Although there are two possible proposals, no research has yet directly investigated the possible mechanisms. Another main aim of the current study, in addition to studying the relationship between metamemory monitoring and control, was to fill this gap, specifically focusing on the role of people's beliefs.

Correcting metamemory illusions

Because self-regulated learning is becoming increasingly important and the merits of effective study strategies are frequently underestimated, many studies have been designed to correct people's metamemory illusions. Two methods are frequently applied to pursue this aim.

The first method involves informing people about the merits of effective study strategies prior to study, and then asking them to make judgements or choices during or after studying. Yan, Bjork, and Bjork (2016) gave some participants information about the spacing effect on inductive learning prior to study. Participants' task was to study artists' painting styles. Before studying, some participants were told that over 90% of individuals learn better when one artist's paintings are presented interleaved with other artists' paintings than when seeing all paintings by one artist together. Other participants were uninformed. After studying all artists' paintings (half in the blocked condition and half in the interleaved condition), participants were shown new paintings and were asked to judge which artist was responsible for each painting. The interleaved artists' paintings were better classified. Informed participants were more likely to judge the interleaving method to be superior than Uninformed participants. Thus, informing people of the spacing effect prior to study enhanced their willingness to judge interleaving as superior.

The other frequently used method is to ask people to make delayed JOLs. Previous research has revealed that delayed JOLs are more accurate than immediate ones (T. O. Nelson & Dunlosky, 1991). Metcalfe and Finn (2008) found that previous test experience impacts subsequent metamemory monitoring, with higher JOLs attached to previously recalled items and lower JOLs to unrecalled items. In contrast no difference in memory was detected in a later test. In their Experiment 3, some participants were asked to make immediate JOLs and others to make delayed JOLs. In the immediate JOL group, participants again showed this metamemory illusion. However this illusion was eliminated in the delayed JOL group.

Overview of the current experiments

To foreshadow, the first two experiments reveal that participants' item-by-item JOLs failed to reflect the errorful generation benefit. Metacognitive misalignment led participants to distribute more

study resources to incorrectly generated pairs, even though superior recall of these pairs was exhibited in the final test. Experiment 3 was designed, using an online survey, to investigate participants' belief about the relative efficacy of generating errors followed by corrective feedback and reading. Participants believed that errorful generation was inferior to reading. Then, in Experiment 4, we tried to calibrate people's item-by-item metamemory monitoring by informing them of the errorful generation benefit before studying. Although participants' metacognitive awareness was partly improved, item-by-item JOLs were still misaligned. In Experiment 5, we tried to calibrate participants' metamemory reports by using delayed JOLs. In the delayed JOL condition, this metacognitive unawareness was countered.

Experiment 1

In Experiment 1, we examined the relationship between metacognitive monitoring and restudy choices in the context of errors committed during learning. We hypothesized that participants would make relatively accurate JOLs and would prefer to restudy low JOL pairs. However, we also hypothesized that participants would underestimate the efficacy of generating errors followed by corrective feedback relative to the efficacy of reading when making item-by-item JOLs and would prefer to choose incorrectly generated pairs to restudy, even though these pairs would be better remembered.

Participants

Twenty native English speakers were recruited from the UCL participant pool (average age = 25.9, *SD* = 7.73, 13 females). Participants received £4 or course credit as compensation. All participants were debriefed after finishing the experiment.

Materials

The experiment employed the errorful generation paradigm and the same 60 weakly associated word pairs (e.g., *pond-frog*) developed by Kornell et al. (2009). The forward relatedness of these pairs is between 0.050 and 0.054, which means that the probability that people can guess the correct target to a given cue is around 5% (D. L. Nelson, McEvoy, & Schreiber, 1998). The minimum word length is four letters. These pairs were divided into two sets, matched for semantic relatedness. One set was

assigned to the Read condition and the other to the Generate condition. Sets were rotated through conditions across participants.

Design and procedure

Study method (Read/Generate) was manipulated within-subjects. The experiment consisted of 3 stages: encoding, distraction, and final recall. During the initial encoding stage, 60 pairs were randomly presented on screen, one pair at a time. Read and Generate pairs were randomly intermixed. In the Read condition, a cue word and corresponding target were presented together for 13 sec. In the Generate condition, a cue word was presented for 8 sec with a blank box displayed below. Participants were instructed to guess the target and type their guess into the text box. Then, alongside the cue word, the correct target was presented for 5 sec. Participants were told to remember the correct answer rather than their guess. After studying each pair, participants were asked to predict the likelihood they could remember that pair 5 min later. JOLs were made on a slider scale ranging from 0 (I'm sure I won't remember it) to 100 (I will definitely remember it). Next, they decided whether or not they wanted to restudy that pair again. They were informed that if they chose 'YES', they could restudy that pair again for 5 sec after studying all 60 word pairs. Participants had unlimited time to make item-by-item JOLs and restudy decisions.

It is important to emphasize that no pairs were restudied regardless of participants' restudy choices. Following the encoding stage, a 5 min distracting task was administered, in which participants were encouraged to solve as many simple arithmetic problems (e.g., *41+28 =*____) as they could. Then, all 60 cue words were presented one by one in a different random order in the test stage, and participants had unlimited time to recall each target and type it via the keyboard.

Results

JOLs and restudy choices

We analysed the data by using normalized JOLs, dividing each participant's JOLs into six levels, following Son (2004). JOL level 1 consisted of the 10 pairs to which a given participant gave the lowest JOLs and JOL level 6 consisted of the 10 pairs to which that participant gave the highest JOLs. When there were ties at the point demarcating a boundary between JOL levels, pairs were randomly divided into the lower JOL and higher JOL levels. The same method was used in all subsequent experiments.

The proportion of pairs that were selected for restudy at each JOL level is shown in Figure 1A. A repeated measures analysis of variance (ANOVA) was conducted to determine the relationship between JOL level and restudy choice. There was a main effect of JOL level, $F(5, 95) = 36.35$, $p < .01$, $\eta_p^2 = .66$. A within-subjects contrast showed that there was a linear regression of restudy choices across JOL levels, $F(1, 19) = 56.93$, $p < .01$, $\eta_p^2 = .75$. Participants preferred to restudy low JOL pairs and, to this extent, regulated their restudy choices according to their metamemory monitoring.

JOLs and final recall

To examine the relationship between JOL level and final recall, a repeated measures ANOVA was conducted. As shown in Figure 1B, participants' JOLs were relatively accurate. There was a main effect of JOL level, $F(5, 95) = 7.08$, $p < .01$, $\eta_p^2 = .27$. The linear regression of final recall across JOL levels was statistically significant, $F(1, 19) = 10.82$, $p < .01$, $\eta_p^2 = .36$. These results confirm that participants' JOLs were relatively well-calibrated.

Restudy choices and final recall

For each participant, we divided all 60 pairs into two sets according to that participant's final recall, the recalled set (comprising all recalled pairs on the final test) and the unrecalled set (comprising all unrecalled pairs on the final test). One participant recalled all 60 pairs in the final test and this participant's data were removed from this analysis. The proportion of unrecalled pairs participants chose to restudy was significantly higher than that of recalled pairs, difference = 18.1%, 95% confidence interval (CI) [9.85, 26.29] (see Figure 1C). Sixteen participants chose a higher proportion of unrecalled pairs to restudy and two showed the reverse pattern. One participant did not select any pairs to restudy. These results reveal that participants were more likely to restudy unrecalled pairs, suggesting that they controlled their restudy choices in a relatively optimal way. To this extent, participants' restudy choices were related to their actual retention.

Were participants always well-calibrated in metamemory monitoring? Is restudy choice related to metamemory monitoring or actual retention? To explore these two issues, in the following sections we analysed the data with study method as a within-subjects variable.

Initial generation performance

Participants correctly guessed 4.2% (*SD* = 4.31) of pairs in the Generate condition. All correctly guessed pairs were removed from subsequent analyses.

Final recall

As shown in Figure 1D, recall of incorrectly generated pairs was significantly greater than that of Read pairs, difference = 7.1%, 95% CI [2.42, 11.75]. Fourteen out of 20 participants recalled a higher proportion of incorrectly generated pairs and 5 showed the reverse pattern (one recalled all pairs). These results confirm past research showing that generating errors followed by corrective feedback enhances retention more effectively than spending the same amount of time on reading.

JOLs

Participants were unaware of the benefit of errorful generation: They attached significantly higher JOLs to Read pairs than they did to incorrectly generated pairs, difference = 5.34, 95% CI [2.67, 8.01] (see Figure 1D). Sixteen participants attached higher JOLs to Read pairs and four showed the opposite pattern.

Restudy choices

The critical interest of the current experiment was to determine whether participants' metacognitive monitoring or actual retention affected their subsequent restudy choices in the context of errors. As shown in Figure 1D, participants preferred to restudy incorrectly generated pairs rather than Read pairs, difference = 9.3%, 95% CI [2.90, 15.64]. Fifteen participants preferred to restudy incorrectly generated pairs, and four showed the opposite pattern (one participant did not choose any pairs to restudy). These results reveal that participants' erroneous metacognitive assessments of learning, rather than their actual retention, influenced their subsequent restudy choices (Metcalfe & Finn, 2008).

Gamma correlations were calculated between JOLs and final recall, between JOLs and restudy choices, and between restudy choices and final recall for Read pairs, incorrectly generated pairs, and all 60 pairs. These are reported in Table A1 (see Appendix A) and reveal significant levels of resolution at the item level.

Discussion

The experiment successfully replicated Potts and Shanks's (2014) findings that people tend to lack metacognitive awareness of the errorful generation benefit and extended it to a case where the

materials to be learned are familiar cue-target associations, unlike the novel vocabulary items used in the Potts and Shanks study. The main aim of Experiment 1 was to investigate the relationships between metamemory monitoring, restudy choice, and actual recall in the context of errors committed during learning. Participants made relatively accurate JOLs overall: they gave low JOLs to pairs which were less likely to be recalled in the final test and preferred to restudy low JOL pairs. In addition they preferred to restudy subsequently unrecalled pairs. When we analysed the data with study method as a within-subjects variable, however, the results showed that participants gave lower JOLs to incorrectly generated pairs, even though these pairs were better recalled in the final test. Participants were also more likely to restudy these pairs. Experiment 1 hence provides novel and striking evidence that restudy choice is related to metamemory monitoring rather than actual retention and is the first study demonstrating that a metacognitive illusion can induce a preference for restudying better- over worse-remembered materials.

Experiment 2

Kornell and Bjork (2007) found that, although learners are often rational in distributing their study time, study time allocation is not always optimal. Only one study has investigated the effect of errorful generation on study time allocation. In Potts and Shanks (2014)'s Experiments 3, participants were allowed to spend as much time as they wanted to study each foreign word. In the Generate condition less time was spent on encoding the correct answer than the encoding time in the Read condition. Potts and Shanks (2014) proposed that participants encoded correct answers more effectively in the Generate condition. Going beyond Potts and Shanks' (2014) Experiment 3, in our Experiment 2, we explored the effect of metacognitive unawareness of the errorful generation benefit on restudy time allocation.

In Experiment 2, participants were given the same amount of time to study Read and Generate pairs in the initial encoding stage. In the restudy stage, all pairs were presented under the Read condition and participants were allowed to restudy all pairs in a self-paced procedure. In this way, we could directly measure the effect of metamemory illusions about the errorful generation benefit on restudy time allocation. In addition, we asked whether this metamemory illusion was long-lasting by adding a short time delay between making JOLs and assessing restudy time allocation.

Participants

Twenty native speakers were recruited from the UCL participant pool (average age = 24.40, *SD* = 6.51, 9 females). Participants received £5 or course credit as compensation. They were debriefed after finishing the experiment.

Materials, design, and procedure

The same materials, experimental design, and procedure were used as in Experiment 1 with the following exceptions. In Experiment 2, participants did not make restudy choices. After studying all 60 pairs and making item-by-item JOLs, a distractor task (arithmetic problem solving for 1 min) was administered. Subsequently, participants were instructed to restudy all pairs under the Read condition (a cue word and its target were presented alongside each other) in a self-paced procedure.

In Experiment 1, about 80% of Read pairs and 90% of incorrectly generated pairs were recalled in the final test. In Experiment 2, to prevent a ceiling effect, a 24 hour delay was implemented between study and the final test.

Results

JOLs and restudy time allocation

We first determined the relationship between metamemory monitoring and restudy time allocation. As in Experiment 1, we analysed the data by using normalized JOLs. Figure 2A shows that restudy time decreased with increasing JOLs. A repeated measures ANOVA was conducted to determine the relationship between JOL level and restudy time allocation. The assumption of sphericity was not met, $\chi^2(14) = 68.69$, $p < .01$, and hence we applied the Huynh-Feldt correction. There was a main effect of JOL level, $F(1.95, 37.14) = 15.88$, $p < .01$, $\eta_p^2 = .45$. A within-subjects contrast showed that the linear regression of restudy time across JOL levels was statistically significant, $F(1, 19) = 21.05$, $p < .01$, $\eta_p^2 = .53$.

Initial generation performance

Participants correctly guessed 3.7% (*SD* =3.40) of pairs in the Generate condition. These pairs were removed from all subsequent analyses.

Final recall

As shown in Figure 2B, final recall of incorrectly generated pairs was significantly better than that of Read pairs, difference = 9.1%, 95% CI [4.58, 13.55]. Sixteen participants recalled a higher proportion of incorrectly generated pairs and four showed the reverse pattern.

JOLs

Consistent with Experiment 1, participants gave higher JOLs to Read pairs, difference = 4.71, 95% CI [1.77, 7.67] (see Figure 2C). Sixteen participants gave higher JOLs to Read pairs while four gave higher JOLs to incorrectly generated pairs.

Restudy time allocation

The critical interest of Experiment 2 was to determine whether metacognitive illusions directly guided subsequent restudy time allocation. Despite the fact that incorrectly generated pairs were better recalled than Read ones, participants spent more time restudying incorrectly generated pairs than Read pairs, difference = 903ms, 95% CI [506.45, 1299.85] (see Figure 2D). Eighteen participants allocated more time to restudying incorrectly generated pairs and two showed the reverse pattern. These results reveal that participants' assessments of learning, instead of their actual learning status, guided their restudy time allocation (Finn, 2008; Metcalfe & Finn, 2008).

Gamma correlations were calculated between JOLs and final recall, and between restudy time and final recall for Read pairs, incorrectly generated pairs, and all 60 pairs. Pearson correlations were calculated between JOLs and restudy time. The correlations are reported in Table A2 (see Appendix A) and again reveal significant levels of resolution at the item level.

Discussion

Participants allocated study time according to their metamemory monitoring. When we included study method as a within-subjects variable, the results indicate that participants gave lower JOLs to incorrectly generated pairs and more restudy time was allocated to these pairs. However in the final test, incorrectly generated pairs were better recalled than Read pairs. The superior final recall of incorrectly generated pairs may be partially due to the benefit of generating errors followed by corrective feedback. Another possible cause is that participants spent more time restudying incorrectly generated pairs in the restudy phase.

Experiment 2's results show that people's erroneous metamemory monitoring leads them to spend more time restudying better remembered information. This illusion's effect on metamemory control is not limited to immediate study resource allocation. Experiment 2 is the first study to observe a situation under which participants allocated more restudy time to better remembered information because of inaccurate metamemory monitoring and the first to show that this metamemory illusion's effect on study time allocation can persist after a short time delay.

Experiment 3

In the first two experiments, participants' item-by-item JOLs failed to reflect the benefit of errorful generation. The mechanisms underlying the unawareness of the errorful generation benefit are still unclear. In the first two experiments, JOLs for correctly generated pairs were removed from data analysis. Presumably the removed pairs (correctly generated pairs) were more likely to come to mind in response to the cue because they represented a subset of pairs that were more closely related for a given individual. People may base their JOLs on the relatedness of word pairs, with higher JOLs to more related pairs and lower JOLs to less related pairs (Koriat & Bjork, 2005; Mueller, Tauber, & Dunlosky, 2013). Although the proportions of items removed were small, this may mean that the incorrectly-generated pairs remaining in the analysis were less semantically related than Read pairs, via an item-selection effect. To determine whether there was any difference in semantic relatedness between incorrectly generated and Read pairs, 16 new participants were asked to rate the semantic relatedness of the cue-target pairs employed in Experiments 1 and 2 after studying each pair. No difference in semantic relatedness ratings was observed between incorrectly generated pairs and Read pairs (see Appendix B for details). Therefore, it seems unlikely that unawareness of the errorful generation benefit can be attributed to perceived differences in relatedness.

Huelser and Metcalfe (2012) proposed that people may hold a bias against believing that errors are beneficial. Accordingly, one possible mechanism underlying the metacognitive unawareness observed in Experiments 1 and 2 is that people's explicit beliefs about the relative learning efficacy of generating errors followed by corrective feedback versus reading drove their JOLs. To our knowledge, no study has yet solicited people's beliefs explicitly. One previous study implies that people may hold beliefs that reading is better than generating. Participants in Froger, Sacher, Gaudouen, Isingrini, and

Taconnat (2011) study predicted that reading would be more effective than generating for learning a future list, but Froger et al. used materials that were designed to elicit correct generations (e.g., *door-win___*?) and, more importantly, collected predictions after participants had actually experienced each learning condition, in which case they may have simply been reporting their experience of the efficacy of each method. People's beliefs about the relative learning efficacy of generating errors followed by corrective feedback versus reading are therefore largely unknown. Experiment 3 was designed to explore this issue.

Participants

One hundred participants, 44 females, were recruited online from Prolific Academic (https://www.prolific.ac/). Their ages ranged from 18 to 60, average age 27.70 (*SD* = 6.33). All participants' first language was English and all of them lived in the United Kingdom. Participants received £0.40 as compensation. The survey took about 5 min.

Materials, design, and procedure

The instructions and questions used in Experiment 3 are attached in Appendix C. Participants were instructed to read the instructions carefully and were told that a test on these instructions would be administered later to check whether they had completely understood them. The instructions explained the aim of the survey and contained full descriptions of the two study methods. Participants were asked to imagine that they would study 60 English word pairs (30 in the Read condition and 30 in the Generate condition). They were informed that the likelihood they would guess correctly in the Generate condition was about 5%. Following the instructions, a short test on the instructions was administered to assess participants' comprehension. The order of response options for each question was randomized. For each question, if an incorrect choice was selected, corrective feedback was provided.

The questionnaire consisted of five questions on two pages. On the first page, participants were asked to choose which method they thought was the more effective way to learn English word pairings: Generate or Read. On the same page, they were asked to estimate the proportion of Read and

Generate pairs they would remember in 24 hours. On the next page, participants were instructed to choose which the more effective method was, Generate (incorrectly) or Read, and then estimate the proportion of incorrectly generated pairs they would remember. All instructions and questions were presented using Qualtrics (http://www.qualtrics.com/).

Results

Education level

8.0% of participants' highest education level was secondary school/GCSE, 34.0% was college/A level, 42.0% was undergraduate degree, 14.0% was graduate level, and 2.0% was doctorate degree.

Instruction test performance

Of all participants, 37.0% answered all five instruction questions correctly, 25.0% answered four instruction questions correctly, 22.0% answered three questions correctly, 11.0% answered two questions correctly, and 5.0% answered one question correctly.

Beliefs about the relative efficacy of generating errors followed by corrective feedback and reading

Of all participants, 65.0% chose Read as the more effective method rather than Generate (see Figure 3A), which is significantly different from chance (50%), $\chi^2(1) = 4.60$, $p = .03$. Of the participants who correctly answered all instruction questions, the result was similar: 64.9% chose Read, although this was not statistically different from chance, $\chi^2(1) = 1.69$, $p = .19$.

Next we analysed their predictions about the proportion of pairs they would remember that were studied in each condition (see Figure 3B). Participants predicted that they would remember a higher proportion of Read than Generate pairs, difference = 8.76, 95% CI [3.90, 13.62]. 55.0% of participants gave higher predictions to Read pairs, 36.0% showed the reverse pattern, and the remaining 9.0% gave equal predictions. Of participants who answered all the instruction questions correctly, they also predicted they would remember more Read than Generate pairs, difference = 8.89,

95% CI [.29, 17.49]. 56.8 % gave higher predictions to Read pairs, 40.5% showed the reverse pattern and the other 2.7% gave the same predictions to these two methods.

The critical concern of the current study was to determine people's beliefs about the relative efficacy between reading and generating errors followed by corrective feedback. Across all participants, 78.0% chose Read as the more effective method compared with Generate (incorrectly), see Figure 3A. This proportion is significantly different from 50%, $\chi^2(1) = 17.01$, $p < .01$. Of participants who answered all instruction questions correctly, 70.3% chose the Read method, $\chi^2(1) = 3.21$, $p = .07$.

Participants gave higher predictions to Read than they did to incorrectly generated pairs (see Figure 3B), difference = 14.79, 95% CI [9.12, 20.45]. 59.0% gave higher predictions to Read pairs, 30.0% showed the reverse pattern, and the remaining 11.0% gave the same predictions to incorrectly generated and Read pairs. Participants who answered all instruction questions correctly showed the same pattern, difference = 14.18, 95% CI [3.83, 24.55]. 54.1% gave higher predictions to Read pairs, 37.8% showed the reverse pattern, and the remaining 8.1% showed no difference in predictions of incorrectly generated and Read pairs.

Discussion

The results show clearly that a majority of people believe that generating errors followed by corrective feedback is inferior to reading for learning English word pairs. This is evident both in choices, where reading was rated more effective than generation in general as well as incorrect generation, and in terms of the proportion of items participants believed they would be able to recall under each encoding format.

Experiment 4

In Experiment 3, the survey results showed that people tend to believe that learning via errorful generation followed by corrective feedback is inferior to reading. In Experiment 4, we tried to counter this metacognitive unawareness by informing people of the errorful generation benefit before studying.

Participants

Forty native English speakers were recruited from the UCL participant pool (average age = 20.80, *SD* = 3.55, 32 females). Participants were randomly divided into two groups (Uninformed/Informed). Participants received £5 or course credit as compensation. All participants were debriefed after finishing the experiment.

Materials, design, and procedure

Except where noted, this experiment is identical to Experiment 1. In the Informed group, participants first read instructions about the benefit of errorful generation. Three possible underlying mechanisms were included in the instructions. Participants read the instructions in a self-paced procedure. To ensure that they understood the instructions completely, a multiple choice test was applied after they finished reading them (see the details of instructions and multiple choice test questions in Appendix D). Participants were allowed to review the instructions during the instruction test. After they answered all instruction questions, an experimenter checked their answers. If they answered some questions incorrectly, the experimenter highlighted the questions they had answered incorrectly and told them to review the instructions to find the correct answers. Only when they had answered all questions correctly could they proceed to the main experiment. In the Uninformed group, participants did not read these instructions and did not take the instruction test.

All participants were asked to make item-by-item JOLs and restudy choices after studying each pair. After studying all 60 pairs, they were also asked to make aggregate JOLs to Read and incorrectly generated pairs on a slider from 0 ("I won't remember any pairs") to 100 ("I will remember every pair"). The final test took place 24 hours after the study phase.

Results

Initial generation performance

2.3% (*SD* = 2.44) of pairs in the Generate condition were correctly guessed in the Uninformed group and 3.3% (*SD* = 3.06) in the Informed group. The difference between the two groups' generation performance was not statistically significant, difference = -1.0%, 95% CI [-2.77, .71]. These pairs were removed from the following analyses.

Final recall

A repeated measures ANOVA, with study method as a within-subjects variable and group as a between-subjects variable, revealed only a main effect of study method, $F(1, 38) = 31.62$, $p < .01$, $\eta_p^2 = .45$. There was no main effect of group, $F(1, 38) = .10$, $p = .76$, and no interaction between study method and group, $F(1, 38) = .33$, $p = .57$. As can be seen in Figures 4A and 4B, for both groups, a lower proportion of Read pairs were recalled than incorrectly generated pairs (Uninformed group: difference = -10.8%, 95% CI [-16.35, -5.31]; Informed group: difference = -8.8%, 95% CI [-13.64, -4.02]). In the Uninformed group, eighteen out of twenty participants recalled a higher proportion of incorrectly generated pairs, and one participant showed the reverse pattern (there was one tie). In the Informed group, fifteen participants recalled a higher proportion of incorrectly pairs, two showed the reverse pattern, and there were three ties.

Item-by-item JOLs

Average item-by-item JOLs for incorrectly generated and Read pairs for both groups are presented in Figures 4A and 4B. A repeated measures ANOVA, with study method as a within-subjects variable and group as a between-subjects variable, revealed a main effect of study method, $F(1, 38) = 32.72$, $p < .01$, $\eta_p^2 = .46$, but no main effect of group, $F(1, 38) = 2.51$, $p = .12$. There was a significant interaction between study method and group, $F(1, 38) = 9.19$, $p < .01$, $\eta_p^2 = .20$. Pairwise comparisons indicated that, in the Uninformed group, higher item-by-item JOLs were given to Read pairs, difference = 7.55, 95% CI [4.71, 10.40]. Nineteen participants gave higher item-by-item JOLs to Read pairs and one showed the reverse pattern. In the Informed group, higher item-by-item JOLs were also given to Read pairs, but to a lesser extent, difference = 2.32, 95% CI [.10, 4.55]. Thirteen participants gave higher item-by-item JOLs to Read pairs and seven showed the reverse pattern. The interaction between study method and group indicates that explicitly telling participants about the benefit of errorful generation partly ameliorated their metacognitive unawareness, but did not eliminate it entirely, let alone reverse it.

Restudy choices

A repeated measures ANOVA with study method as a within-subjects variable and group as a between-subjects variable showed a main effect of study method, $F(1, 38) = 17.88$, $p < .01$, $\eta_p^2 = .32$,

but no main effect of group, $F(1, 38) = .03$, $p = .87$. The group main effect was qualified by an interaction between study method and group, $F(1, 38) = 4.57$, $p = .04$, $\eta_p^2 = .11$. Pairwise comparisons indicated that, in the Uninformed group, participants selected a lower proportion of Read pairs to restudy, difference = -11.4%, 95% CI [-17.87, -4.87]. Fourteen participants preferred to restudy incorrectly generated pairs, two showed the reverse pattern, there was one tie, and three did not choose any pairs to restudy. In the Informed group, similarly, a lower proportion of Read pairs were selected for restudy, difference = -3.7%, 95% CI [-7.44, -.03]. Eleven participants preferred to restudy incorrectly generated pairs, and two showed the reverse pattern. Four did not choose any pairs to restudy and three chose all pairs to restudy.

Aggregate JOLs

Averages of aggregate JOLs for incorrectly generated and Read pairs for both groups are presented in Figures 4A and 4B. A repeated measures ANOVA, with study method as a within-subjects variable and group as a between-subjects variable, revealed no main effect of study method, $F(1, 38) = .57$, $p = .45$, and no main effect of group, $F(1, 38) = 2.55$, $p = .12$. There was however a significant interaction between study method and group, $F(1, 38) = 18.50$, $p < .01$, $\eta_p^2 = .33$. Pairwise comparisons indicated that, in the Uninformed group, higher aggregate JOLs were made to Read pairs, difference = 10.70, 95% CI [3.55, 17.85]. Sixteen participants gave higher aggregate JOLs to Read pairs and three showed the reverse pattern (there was one tie). In stark contrast, in the Informed group, lower aggregate JOLs were given to Read pairs, difference = -7.50, 95% CI [-12.73, -2.27]. Sixteen participants reported lower aggregate JOLs to Read pairs and four showed the reverse pattern. The 16 ‘believers’, who gave lower aggregate JOLs to Read pairs than to incorrectly generated pairs in the Informed group, also gave marginally lower item-by-item JOLs to incorrectly generated pairs than to Read pairs, difference = 1.65, 95% CI [-.03, 3.32], and preferred to restudy incorrectly generated pairs over Read pairs, difference = 5.0%, 95% CI [.11, 9.82] .

Gamma correlations were calculated between JOLs and final recall, between JOLs and restudy choices, and between restudy choices and final recall for Read pairs, incorrectly generated pairs, and all 60 pairs for both groups (see Table A3 in Appendix A). In addition, the mean difference between the two groups’ gamma values is reported also.

Discussion

The results from the Uninformed group fully replicate those from Experiment 1, extended to a situation in which retention is tested after 24h. In the Informed group, participants' aggregate JOLs revealed that they recognized the positive benefit of generating errors followed by corrective feedback and their metacognitive awareness was better aligned with their true recall. However, their item-by-item JOLs still reflected metacognitive unawareness, albeit to a reduced level. Although the instructions altered the distribution of Informed participants' restudy choices, they still preferred overall to restudy incorrectly generated than Read pairs.

Two possible reasons may account for this pattern. The first is that participants' beliefs may not be the only source of their lack of awareness. For example, item-by-item processing fluency could be another possible source. It is reasonable to assume that in the Generate condition, the encoding process was less fluent because of the demands of generating a response (Froger et al., 2011; Potts & Shanks, 2014). Another possible reason is that Informed participants did not know to what extent they should adjust their item-by-item JOLs to reflect this benefit. Although the instructions informed participants of the errorful generation benefit, the instructions did not say anything about the magnitude of this benefit.

Experiment 5

In Experiments 1, 2, and 4, item-by-item JOLs failed to reflect the errorful generation benefit and, counterproductively, more study resources were allocated to incorrectly generated pairs. In Experiment 4's Informed group, even when participants were informed of the benefit before study, they still tended to give lower item-by-item JOLs to incorrectly generated pairs. In Experiment 5, our first aim was to find another method for bringing metacognition into line with actual memory performance.

Previous research has found that delayed JOLs are more accurate than immediate ones, because the former are based on people's attempts to retrieve information from memory (T. O. Nelson & Dunlosky, 1991). Therefore, in Experiment 5, we hypothesised that in a delayed JOL condition, higher JOLs would be attached to incorrectly generated pairs.

In our Experiments 1, 2, and 4, the fact that participants allocated more restudy resources to incorrectly generated pairs might be a consequence of their lower confidence that they would remember

incorrectly generated pairs, as reflected in their lower JOLs. Another possible reason might concern participants' explicit beliefs. Participants may have allocated more restudy resources to incorrectly generated pairs because they believed that generating errors followed by corrective feedback is inferior to reading. Therefore, it is still unclear whether metamemory control is related to metamemory monitoring or to people's belief. Another aim of Experiment 5 is to explore this issue.

Participants

Thirty two native English speakers were recruited from the UCL participant pool (average age = 23.41, *SD* = 4.16, 21 females) and were randomly divided into two groups (immediate JOL/ delayed JOL). Participants received £5 or course credit as compensation. All participants were debriefed after finishing the experiment.

Materials, design and procedure

Except where noted, this experiment is identical to Experiment 1. The same 60 word pairs were used as in the previous experiments. A 2 (study method: Read/Generate) × 2 (JOL type: immediate JOL/delayed JOL) mixed factorial design was implemented, with study method as a within-subjects variable and JOL type as a between-subjects variable.

In the immediate JOL group, after encoding each pair, participants were asked to estimate the likelihood that they would remember that pair in 24 hours, and then make a restudy decision. In the delayed JOL group, the procedure was similar with the following exceptions. In the study phase, participants did not make item-by-item JOLs and did not make restudy decisions following encoding of each pair. Instead, after they studied all 60 pairs, all 60 cue words were randomly presented one by one (target words were omitted) and participants were asked to make item-by-item delayed JOLs and restudy decisions. Participants had no opportunity to restudy any pairs regardless of their restudy choices. The final test took place 24 hours later.

Results

Initial generation performance

3.5% (*SD* = 2.85) of pairs in the Generate condition were correctly guessed in the immediate JOL group and 3.5% (*SD* = 4.12) in the delayed JOL group. There was no difference in generation performance between groups. These pairs were removed from the following analyses.

Final recall

A repeated measures ANOVA, with study method as a within-subjects variable and JOL type as a between-subjects variable, showed only a main effect of study method, $F(1, 22) = 54.33$, $p < .01$, $\eta_p^2 = .64$. There was no main effect of JOL type, $F(1, 22) = 1.21$, $p = .28$, and no interaction between study method and JOL type, $F(1, 22) = .09$, $p = .77$. As shown in Figure 5A, in both groups incorrectly generated pairs were better recalled than Read pairs: immediate JOL group, difference = 17.5%, 95% CI [9.87, 25.13]; delayed JOL group, difference = 19.0%, 95% CI [11.69, 26.23]. In the immediate JOL group, 15 participants recalled a higher proportion of incorrectly generated pairs and there was one tie. In the delayed JOL group, 14 participants recalled a higher proportion of incorrectly generated pairs and two showed the reverse pattern.

JOLs

Average JOLs for incorrectly generated and Read pairs for both groups are shown in Figure 5B. A repeated measures ANOVA, with study method as a within-subjects variable and JOL type as a between-subjects variable, showed a main effect of study method, $F(1, 30) = 4.58$, $p = .04$, $\eta_p^2 = .13$, but no main effect of JOL type, $F(1, 30) = 2.56$, $p = .11$. These effects were moderated by a significant interaction between study method and JOL type, $F(1, 30) = 33.98$, $p < .01$, $\eta_p^2 = .53$. Pairwise comparisons indicated that, in the immediate JOL group, higher JOLs were made to Read pairs, difference = 5.82, 95% CI [2.37, 9.28]. Fourteen participants gave higher JOLs to Read pairs and two showed the reverse pattern. In contrast, in the delayed JOL group, lower JOLs were given to Read pairs, difference = -12.58, 95% CI [-18.36, -6.81]. Fifteen participants gave higher JOLs to incorrectly generated pairs and only one participant gave higher JOLs to Read pairs. The interaction between study method and JOL type indicates that metacognitive unawareness of the benefit of errorful generation is reversed by replacing immediate JOLs with delayed ones.

Restudy choices

Figure 5C depicts participants' restudy choices. A repeated measures ANOVA, with study method as a within-subjects variable and group as a between-subjects variable, revealed that there was no overall main effect of study method, $F(1, 30) = 3.12$, $p = .09$, and no main effect of JOL type, $F(1, 30) = .78$, $p = .39$, whereas the interaction between study method and JOL type was statistically

significant, $F(1, 30) = 43.67$, $p < .01$, $\eta_p^2 = .59$. In the immediate JOL group, a higher proportion of incorrectly generated pairs were selected for restudy, difference = 10.0%, 95% CI [4.61, 15.39]. Twelve participants chose a higher proportion of incorrectly generated pairs to restudy and one showed the reverse pattern (one participant did not choose any pairs and one chose every pair to restudy). In contrast, participants in the delayed JOL group preferred to restudy Read pairs rather than incorrectly generated pairs, difference = 17.9%, 95% CI [10.33, 24.25]. Fifteen participants preferred to restudy Read pairs and only one participant chose to restudy a higher proportion of incorrectly generated pairs. These results provide convincing evidence that people's restudy choices are related to their metacognitive assessments of learning, rather than actual retention.

Gamma correlations were calculated between JOLs and final recall, JOLs and restudy choices, and restudy choices and final recall for Read pairs, incorrectly generated pairs, and all 60 pairs for both groups (see Table A4 in Appendix A). These reveal significant levels of resolution at the item level. In addition, the mean difference between the two groups' gamma values is also reported.

Discussion

In the immediate JOL group, the findings of Experiment 1 were once again replicated. Participants gave lower JOLs to incorrectly generated pairs and preferred to restudy these pairs, and in the final test, better recall of incorrectly generated pairs was observed. In the delayed JOL group, participants' item-by-item JOLs were successfully calibrated and restudy decisions were aligned with these JOLs: they preferred to restudy Read pairs.

These results indicate that, under these conditions, people's metamemory monitoring rather than their explicit beliefs or actual retention guides metamemory control. Forcing participants to make delayed instead of immediate JOLs rendered participants' JOLs better tuned to the true level of memory strength. The sensitivity of delayed JOLs to the errorful generation benefit does not imply that participants in our Experiment 5's delayed JOL group knew that generating errors followed by corrective feedback was more effective than reading, because they might not have explicitly attributed their increased confidence in memory for incorrectly generated pairs to the errorful generation strategy. Future research could ask participants to rate the learning efficacies of generating errors

followed by corrective feedback versus reading after giving delayed JOLs to explore whether they attribute increased JOLs to the errorful generation strategy. Our findings show that it is possible to reverse the pattern of participants' JOLs and restudy decisions for read and incorrectly generated pairs, not by addressing an explicit belief but simply by presenting the cues again after a delay and having participants reflect on their state of learning of those items.

In Experiment 2, when participants were presented with cue-target pairs following initial study, they chose to allocate more restudy time to items for which they had generated an incorrect response at study, despite these being better remembered at final test: Presenting intact cue-target pairs did not alter participants' suboptimal study strategies. By contrast, presenting cues alone in Experiment 5 reversed the misalignment of metacognition with actual learning.

General discussion

The first aim of the current study was to test the reliability and reproducibility of metacognitive unawareness of the errorful generation benefit by using weakly associated pairs. The errorful generation benefit was observed across Experiments 1, 2, 4, and 5. However, participants' immediate item-by-item JOLs showed unawareness of this benefit, with lower JOLs attached to incorrectly generated pairs and higher JOLs to Read pairs.

Unawareness of the errorful generation benefit

The second aim of the current study was to examine possible mechanisms underlying this metacognitive unawareness. In the current study, correctly generated pairs were removed from the data analysis. The correctly generated pairs were presumably more likely to come to mind in response to a cue because they represented a subset of pairs that were more closely related for a given individual. However, there was no statistically significant difference in semantic relatedness ratings between Read and incorrectly generated pairs. Therefore, this unawareness cannot be attributed to an item-selection effect. Other supporting evidence comes from the fact that JOLs for Read pairs were significantly higher than for Generate pairs (including both correctly and incorrectly generated pairs) across Experiments 1 (difference = 4.59, 95% CI [1.99, 7.19]), 2 (difference = 3.78, 95% CI [.44, 7.12]), 4 (uninformed group: difference = 6.36, 95% CI [3.48, 9.24]), and 5 (immediate JOL group: difference = 5.14, 95% CI [2.11,

8.17]), indicating that it is the study method rather than any item-selection effect that determines this metacognitive illusion.

People's intrinsic beliefs about the relative efficacy of errorful generation and reading may be a potential source of this unawareness. Experiment 3 shows that people do tend to believe that generating errors followed by corrective feedback is inferior to reading for learning. Intrinsic beliefs about the negative effects of committing errors then bias people's item-by-item JOLs. We hypothesize that participants believe that their incorrectly generated response may interfere with correct target recall at later test. To test this idea, we divided incorrectly generated pairs into two types: omission pairs (to which a participant did not generate a response in the permitted time window) and commission pairs (to which a participant generated an incorrect response). We computed a Gamma correlation between JOLs and error types across Experiments 1, 2, 4 (Uninformed group), and 5 (immediate JOL group). There were 2 participants in Experiment 1, 4 participants in Experiment 2, and 7 participants in Experiment 4's Uninformed group who made commission errors but no omission errors, and there was one participant in Experiment 2 and one participant in Experiment 4's Uninformed group who made omission errors but no commission errors. These participants' data were removed from this analysis.

There were negative Gamma correlations between error type and JOLs across experiments (Rhodes & Castel, 2008): significant in Experiment 1: r = -.39, p = .002, Experiment 2: r = -.48, p = .007, Experiment 4's Uninformed group: r = -.53, p = .026, and marginally significant in Experiment 5's immediate JOL group: r = -.23, p = .09. These results reveal that participants gave significantly lower JOLs to commission pairs than to omission pairs, supporting our assumption that people's concern about interference of an incorrect generation with memory for the correct target at a later test partially contributes to this metacognitive illusion. We should be cautious about drawing conclusions from these results because only about 17% of incorrectly generated pairs were omission pairs, the majority being commission pairs.

In Experiment 4, participants in the Informed group did accept guidance that generating errors followed by corrective feedback is more effective than reading for learning related pairs, as revealed by

their aggregate JOLs. Informing participants of the benefit of generating errors followed by corrective feedback prior to study partially alleviated the metacognitive unawareness, but participants still tended to adjust their item-by-item JOLs insufficiently, and still preferred to restudy incorrectly-generated than Read items (albeit to an attenuated degree). One possible reason is that, although participants in the Informed group held the belief that generating errors followed by corrective feedback is more effective, they did not know to what extent to adjust their item-by-item JOLs to reflect this benefit. No specific magnitude of the benefit of generating errors followed by corrective feedback over reading was mentioned in the instructions.

Another possible reason is that people's beliefs may not be the only mechanism underlying this metacognitive unawareness. Of participants in Experiments 1, 2, 4 (Uninformed group), and 5 (immediate JOL group), 85.5% gave higher immediate item-by-item JOLs to Read pairs, which was numerically (although not significantly) larger than the proportion of participants in Experiment 3 who chose Read rather than Generate (incorrectly) as the more effective study method (78%), $\chi^2(1) = 1.61$, $p = .21$. In Experiment 3, 59.0% of participants predicted that they would remember a higher proportion of Read pairs over incorrectly generated pairs, which is significantly lower than 85.5%, $\chi^2(1) = 14.96$, $p < .01$. These results reveal that a higher proportion of participants gave lower immediate item-by-item JOLs to incorrectly generated pairs than the proportion of participants who held the belief that incorrectly generated pairs would be less likely to be remembered. More convincing evidence comes from Experiment 4's Informed group. 'Believers', who gave higher aggregate JOLs to Incorrectly generated pairs over Read pairs, still gave higher item-by-item JOLs to Read pairs over incorrectly generated pairs, indicating that, even when people's beliefs change, they still show this metacognitive illusion. Therefore, people's prior beliefs cannot be the only mechanism underlying this metacognitive unawareness. Other possible mechanisms are further discussed below.

One possibility is that people's experience of generation performance could affect their later JOLs. According to the *memory of past test theory* (MPT), if people answer a question correctly on a previous test, a high JOL will be assigned to that question; correspondingly, if they fail to answer correctly, a low JOL will be given to that question (Finn & Metcalfe, 2007, 2008). Kornell and

Rhodes (2013) explored the effect of feedback on JOLs and found that people tend to discount subsequent learning from feedback. They claimed that memory of a past test anchors participants' JOLs and leads to underestimation of subsequent learning from feedback. In the current study, the experience of incorrect generation may have led participants to underestimate the positive effect of generating errors followed by corrective feedback relative to reading. According to the *anchoring hypothesis*, this metacognitive unawareness effect may be attributed to people's inadequate adjustments of JOLs from anchors set by themselves (England & Serra, 2012; Scheck, Meeter, & Nelson, 2004; Scheck & Nelson, 2005). People overestimate their learning status when an anchor exceeds actual memory outcomes, and, if an anchor is lower than real memory outcomes, underestimation of studying status emerges. Low anchors might be attached to incorrectly generated pairs in the initial generation phase and participants' adjustments away from these low anchors during learning were inadequate to tune their JOLs in line with actual retention.

To test this, we separated correctly recalled pairs in the Generate condition into two sets: correctly generated and incorrectly generated. Eight participants in Experiment 1, 5 in Experiment 2, 9 in Experiment 4's Uninformed group, and 5 in Experiment 5's immediate JOL group did not generate any correct targets. Their data were removed from this data analysis. Although the pairs in both sets were correctly recalled in the final test, participants gave higher JOLs to pairs correctly generated at study than to pairs incorrectly generated at study: Experiment 1: difference = 23.95, 95% CI [18.26, 29.63]; Experiment 4's Uninformed group: difference = 18.53, 95% CI [9.19, 27.87]; Experiment 5's immediate JOL group: difference = 18.48, 95% CI [6.41, 30.52]. These results show that participants' experience of their generation performance affected their judgements of learning. We should be cautious about this conclusion because there was only a small proportion (about 3%) of pairs in the Generate condition that were correctly generated. Similarly, Potts and Shanks (2014) found that participants gave much higher JOLs to items they had correctly selected when choosing from several options at study than to items for which they had made an incorrect choice, even though the stimuli were novel vocabulary items and responses at study could only be guesses.

According to the *fluency effect*, people's metamemory can be influenced by the *ease of encoding* (Besken & Mulligan, 2013; Hertzog, Dunlosky, Robinson, & Kidder, 2003; Rhodes & Castel, 2008; Undorf & Erdfelder, 2011). Generating errors followed by corrective feedback might be an elaborative and strengthening process between a cue and target (Huelser & Metcalfe, 2012). Following incorrect generation, a more elaborative and deeper encoding will take place, which is beneficial for future retrieval (Craik & Lockhart, 1972). However, generating is less fluent than reading, especially when errors are committed frequently (Potts & Shanks, 2014), which leads to lower JOLs being attached to incorrectly generated pairs. To determine the role of the fluency effect in this metacognitive illusion, we calculated correlations between the generation time of incorrectly generated pairs (time interval to typing the first letter) and corresponding JOLs. Only commission pairs could be included in this analysis. There was one participant in Experiment 2 and one in Experiment 4's Uninformed group who did not generate any responses (all errors were omissions). Therefore, their data were omitted from this analysis. For each participant, we calculated an *r* value, and then transformed it to a Fisher *z* score. The estimated mean *z* score was then transformed back to an *r* value (Silver & Dunlap, 1987).

There were negative correlations between generation time and JOLs across Experiments 1, 2, 4 (Uninformed group), and 5 (immediate JOL group), but none of them was significant: Experiment 1: $r = -.08$, $p = .14$; Experiment 2: $r = -.08$, $p = .13$; Experiment 4's Uninformed group: $r = -.06$, $p = .31$; Experiment 5's immediate JOL group: $r = -.06$, $p = .18$. To increase power, we collapsed data across experiments, revealing a significantly negative correlation, $r = -.07$, $p = .007$, indicating that the longer the generation time, the lower the JOL given to that pair. These results provide evidence supporting the fluency effect on this metamemory illusion, although the effect is fairly weak.

Overall, our data suggest that beliefs, generation fluency, experience of generation performance, and concern that incorrectly generated items will interfere with memory for the correct answer, all contribute to metacognitive unawareness of the errorful generation benefit.

Reversing unawareness of the errorful generation benefit

The third aim of the current study was to find a way to overcome this instance of metacognitive unawareness. In Experiment 4, modifying people's beliefs about the relative efficacy of generating errors followed by corrective feedback versus reading partially alleviated this unawareness, but this method was not strong enough to totally counter it. In Experiment 5, in the immediate JOL group, lower JOLs were attached to incorrectly generated pairs. However, in the delayed JOL group, the pattern of item-by-item metacognitive judgements was reversed. Delayed JOLs are more accurate for assessing retention status (Finn & Metcalfe, 2008; T. O. Nelson & Dunlosky, 1991; Scheck et al., 2004). When making delayed JOLs, people try to retrieve information from their memory. Making delayed JOLs is an effective way to reverse the pattern of item-by-item metacognitive judgements and leads to better restudy decisions.

Metamemory monitoring and control

The fourth and primary aim of the current study was to explore a situation under which people give lower JOLs to better remembered information and select more of the better remembered items to restudy or spend longer restudying better remembered items. The current study employed the errorful generation paradigm to investigate the effect of metamemory illusions on learning management. Participants' self-regulated learning (study time allocation and restudy choices) was closely related to their metamemory monitoring. In Experiments 1, 2, and 4, and in the immediate JOL group of Experiment 5, lower JOLs were made to incorrectly generated pairs, but final recall showed the reverse pattern. More study time and restudy choices were allocated to these pairs. However, in the delayed JOL group in Experiment 5, participants gave higher JOLs to incorrectly generated pairs and their recall showed the same pattern. They preferred to restudy Read pairs. Consistent with the "monitoring affects control" hypothesis (MC), the current experiments' results provide new and convincing evidence to support the idea that metamemory control is intimately related to metamemory monitoring.

Conclusion

Metamemory control is related to metamemory monitoring. Generating followed by corrective feedback, even when it produces many errors, leads to better subsequent memory than reading but people tend to be unaware of this benefit when making immediate item-by-item JOLs. Moreover this metamemory illusion affects people's self-regulated learning, including choices about which items to

restudy and how long to study for. People hold a strong belief that generating errors followed by corrective feedback is inferior to reading, which may contribute to the metacognitive misalignment demonstrated in their immediate item-by-item JOLs. Moreover, people's experience of their generation performance and generation fluency also partially contribute to this metamemory illusion. Informing people of the errorful generation benefit before study partially (but not totally) overcame this metacognitive illusion. Delayed JOLs are more accurate than immediate ones and making delayed JOLs is an effective way to overcome the negative consequences of faulty memory monitoring following errorful generation and to help learners make more effective study choices.

References

Besken, M., & Mulligan, N. W. (2013). Easily perceived, easily remembered? Perceptual interference produces a double dissociation between metamemory and memory performance. *Memory & Cognition, 41*, 897-903. doi: 10.3758/s13421-013-0307-8

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In M. J. & S. A. (Eds.), *Metacognition: Knowing about Knowing* (pp. 185-205): Cambridge, MA: MIT Press.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology, 64*, 417-444. doi: 10.1146/annurev-psych-113011-143823

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

Dunlosky, J., & Ariel, R. (2011). Self-regulated learning and the allocation of study time. *Psychology of Learning and Motivation-Advances in Research and Theory, 54*, 103-140. doi: 10.1016/b978-0-12-385527-5.00004-8

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14*, 4-58. doi: 10.1177/1529100612453266

England, B. D., & Serra, M. J. (2012). The contributions of anchoring and past-test performance to the underconfidence-with-practice effect. *Psychonomic Bulletin & Review, 19*, 715-722. doi: 10.3758/s13423-012-0237-7

Finn, B. (2008). Framing effects on metacognitive monitoring and control. *Memory & Cognition, 36*, 813-821. doi: 10.3758/mc.36.4.813

Finn, B., & Metcalfe, J. (2007). The role of memory for past test in the underconfidence with practice effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 238-244. doi: 10.1037/0278-7393.33.1.238

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

Froger, C., Sacher, M., Gaudouen, M. S., Isingrini, M., & Taconnat, L. (2011). Metamemory judgments and study time allocation in young and older adults: Dissociative effects of a generation task. *Canadian Journal of Experimental Psychology, 65*, 269-276. doi: 10.1037/a0022429

Grimaldi, P. J., & Karpicke, J. D. (2012). When and why do retrieval attempts enhance subsequent encoding? *Memory & Cognition, 40*, 505-513. doi: 10.3758/s13421-011-0174-0

Hays, M. J., Kornell, N., & Bjork, R. A. (2013). When and why a failed test potentiates the effectiveness of subsequent study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*, 290-296. doi: 10.1037/a0028468

Hertzog, C., Dunlosky, J., Robinson, A. E., & Kidder, D. P. (2003). Encoding fluency is a cue used for judgments about learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 22-34. doi: 10.1037/0278-7393.29.1.22

Huelser, B. J., & Metcalfe, J. (2012). Making related errors facilitates learning, but learners do not know it. *Memory & Cognition, 40*, 514-527. doi: 10.3758/s13421-011-0167-z

Kang, M. J., Hsu, M., Krajbich, I. M., Loewenstein, G., McClure, S. M., Wang, J. T. Y., & Camerer, C. F. (2009). The wick in the candle of learning epistemic curiosity activates reward circuitry and enhances memory. *Psychological science, 20*, 963-973. doi: 10.1111/j.1467-9280.2009.02402.x

Kang, S. H., Pashler, H., Cepeda, N. J., Rohrer, D., Carpenter, S. K., & Mozer, M. C. (2011). Does incorrect guessing impair fact learning? *Journal of Educational Psychology, 103*, 48-59. doi: 10.1037/a0021977

Knight, J. B., Hunter Ball, B., Brewer, G. A., DeWitt, M. R., & Marsh, R. L. (2012). Testing unsuccessfully: A specification of the underlying mechanisms supporting its influence on retention. *Journal of Memory and Language, 66*, 731-746. doi: 10.1016/j.jml.2011.12.008

Koriat, A., & Bjork, R. A. (2005). Illusions of competence in monitoring one's knowledge during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 187 - 194. doi 0.1037/0278-7393.31.2.187

Koriat, A., & Ma'ayan, H. (2005). The effects of encoding fluency and retrieval fluency on judgments of learning. *Journal of Memory and Language, 52*, 478-492. doi: 10.1016/j.jml.2005.01.001

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review, 14*, 219 -224. doi:10.3758/BF03194055

Kornell, N., & Bjork, R. A. (2008a). Learning concepts and categories is spacing the "enemy of induction"? *Psychological science, 19*, 585-592. doi: 10.1111/j.1467-9280.2008.02127.x

Kornell, N., & Bjork, R. A. (2008b). Optimising self-regulated study: The benefits—and costs—of dropping flashcards. *Memory, 16*, 125-136. doi: 10.1080/09658210701763899

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 989-998. doi: 10.1037/a0015729

Kornell, N., & Metcalfe, J. (2006). Study efficacy and the region of proximal learning framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*, 609-622. doi: 10.1037/0278-7393.32.3.609

Kornell, N., & Rhodes, M. G. (2013). Feedback reduces the metacognitive benefit of tests. *Journal of Experimental Psychology. Applied, 19*, 1-13. doi: 10.1037/a0032147

Mazzoni, G., & Cesare, C. (1993). Strategies in study time allocation: Why is study time sometimes not effective? *Journal of Experimental Psychology: General, 122*, 47-60. doi: 10.1037/0096-3445.122.1.47

Metcalfe, J. (2002). Is study time allocated selectively to a region of proximal learning? *Journal of Experimental Psychology: General, 131*, 349-363. doi: 10.1037//0096-3445.131.3.349

Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review, 15*, 174-179. doi: 10.3758/pbr.15.1.174

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*, 378-384. doi: 10.3758/s13423-012-0343-6

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

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The “delayed-JOL effect”. *Psychological science, 2*, 267-270. doi: 10.1111/j.1467-9280.1991.tb00147.x

Nelson, T. O., & Leonesio, R. J. (1988). Allocation of self-paced study time and the "labor-in-vain effect". *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 676-686. doi: 10.1037/0278-7393.14.4.676

Potts, R., & Shanks, D. R. (2014). The benefit of generating errors during learning. *Journal of Experimental Psychology: General, 143*, 644-667. doi: 10.1037/a0033194

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

Rhodes, M. G., & Castel, A. D. (2009). Metacognitive illusions for auditory information: Effects on monitoring and control. *Psychonomic Bulletin & Review, 16*, 550-554. doi: 10.3758/PBR.16.3.550

Scheck, P., Meeter, M., & Nelson, T. O. (2004). Anchoring effects in the absolute accuracy of immediate versus delayed judgments of learning. *Journal of Memory and Language, 51*, 71-79. doi: 10.1016/j.jml.2004.03.004

Scheck, P., & Nelson, T. O. (2005). Lack of pervasiveness of the underconfidence-with-practice effect: Boundary conditions and an explanation via anchoring. *Journal of Experimental Psychology: General, 134*, 124-128. doi: 10.1037/0096-3445.134.1.124

Silver, N. C., & Dunlap, W. P. (1987). Averaging correlation coefficients: Should Fisher's z transformation be used? *Journal of Applied Psychology, 72*(1), 146-148. doi: 10.1037/0021-9010.72.1.146

Soderstrom, N. C., & Bjork, R. A. (2014). Testing facilitates the regulation of subsequent study time. *Journal of Memory and Language, 73*, 99-115. doi: 10.1016/j.jml.2014.03.003

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

Son, L. K. (2004). Spacing one's study: Evidence for a metacognitive control strategy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 601 - 604. doi: 10.1037/0278-7393.30.3.601

Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 204-221. doi: 10.1037/^)278-7393.26.1.204

Soraci, S. A., Franks, J. J., Bransford, J. D., Chechile, R. A., Belli, R. F., Carr, M., & Carlin, M. (1994). Incongruous item generation effects: A multiple-cue perspective. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 67-78. doi: 10.1037/0278-7393.20.1.67

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*, 1264-1269. doi: 10.1037/a0023719

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*, 918-933. doi: 10.1037/xge0000177

Yan, V. X., Thai, K.-P., & Bjork, R. A. (2014). Habits and beliefs that guide self-regulated learning: Do they vary with mindset? *Journal of Applied Research in Memory and Cognition, 3*, 140-152. doi: 10.1016/j.jarmac.2014.04.003

Yan, V. X., Yu, Y., Garcia, M. A., & Bjork, R. A. (2014). Why does guessing incorrectly enhance, rather than impair, retention? *Memory & Cognition, 42*, 1373-1383. doi: 10.3758/s13421-014-0454-6

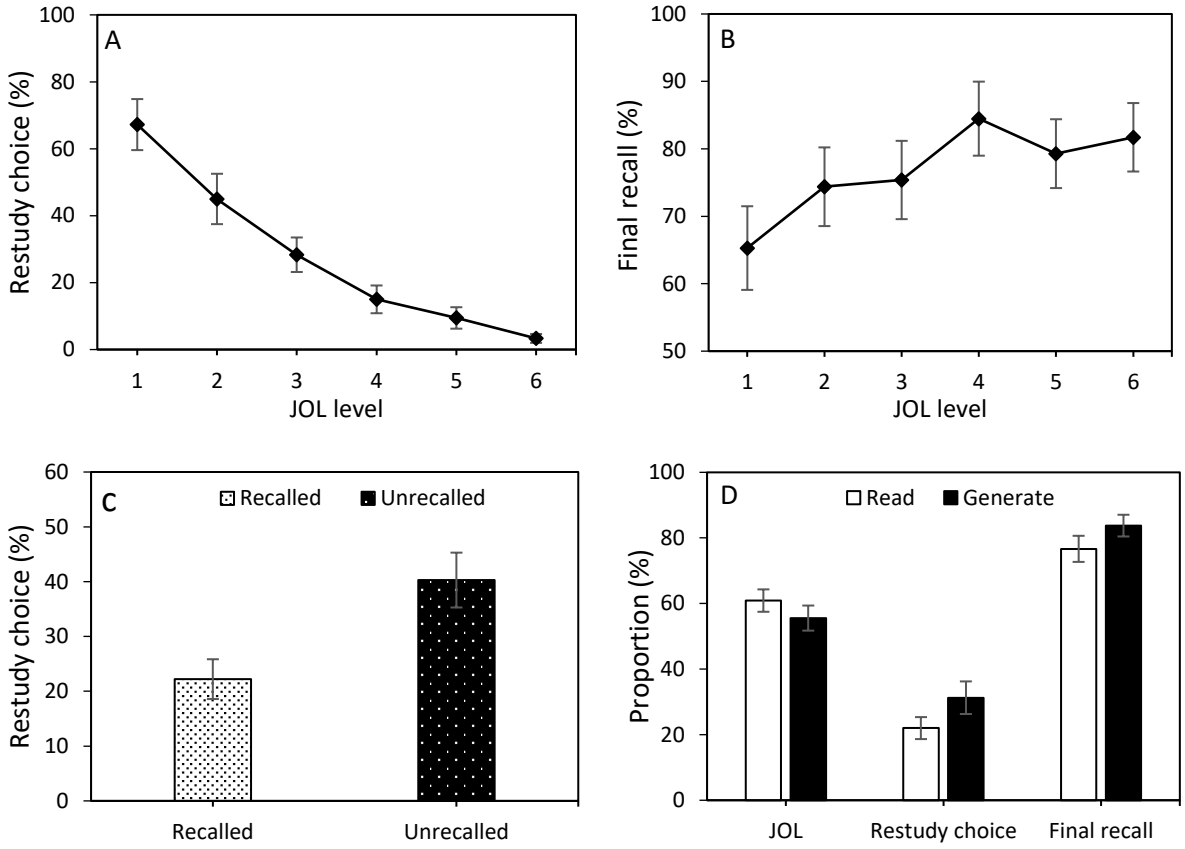


Figure 1. Experiment 1. Panel A: JOL levels and restudy choices. Panel B: JOL levels and final recall. Panel C: Final recall (recalled and unrecalled) and restudy choices. Panel D: JOLs, restudy choices and final recall for the Read and Generate (incorrectly generated) pairs. Error bars represent ± 1 standard error.

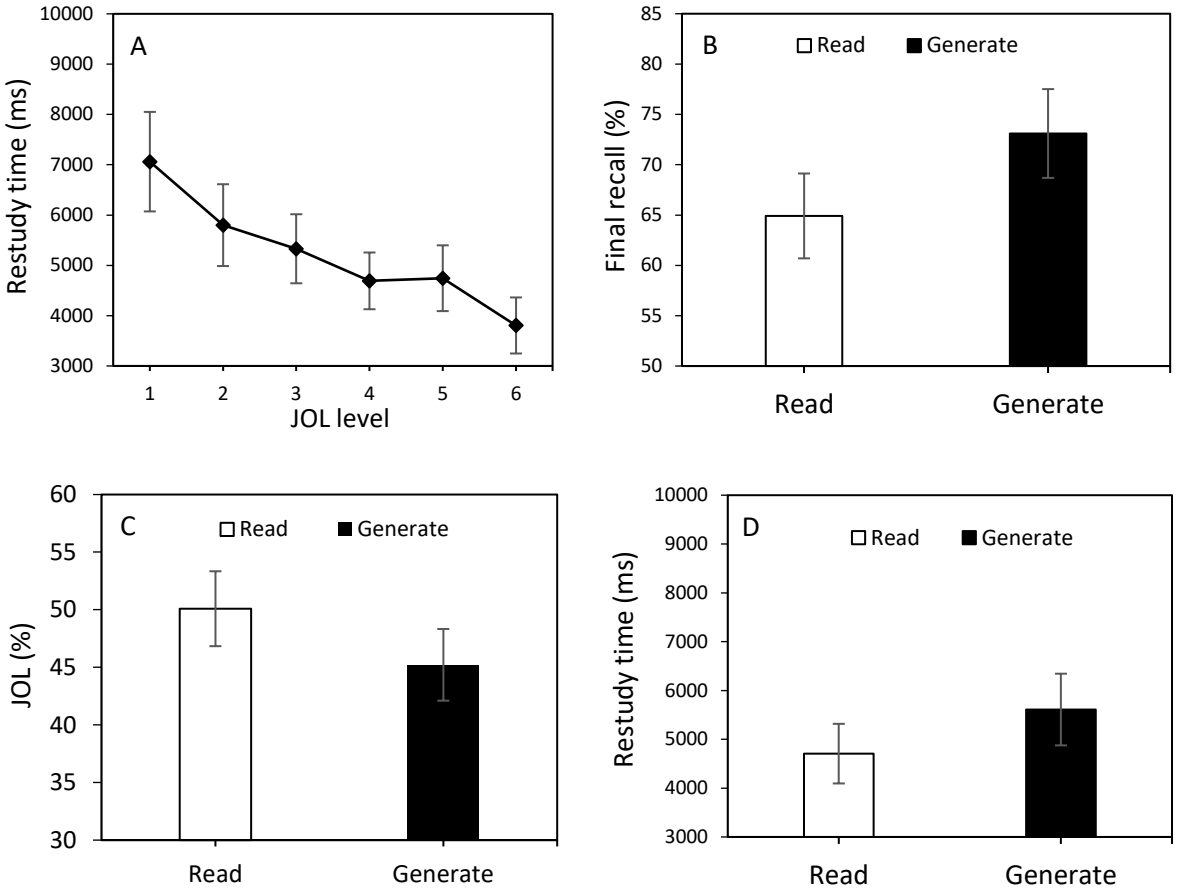


Figure 2. Experiment 2. Panel A: JOL levels and restudy time. Panel B: Final recall for the Read and Generate (incorrectly generated) pairs. Panel C: JOLs for the Read and Generate (incorrectly generated) pairs. Panel D: Restudy time for the Read and Generate (incorrectly generated) pairs. Error bars represent ± 1 standard error.

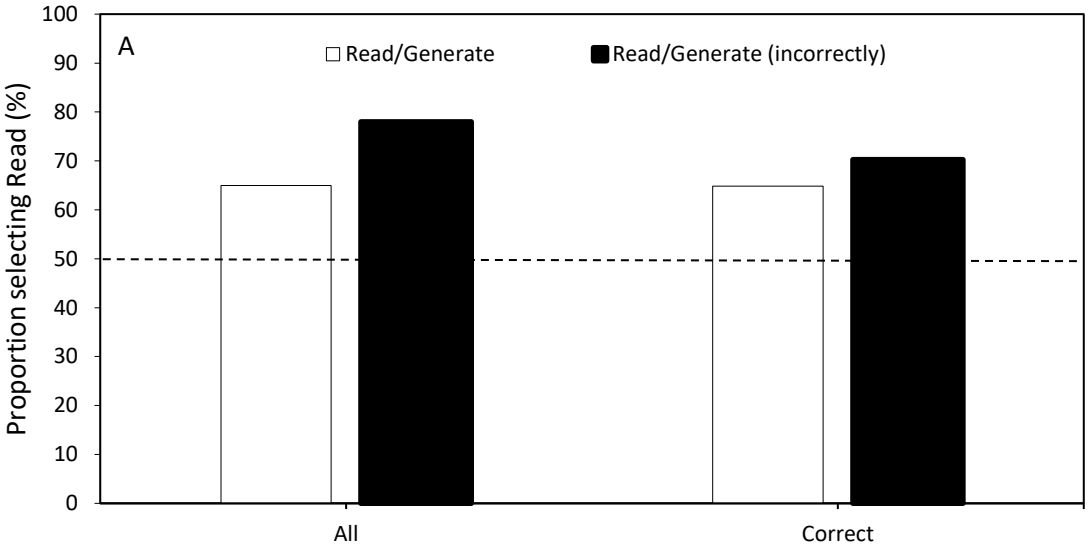


Figure 3A. Experiment 3. Proportion of participants (all participants (All), participants who answered all instruction questions correctly (Correct)) who selected Read as the more effective method when judging Read/Generate and Read/Generate (incorrectly).

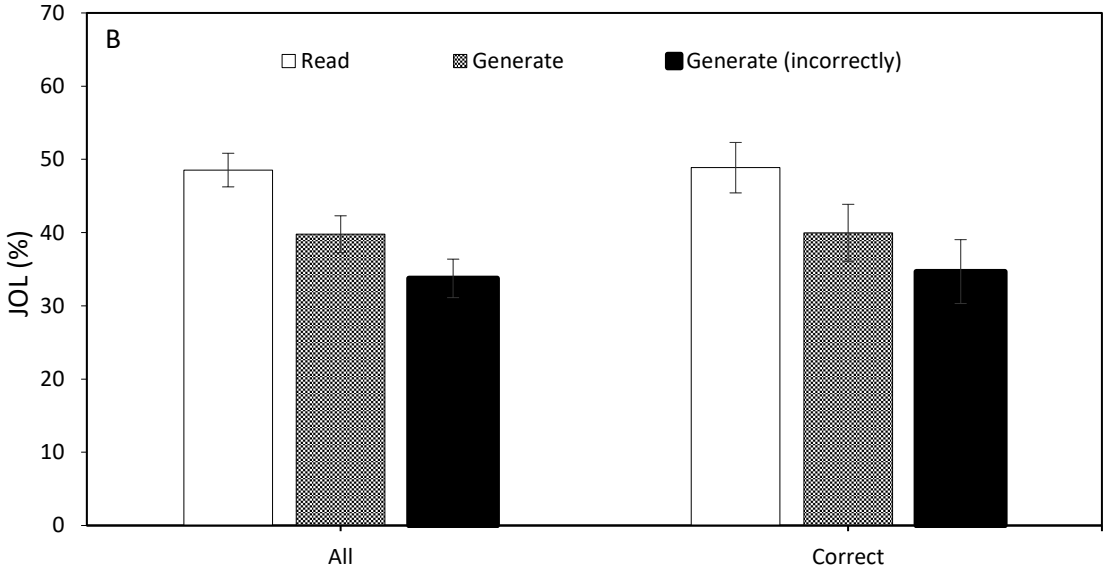


Figure 3B. Experiment 3. Mean proportion of word pairs in different conditions [Read/Generate/Generate (incorrectly)] that participants (all participants (All), participants who answered all instruction questions correctly (Correct)) estimated they would remember. Error bars represent ± 1 standard error.

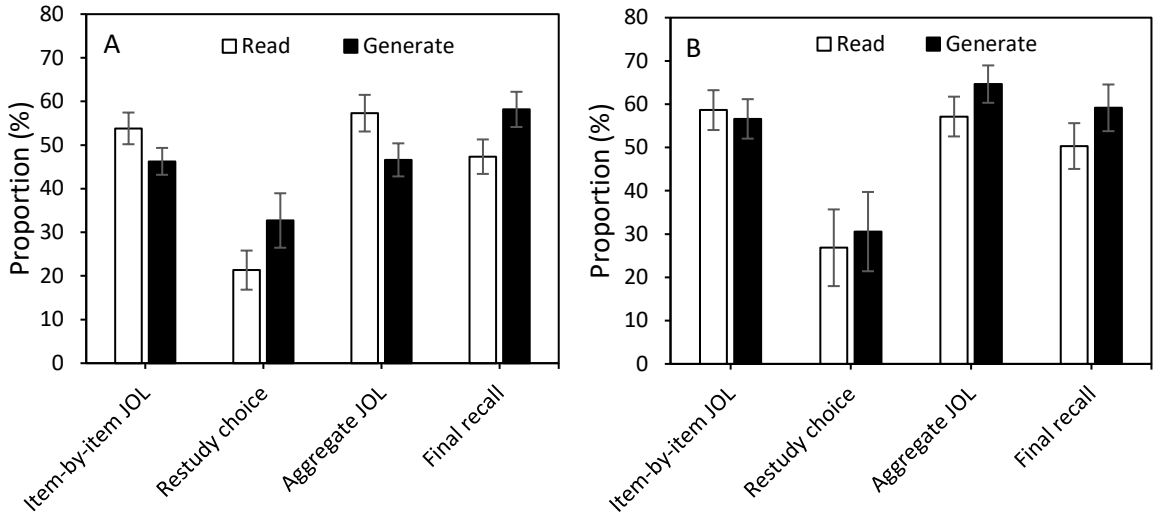


Figure 4. Experiment 4. Item-by-item JOLs, restudy choices, aggregate JOLs, and final recall for the Read and Generate (incorrectly generated) pairs. Data for the Uninformed group are shown in Panel A and data for the Informed group are shown in Panel B. Error bars represent ± 1 standard error.

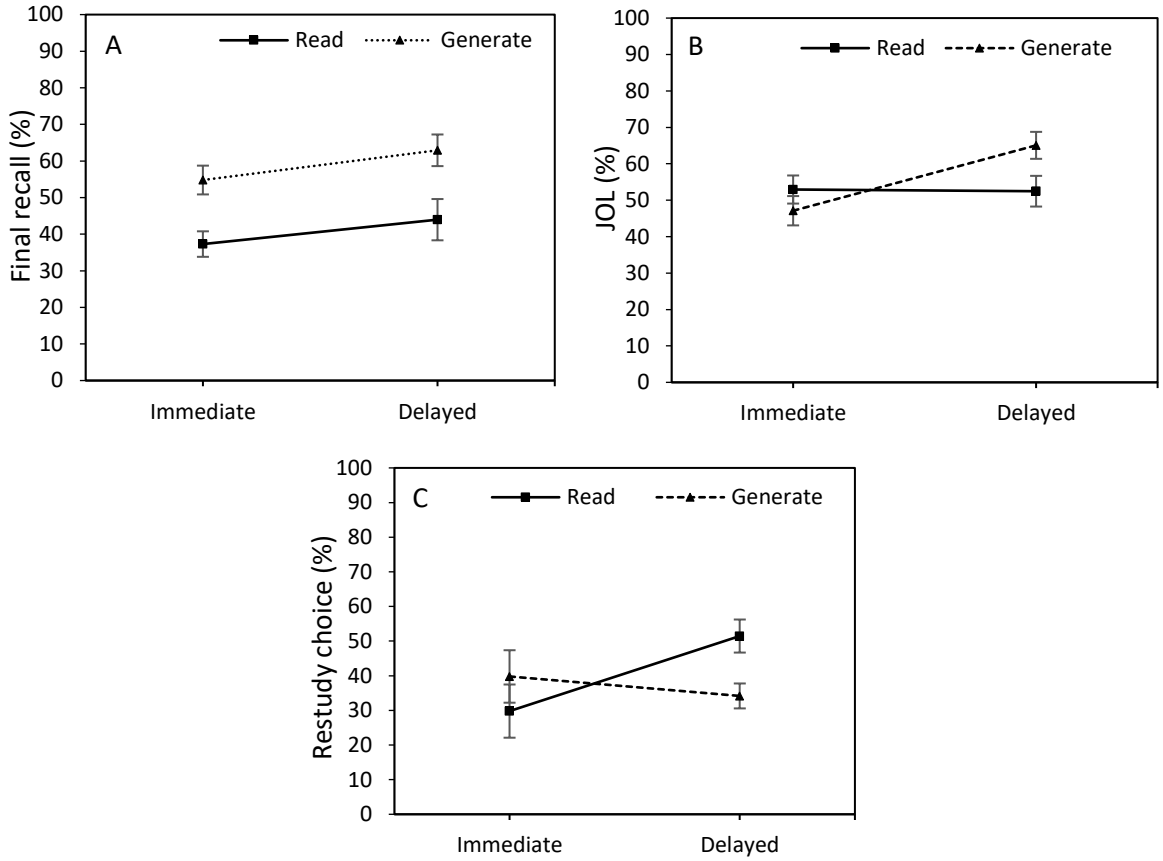


Figure 5. Experiment 5. Panel A: Final recall for the Read and Generate (incorrectly generated) pairs for the immediate and delayed JOL groups. Panel B: JOLs for the Read and Generate (incorrectly generated) pairs for the immediate and delayed JOL groups. Panel C: Restudy choices for the Read and Generate (incorrectly generated) pairs for the immediate and delayed JOL groups. Error bars represent ± 1 standard error.

Appendix A

Table A1. Gamma correlations in Experiment 1

Condition	JOL-Recall	JOL-Restudy choice	Restudy choice-Recall
Read	.32 [.12, .52]	-.77 [-.88, -.67]	-.55 [-.71, -.37]
Generate	.23 [.07, .45]	-.72 [-.84, -.61]	-.39 [-.75, -.04]
All	.21 [.05, .38]	-.77 [-.85, -.69]	-.48 [-.62, -.34]

Note: Experiment 1. *M* [95% CI] gamma values for the correlations for Read, Generate (incorrectly generated), and All (all 60) pairs.

Table A2. Gamma and Pearson correlations in Experiment 2

Condition	JOL-Recall (Gamma)	JOL-Restudy time (Pearson)	Restudy time-Recall (Gamma)
Read	.25 [.06, .43]	-.22 [-.29, -.15]	-.18 [-.29, -.06]
Generate	.15 [.02, .28]	-.25 [-.31, -.18]	-.22 [-.37, -.08]
All	.19 [.09, .28]	-.26 [-.32, -.20]	-.17 [-.27, -.06]

Note: Experiment 2. *M* [95% CI] gamma and Pearson values for the correlations for Read, Generate (incorrectly generated), and All (all 60) pairs.

Table A3. Gamma correlations in Experiment 4

Condition	Item-by-item JOL-Recall	Item-by-item JOL-Restudy choice	Restudy choice-Recall
Read (Informed)	.34 [.17, .51]	-.82 [-.1.04, -.60]	-.43 [-.66, -.19]
Read (Uninformed)	.25 [.10, .40]	-.54 [-.97, -.10]	-.30 [-.77, .17]
Read (Difference)	.09 [-.11, .29]	-.28 [-.67, .11]	-.13 [-.57, .31]
Generate (Informed)	.34 [.09, .60]	-.55 [-.79, -.30]	-.45 [-.69, -.22]
Generate (Uninformed)	.33 [.20, .46]	-.54 [-.86, -.23]	-.43 [-.83, -.03]
Generate (Difference)	-.01 [-.24, .22]	-.01 [-.37, .35]	-.02 [-.43, .10]
All (Informed)	.21 [.13, .28]	-.62 [-.78, -.47]	-.35 [-.51, -.19]
All (Uninformed)	.27 [.16, .39]	-.53 [-.85, -.22]	-.38 [-.75, -.01]
All (Difference)	.06 [-.20, .06]	-.11 [-.42, .20]	-.01 [-.38, .36]

Note: Experiment 2. *M* [95% CI] gamma values for the correlations for Read, Generate (incorrectly generated), and All (all 60) pairs for the Informed and Uninformed groups. Differences between the two groups' gamma values are also reported.

Table A4. Gamma correlations in Experiment 5

Condition	JOL-Recall	JOL-Restudy choice	Restudy choice-Recall
Read (Immediate)	.29 [.07, .51]	-.68 [-.29, -.15]	-.25 [-.62, .11]
Read (Delayed)	.66 [.53, .79]	-.90 [-.1.03, -.77]	-.79 [-.94, -.65]
Read (Difference)	-.37 [-.61, -.12]	.22 [-.01, .45]	.54 [.19, .88]
Generate (Immediate)	.28 [.16, .40]	-.59 [-.85, -.34]	-.28 [-.53, -.04]
Generate (Delayed)	.64 [.51, .77]	-.93 [-.99, -.87]	-.77 [-.90, -.63]

Generate (Difference)	-.36 [-.53, -.20]	.34 [.11, .56]	.48 [.23, .74]
All (Immediate)	.24 [.13, .35]	-.69 [-.87, -.51]	-.24 [-.51, .04]
All (Delayed)	.63 [.49, .77]	-.93 [-1.01, -.85]	-.55 [-.93, -.17]
All (Difference)	-.39 [-.56, -.22]	.24 [.06, .41]	.31 [-.14, .77]

Note: Experiment 5. *M* [95% CI] gamma values for the correlations for Read, Generate (incorrectly generated), and All (all 60) pairs for the immediate and delayed JOL groups. Differences between the two groups' gamma values are also reported.

Appendix B

Perceived differences in relatedness

In Experiments 1, 2, 4, and 5, correctly generated pairs were removed from the data analysis. It seems reasonable to assume that these were more likely to come to mind in response to the cue because they represented a subset that were more closely related for a given individual. This raises the possibility of an item-selection artefact: If the remaining pairs in the Generate condition were more difficult for participants to remember, the difficulty difference between Read and incorrectly generated pairs could lead to the observed metacognitive unawareness. We examined whether there is a difference in difficulty between Read pairs and incorrectly generated pairs by asking participants to rate their semantic relatedness.

Participants

16 native English speakers were recruited from the UCL participant pool (average age = 23.25, *SD* = 4.80, 14 females). Participants received £4 or course credit for participation. All were debriefed after finishing the experiment.

Materials, design and procedure

The same 60 word pairs were used as in Experiment 1, 30 in the Read condition and 30 in the Generate condition. Participants were instructed to rate, on a slider from 0 to 100, each pair's semantic relatedness after studying it. They were informed that 0 indicates totally unrelated and 100 indicates very highly related. After studying all 60 pairs, participants were instructed to solve arithmetic problems for 5 min and then took a final test.

Results

Initial generation performance

4.0% (*SD* = 4.08) of pairs in the Generate condition were correctly guessed and these pairs were removed from the following analyses.

Semantic relatedness ratings, final recall, and correlation

No statistically significant difference between Read and incorrectly generated pairs' semantic relatedness ratings was detected, difference = 0.83, 95% CI [-2.71, 4.37] (Read pairs: *M* = 64.51, *SD* = 13.68; incorrectly generated pairs: *M* = 63.68, *SD* = 13.83). 8 participants gave higher relatedness

ratings to Read pairs and 8 showed the reverse pattern. Final recall of incorrectly generated pairs was significantly better than that of Read pairs, difference = 7.8 %, 95% CI [1.83, 13.27] (for Read pairs, *M* = 76.9%, *SD* = 18.52, and for the incorrectly generated pairs, *M* = 84.4%, *SD* = 13.57). 13 participants recalled a higher proportion of incorrectly generated pairs, while 3 showed the reverse pattern. Gamma correlations between semantic ratings and final recall were calculated for each individual. There was a moderate correlation between semantic relatedness ratings and final recall, *r* = .24, 95% CI = [.16, .32].

Discussion

No difference in semantic relatedness ratings was detected between Read and incorrectly generated pairs. Therefore, the metacognitive unawareness effect cannot be attributed to the degree of semantic relatedness.

Appendix C

Instructions for Experiment 3

Thank you for taking part in this survey.

In this survey, we are investigating people's beliefs about the most effective way to learn new information. Specifically, imagine that you need to learn 60 word pairings, such as *pond-frog*. Your task is to commit this pairing to memory so that when given the word *pond*, you can immediately respond *frog*.

We are interested in two different methods which you might use to learn these pairs. The first method is the Read method. This simply involves studying each word pair comprising a cue word (the first word, e.g., *pond-*) and a target word (the second word, e.g., *frog*). In this case, imagine that you have 13 sec to remember each word pair. In the second method, which is called as the Generate method, a cue word will be presented first (e.g., *pond-?*) and you will have 8 sec to guess the correct target. Then, the correct answer will be presented alongside the cue word for 5 sec for you to study. Assume that the likelihood you can guess correctly is about 5%. Therefore, most of your generations will be incorrect. 24 hours after studying all 60 word pairs (30 pairs by the Read method and the other 30 pairs by the Generate method), there will be a final memory test. All the first words of each pair will be presented one by one and you will have unlimited time to recall the correct answers.

In the following survey, we are interested to know which method (Generate or Read) you think is more effective to learn word pairs. To make sure that you understand the scenario completely, we will give you a short test. If you are not sure you understand our instructions completely, please read them again. When you are ready, please proceed to the test.

Test questions

1. What's the aim of the survey?

   A. To study people's beliefs about the best way to learn new information.

   B. To study the relation between mental illness and aging.

   C. To study the role of environment in personality development.

   D. To study addictive behaviours.

2. In the Read condition, how will the cue and target words be presented?

   A. The cue word and the target word will be presented together for 5 seconds.

   B. The cue word and the target word will be presented together for 13 seconds.

   C. The cue word will be presented first for 8 seconds and you need to take a guess of the target word. Then the cue word and correct answer will be presented together for 5 seconds.

   D. The cue word will be presented first for 5 seconds and you need to take a guess of the target word. Then the cue word and correct answer will be presented together for 8 seconds.

3. In the Generate condition, how will the cue and target words be presented?

   A. The cue word and the target word will be presented together for 5 seconds.

   B. The cue word and the target word will be presented together for 13 seconds.

   C. The cue word will be presented first for 8 seconds and you need to take a guess of the target word. Then the cue word and correct answer will be presented together for 5 seconds.

   D. The cue word will be presented first for 5 seconds and you need to take a guess of the target word. Then the cue word and correct answer will be presented together for 8 seconds.

4. In the Generate condition, what's the likelihood that you can successfully guess the correct answer?

A. 1%

B. 5%

C. 10%

D. 20%

5. When will the final memory test happen after studying the word pairs?

A. Immediately.

B. 5 minutes later.

C. 24 hours later.

D. 1 week later.

Survey questions

1. Which way do you think is more effective to learn word pairs?

A. Generate

B. Read

2. Please make a prediction about the proportion of the word pairs studied in the Read condition you think you will remember in the final memory test. 0 means that you cannot remember any pairs. 100 indicates that you can remember every pair.

0-100

3. Please make a prediction about the proportion of the word pairs studied in the Generate condition you think you will remember in the final memory test. 0 means that you cannot remember any pairs. 100 indicates that you can remember every pair.

0-100

4. A majority (about 95%) of your guesses in the Generate condition will be incorrect. Comparing Generate (incorrectly) with Read, which way do you think is more effective? Generate (incorrectly) is the method used for studying incorrectly guessed pairs.

A. Generate (incorrectly)

B. Read

5. A majority (about 95%) of your guesses in the Generate condition will be incorrect. Please make a prediction about the proportion of the incorrectly guessed word pairs you think you will remember in the final memory test. 0 means that you cannot remember any pairs. 100 indicates that you can remember every pair.

0-100.

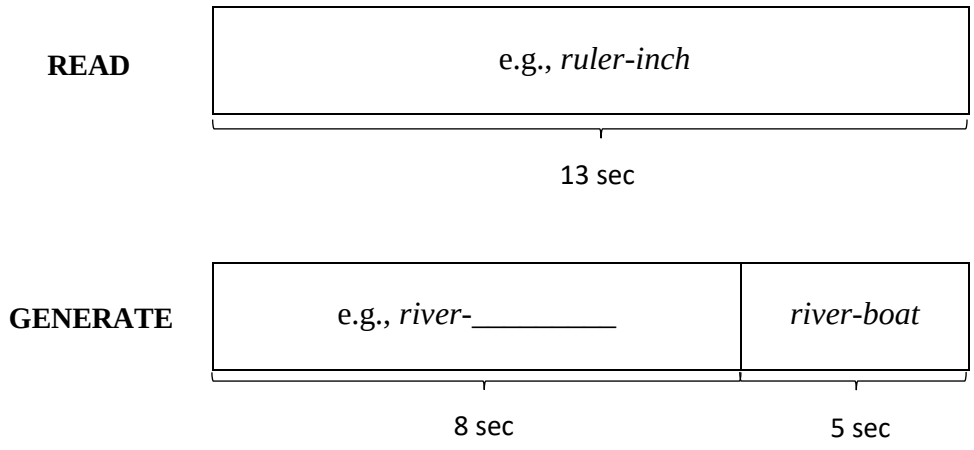
Appendix D

Instructions for Experiment 4

Imagine that you have to learn some English word pairings (e.g., *ruler-inch*), each pairing comprising a 'cue' word (the first word of the pair) and a 'target' word (the second word). Your task is to commit these pairings to memory so that when later given the cue word (*ruler*), you can immediately respond with the target (*inch*).

There are two different methods by which you might learn these pairs. We call the first method the READ method. This simply involves studying the two words side-by-side and trying to commit them to memory. In this case, imagine that the words are presented together for 13 sec.

In the second method, which we call the GENERATE method, a cue word is presented first (e.g., *river-?*) and you have 8 seconds to guess what the target might be. Then, the correct target (*boat*) will be shown alongside the cue word for 5 seconds for you to study. Assume that the likelihood you will guess correctly is about 5%. Therefore, for the majority (about 95%) of pairs in the GENERATE condition your guesses will be incorrect.



Surprisingly, researchers have found that word pairs studied in the GENERATE condition are better recalled than the ones studied in the READ condition in a later test. This is true both when the guess is correct and when it is incorrect.

Many theories have been proposed to explain why this might be the case. One possibility is that, when people see a cue word (e.g., *river-?*) and try to guess the target, many related words will be activated in memory (e.g., *lake, water, sea*). The correct answer (*boat*) will also be activated during the guessing process, and this makes it easier to learn the correct answer when it is shown after the guess. Another possibility is that an incorrect guess can favour later recall of the correct answer. For instance, when shown *river-?*, people may guess *lake*. In a later test, when people see the cue word (*river-?*) they will first recall their guess (*lake*) and then their guess will help them to recall the correct answer (*boat*). A final possibility is that, if people guess incorrectly, the correct answer will surprise them, and capture their attention, which enhances learning of the correct answer.

Test questions

1. According to the above instructions, please choose which pairs can be better remembered.

   A. **Correctly guessed pairs.**

   B. Pairs studied in the READ condition.

2. According to the above instructions, please choose which pairs can be better remembered.

   A. **Incorrectly guessed pairs.**

   B. Pairs studied in the READ condition.

3. What are the possible reasons why incorrectly guessed pairs are better remembered than the ones studied in the READ condition? You can choose more than one option.

A. The correct answer is activated during the guessing process, which makes it easier to learn the correct answer.

B. In a later test, people may first recall their incorrect guess, and then their guess may help them to recall the correct answer.

C. The guessing process is very time-consuming and little time is left for learning the correct answer.

D. The likelihood people guess correctly is very low. Therefore, the incorrectly generated pairs may be very difficult to remember.

E. People may recall their incorrect guess as the correct answer in a later test.

F. If people guess incorrectly, the correct answer may surprise them and capture their attention.