2019-02

Learning from failure: Errorful generation improves memory for items, not associations

Seabrooke, T

10.1016/j.jml.2018.10.001
Journal of Memory and Language
Elsevier

Learning from failure: Errorful generation improves memory for items, not associations

Tina Seabrooke[1], Timothy J. Hollins[1], Christopher Kent[2], Andy J. Wills[1], and Chris J. Mitchell[1]

[1] Plymouth University, UK

[2] University of Bristol, UK

Declarations of interest: none

Please address correspondence to:

Dr Tina Seabrooke

School of Psychology

Plymouth University

Devon

PL4 8AA

United Kingdom

Email: tina.seabrooke@plymouth.ac.uk

Abstract

Potts and Shanks (2014) recently reported that making mistakes improved the encoding of novel information compared with simply studying. This benefit of generating errors is counterintuitive, since it resulted in less study time and more opportunity for proactive interference. Five experiments examined the effect of generating errors versus studying on item recognition, cued recall, associative recognition, two-alternative forced choice and multiple-choice performance. Following Potts and Shanks (2014), participants first attempted to learn the English definitions of either very rare English words or Euskara nouns. During encoding, participants either guessed the definition (and almost always made an error) before the correct definition was revealed, or simply studied the words for an equivalent period. Experiments 1-4 used rare English words. In these experiments, generating errors led to better subsequent recognition of both the cues and targets compared with studying (Experiments 1 and 3). Tests of cued recall and associative recognition, by contrast, revealed no significant benefit of generating errors over studying (Experiments 1-3). Generating errors during encoding also improved performance on a two-alternative forced choice test when the correct target was presented with a novel foil, but not when the familiarity of the target and the foil was matched (Experiment 4). In Experiment 5, a different set of materials – Euskara nouns – and a different (intermixed) encoding procedure was adopted. Here, guessing improved target recognition (performance was improved on a multiple-choice test with unfamiliar foils), but impaired cued recall performance. These results suggest that, when learning word pairs that do not have a pre-existing semantic association, generating errors strengthens the cues and targets in isolation, but does not strengthen the cue-target associations.

Keywords: errors, learning, memory, testing, education

Optimising the learning of educational materials is of critical importance to educators and students alike. Testing is one technique that has been particularly endorsed in recent years (Ariel & Karpicke, 2017; Kornell & Vaughn, 2016; Metcalfe, 2017; Pashler, Rohrer, Cepeda, & Carpenter, 2007). It is now well-established that the process of retrieving information from memory in an initial test can improve retrieval in a later test relative to simply restudying information. This effect is known as the *testing effect* (see Roediger & Karpicke, 2006, for a review).

Students do not always do well on tests though, which might lead educators to worry whether failed tests could do more harm than good. What has become increasingly clear in recent years, however, is that even *unsuccessful* tests improve retention (e.g., Kane & Anderson, 1983; Slamecka & Fevreiski, 1983). Kornell et al. (2009), for example, asked participants to study and remember weakly associated word pairs (e.g., *freckle-mole*). Sometimes, the cue (*freckle*) was first presented alone and the participants were asked to guess the target (*mole*) before it was revealed. Other times, the participants simply studied the intact word pair. In a subsequent cued recall test, the cues were presented and the participants had to generate the targets. Critically, the targets that were presented after an initial unsuccessful guess were better recalled than the targets that were studied without a guess. Kornell et al. (2009) therefore demonstrated that generating errors during learning enhanced memory of the correct answer, even when the original guess was incorrect. We refer to Kornell et al.'s (2009) effect as an example of the benefits of *unsuccessful retrieval* because, although the cues and the targets were weakly related, the participants failed to retrieve the target that the experimenter had (arbitrarily) deemed correct.

Kornell et al.'s (2009) findings are striking for two reasons. First, generating errors improved cued recall even when there was less time to encode the correct target in the test condition than in the study condition. Second, there was more scope for incorrect guesses to have interfered with the participants' memory of the targets (Huelser & Metcalfe, 2012; Yan, Yu, Garcia, & Bjork, 2014). Thus,

the benefits of generating errors appeared to outweigh the costs of less study time and more interference.

The notion that making errors aids learning speaks against the influential view that *errorless* learning is optimal (e.g., Baddeley & Wilson, 1994; Skinner, 1958; Terrace, 1963). This stance is backed by a large literature that suggests that errors are best avoided during learning, at least in memory-impaired individuals (e.g., see Middleton & Schwartz, 2012, for a review). Kornell et al.'s (2009) findings are also intriguing in light of research demonstrating that students often fail to appreciate the value of tests as a learning tool (Karpicke, Butler, & Roediger III, 2009; Kornell & Bjork, 2007; Kornell & Son, 2009; McCabe, 2011). This latter finding suggests that informing students about the value of tests could be an effective way to improve study strategies (Yang, Potts, & Shanks, 2017). Unsurprisingly, then, Kornell et al.'s (2009) results have generated an explosion of interest in recent years (Grimaldi & Karpicke, 2012; Hays, Kornell, & Bjork, 2013; Huelser & Metcalfe, 2012; Knight, Ball, Brewer, DeWitt, & Marsh, 2012; Kornell, 2014; Potts & Shanks, 2014; Richland, Kornell, & Kao, 2009; Vaughn & Rawson, 2012; Yang et al., 2017).

Generating errors might aid learning for several reasons. One possibility, outlined in what we call *search set* theory, is that guessing during learning activates a host of items that are related to the cue, including the correct target (e.g., Grimaldi & Karpicke, 2012; Yang et al., 2017). Even though this activation does not lead to the correct answer being output as a guess, it is suggested to enhance encoding of the correct target when it is revealed. Search set theory therefore naturally predicts that generating errors should only be beneficial for cues and targets that have a pre-existing semantic association, because unrelated targets would not come to mind during the initial search when guessing.

Grimaldi and Karpicke (2012) recently tested this prediction from search set theory. Their participants first attempted to memorise a series of related (e.g., *pond-frog*) and unrelated (e.g., *pillow-leaf*) word pairs. Half of the participants guessed the target for each cue before receiving

corrective feedback. The remaining participants simply studied each word pair. In a subsequent cued recall test, generating errors only improved retrieval of the related word pairs. This finding has led to a prevailing view that generating errors only aids learning when the cue and the target have an existing semantic association (Grimaldi & Karpicke, 2012; Huelser & Metcalfe, 2012; Knight et al., 2012).

Potts and Shanks (2014) recently challenged the view that generating errors only benefits the learning of related word pairs. Their first three experiments involved asking participants to learn the definitions of very rare English words (or the English translations of Euskara – a language isolate – nouns). During encoding, the participants were sometimes shown the cue (the rare word, e.g., *valinch*) and were asked to guess the target (the definition), before the correct target (*tube*) was revealed. Other times, the participants simply studied the intact word pairs (e.g., *valinch = tube*) for the entire trial[1]. In a subsequent multiple-choice test, the cues were presented one at a time, and the participants were asked to select the correct target, which was placed among three novel foils that were created for each word pair. Since the cues were archaic English words, they were very unlikely to have had a pre-existing semantic association with the targets. In line with this assumption, the participants' guesses during encoding were almost always incorrect. Nevertheless, in a series of studies, generating errors enhanced performance on the final test compared with studying. Thus, Potts and Shanks (2014) demonstrated a benefit of generating errors over studying with novel word pairs. Furthermore, the participants seemed to be strikingly unaware of this benefit, since they consistently believed that they were less likely to remember the guessed items than the studied items. Following Potts and Shanks (2014), we refer to their effect as an *errorful generation* effect because, in their procedure, the participants made a genuine error when they failed to

[1]The participants also completed a Choice condition, where they were shown a rare word and were required to choose between two different potential definitions. Incorrect answers in the Choice condition did not lead to better memory than the Study condition on the final test in any of the experiments.

produce the correct target for a cue (rather than failing to retrieve the target that the experimenter had chosen to be correct).

Potts and Shanks' (2014) results spoke against search set theory; generating errors was beneficial for novel cues, for which the search set was unlikely to have included the target. Instead, they proposed that generating errors fosters more curiosity, interest and motivation to learn the answer than simply studying word pairs. They also suggested that the correct target might generate surprise when it contradicts the original guess. This latter idea fits with the influential view that learning is driven by discrepancies between one's expectations and the actual outcome (e.g., Rescorla & Wagner, 1972). These factors were all suggested to improve encoding by directing attention to the target more effectively than studying.

Potts and Shanks' (2014) attentional/motivational account provides a persuasive and intuitive explanation of their results. A question remains, however, as to why generating errors improved the learning of novel word pairs in their task, but did not boost the learning of unrelated word pairs in previous studies (Grimaldi & Karpicke, 2012; Huelser & Metcalfe, 2012; Knight et al., 2012). One notable difference between the experiments concerns the final test procedure. Grimaldi and Karpicke (2012), Huelser and Metcalfe (2012) and Knight et al. (2012) all used cued recall tests, which require associative knowledge (because the cue should only bring the target to mind once a cue-target association has formed). Potts and Shanks' (2014) multiple-choice tests, by contrast, did not require associative knowledge, because the targets were placed among novel foils[2]. This meant that final test performance could have been driven by target strength alone, because the target was the only option that was presented during encoding. Hence, generating errors during encoding might have strengthened the targets in isolation (and potentially the cues as well), without

[2]Potts and Shanks' (2014) final experiment employed a complex testing procedure that involved presenting the correct target among foils that served as targets in the other encoding conditions. The foils were not, therefore, novel. We shall return to this experiment in the General Discussion.

necessarily strengthening the cue-target *associations*. This would explain why generating errors was unhelpful for unrelated word pairs in cued recall, yet helpful for novel word pairs in Potts and Shanks' (2014) multiple-choice tests.

Potts and Shanks' (2014) results have clear potential to inform practices in educational settings. However, it is important to clarify the mechanisms that underlie the effect, in order to identify the conditions under which it generalises. Their effect looks like a demonstration that generating errors potentiates the learning of associations, since their task involved attempting to learn cue-target associations. However, their multiple-choice testing procedure leaves open the possibility that generating errors boosted the strength of the individual cues and targets, without simultaneously strengthening the cue-target *associations*. The present experiments aimed to discriminate between these two possibilities, in order to better understand the extent to which generating errors is a useful strategy when learning novel word pairs.

Experiment 1

Experiment 1 explored whether Potts and Shanks' (2014) observed benefit of generating errors reflects an increase in item strength, associative strength, or both. We used a cued recall test to examine whether generating errors strengthens cue-target *associations*, since cued recall requires associative knowledge. We also used an item recognition test, which assesses item strength but does not require associative knowledge. Thus, the two tests (cued recall and item recognition) allowed us to test whether errorful generation improves memory of the cues and targets, the cue-target associations, or both.

Most studies to date have examined how generating errors affects subsequent memory of the targets. Kornell et al. (2009), for example, asked participants to recall the target for each cue. Relatively few published studies, by contrast, have examined memory for the *cue*. Hays et al. (2013) provided one exception; they found that unsuccessful retrieval produced better recall of the cue when prompted by the target than studying in a weak-associates (e.g., *frog-pond*) task that required

associative knowledge. To our knowledge, no published studies have examined the effect of errorful generation on memory for novel vocabulary cues. From an applied perspective, it is important to know whether generating errors improves memory of the novel words as well as the definitions. Perhaps after generating errors, participants attend to the target at the expense of the cue, leading to *worse* recognition of the cue than studying. Experiment 1 tested this possibility by examining both cue and target recognition.

Experiment 1 also tested whether attempting to generate the correct answer is necessary to improve memory relative to studying, or whether generating *any* answer is sufficient. Perhaps generating words is simply more engaging than studying, leading participants to attend to and encode the word pairs more effectively in the generate condition. To test this possibility, we asked participants to type the first word that came to mind for some of the cues, before the target was revealed. If generating errors improves learning because typing anything is more interesting than studying, then typing the first word that comes to mind should produce a similar benefit to generating errors. If, however, guessing the target improves performance on the final test relative to stating the first word that comes to mind, it would suggest that guessing the target (rather than generating anything) is crucial.

The encoding phase of Experiment 1 was based on Potts and Shanks' (2014) first experiment, in which participants studied rare English words and their definitions. In a Study condition, the cues and targets were presented together (e.g., *picaroon = cheat*) for 17 seconds, and the participants simply studied and tried to remember them. In a Meaning condition, the cues were presented on their own for 10 seconds (e.g., *spoffish*), and the participants were encouraged to guess the definition. The cue and the target were then presented together (*spoffish = fussy*) for a further seven seconds. The First Word condition was the same as the Meaning condition, except that the participants were simply asked to state the first word that the cue made them think of. Thus, while the total duration of the trials in each encoding condition was matched, the target was

presented for longer in the Study condition than in the Meaning and First Word conditions. We chose to match the overall trial duration of the encoding conditions rather than the target presentation alone to be consistent with both Potts and Shanks' experiments and many others on the topic of unsuccessful retrieval (e.g., Knight et al., 2012; Kornell et al., 2009; Yan et al., 2014; Yang et al., 2017). Indeed, Potts and Shanks' results are especially interesting precisely *because* errorful generation was beneficial even though it was at the expense of study time. We therefore felt that it was important to uphold this critical feature of their design.

After the encoding phase, half of the participants completed a cued recall test, where the cues were presented one by one and the participants had to generate the target. The remaining participants completed an item recognition test, where the cues and targets were presented (plus novel foils) and the participants had to decide whether each item was presented previously. The question was whether Potts and Shanks' (2014) results would extend to item recognition (for both cues and targets) and cued recall tests.

Method

Participants. We chose a sample size of 48 participants per group before data collection. This sample size has good power to detect within-subject effects of the size reported in Potts and Shanks (2014) for the English-words version of the task used here (> 90% power at Cohen's d_z = 0.49). Thus, 96 Plymouth University psychology undergraduates (19 males and 77 females, aged between 18 and 50, *M* = 21.19 years, *SEM* = 0.56 years) took part for course credit. The participants were randomly allocated to the item recognition (*N* = 48) or cued recall (*N* = 48) group before the experiment. We did not record whether participants were native English speakers in Experiments 1-4. Experiments 1-4 were approved by the Plymouth University Ethics Committee.

Apparatus and materials. The experiment was programmed in E-Prime 2.0 and was presented on a 22-inch computer monitor. Stimuli were presented on a white background, and responses were made using a standard keyboard. The word pairs were taken from Potts (2013). Sixty

word pairs, which were randomly allocated to the Study, Meaning, First Word or foil trials for each participant, were used for the main task. Nine additional word pairs were randomly allocated to the practice trials for each encoding condition.

Procedure. At the start of the experiment, the participants were told that they would study rare English words and their definitions in three formats. The order of the three encoding conditions (Study, Meaning and First Word) was counterbalanced between-subjects. Before each encoding condition, the experimenter read aloud on-screen instructions stating that the participants would see rare English words and that they should try to remember the correct definitions, because they would be tested afterwards. The critical instruction for each stage (which told the participants to study the words, guess the meaning of the rare word or type the first word that came to mind) was presented in red. All other text was presented in black.

Each encoding condition (Study, Meaning and First Word) consisted of one presentation of each of 15 different word pairs. In the Study condition, each word pair was presented centrally for 17 seconds (e.g., *gadoid = fish*). The participants simply studied the words. In the Meaning condition, the cue was first presented alone for 10 seconds, above the question, “What do you think this word means?” The participants were firmly encouraged to type a one-word definition, which appeared beneath the question. The participants could use the Backspace key to change their answer during this time. The target appeared after the 10 seconds, together with the cue, for a further seven seconds. The First Word condition was the same as the Meaning condition, except that the question read “What does this word make you think of?”, and the participants simply typed the first word that came to mind. It was impossible to force the participants to respond whilst controlling the trial duration, but we did emphasise the importance of responding on every Meaning/First Word trial. We also stressed that there was no right or wrong answer before the First Word condition. Each condition was preceded by three practice trials.

The test phase immediately followed the encoding phase. During the cued recall test, the 45 cues from the encoding phase were presented in a random order. Fifteen novel foil cues were also randomly intermixed (so that the total number of word pairs was equated for both groups), creating 60 trials in total. The question, "What does this word mean?" was presented beneath each cue, and the participants were firmly encouraged (but not forced) to type the correct target (in the same way as in the Meaning encoding condition). Responses were not time-limited. Three practice trials were administered before the main test, using cues from the encoding practice trials.

During the item recognition test, the 45 cues and 45 targets from the encoding phase were presented one at a time, in a random order. Fifteen foil cues and 15 foil targets were intermixed, creating 120 trials in total. On each trial, the question "Did you see this word before?" was presented below the word. The participants chose between "YES" and "NO" options with the mouse. Six practice trials, using three cues and three targets from the encoding practice trials, preceded the main test.

The word pairs were presented in size 16 Verdana font and in lowercase in Experiments 1-4. The trials were separated by 3-4 second intervals.

Results

The trial-level raw data for this experiment, including the complete list of word pairs, are publicly available at https://osf.io/bwyr8/.

One participant correctly guessed the target for one cue during encoding. Another participant reported that the target was the first word that came to mind for another cue. These trials were removed from all subsequent analyses. Although excluding the correct generations did not alter the results in any of the reported experiments, we report the complete results (correctly-generated items included) in Supplementary Materials A for completeness. The foil cues were rarely correctly defined in the recall test (M = 0.83%, SEM = 0.47%), but the participants were very good at

recognising that they were novel in the item recognition test (cue accuracy: *M* = 85.56%, *SEM* = 2.03%; target accuracy: *M* = 89.72%, *SEM* = 1.93%).

Figure 1a shows the item recognition data. A repeated-measures ANOVA on the encoding condition (First Word, Meaning, Study) and item type (cue, target) factors revealed a main effect of encoding condition, F (2, 94) = 23.64, mean square error (*MSE*) = 214.24, $p < .001$, generalised eta squared (η_g^2) = .07, but not of item type, $F < 1$. There was a significant encoding condition × item type interaction, F (2, 94) = 3.90, *MSE* = 176.07, $p = .02$, $\eta_g^2 = .01$. For the cues, incorrectly guessing the meaning of the cue led to better recognition of that cue than studying, t (47) = 6.34, $p < .001$, $d_z = 0.92$. Similarly, typing the first word that came to mind facilitated better cue recognition than studying alone, t (47) = 5.25, $p < .001$, $d_z = 0.76$. No significant difference in cue recognition was observed between the Meaning and First Word conditions, t (47) = 1.28, $p = .21$, $d_z = 0.18$. For the targets, generating errors produced better target recognition than studying, t (47) = 3.53, $p < .001$, $d_z = 0.51$, or stating the first word that came to mind, t (47) = 2.62, $p = .01$, $d_z = 0.38$. No significant difference was observed in recognition of targets in the First Word and Study conditions, t (47) = 1.20, $p = .24$, $d_z = 0.17$[3].

The above analyses suggest that, relative to the Study condition, the First Word condition differentially affected cue and target recognition. To confirm this conclusion, we examined the effect of item type on just the First Word and Study conditions. An item type (cue, target) × encoding condition (First Word, Study) repeated-measures ANOVA revealed a significant effect of encoding condition, F (1, 47) = 18.20, *MSE* = 218.32, $p < .001$, $\eta_g^2 = .04$, but not of item type, $F < 1$. Most importantly, there was a significant encoding condition × item type interaction, F (1, 47) = 6.90, *MSE*

[3] We also calculated and conducted a repeated analysis on *d'* and response bias (*c*) scores for each encoding condition in the item recognition test. These analyses are reported in Supplementary Materials B.

= 198.93, p = .01, η_g^2 = .01. This interaction demonstrates that, relative to the Study condition, the First Word condition had a greater effect on cue recognition than on target recognition.

A comparable analysis was conducted on the Meaning and Study conditions. There was a main effect of encoding condition (Meaning, Study), F (1, 47) = 37.21, MSE = 265.86, p < .001, η_g^2 = .10, but not of item type (cue, target), F (1, 47) = 2.58, MSE = 200.79, p = .12, η_g^2 = .006. The encoding condition × item type interaction was not significant, F (1, 47) = 2.05, MSE = 169.03, p = .16, η_g^2 = .004. Thus, there was no evidence of a differential effect of encoding condition on cue and target recognition.

Figure 1b shows the cued recall data. Responses were submitted, on average, on 99.68% of trials. Thus, incorrect responses were usually errors of commission rather than omission. Mean accuracy across encoding conditions was 12.92% (*SEM* = 1.57%). The graph shows the results with a conservative scoring approach, where participants were required to recall the target exactly as it was presented during encoding. Our conservative scoring approach revealed no significant effect of encoding condition, F (2, 94) = 1.82, MSE = 115.58, p = .18, η_g^2 = .02[4]. We subsequently adopted a liberal scoring approach, where responses were deemed correct so long as the first five letters matched those of the target (the fifth letter was the point at which all of the targets differed). However, this approach did not reveal a significant effect of encoding condition either, F (2, 94) = 1.96, MSE = 117.10, p = .16, η_g^2 = .01.

Non-significant results might reflect the absence of a difference between conditions. Alternatively, the data might simply be insufficiently sensitive to differentiate the experimental and null hypotheses. Bayes factors are useful for distinguishing these possibilities. Bayes factors of more than three (BF_{10} > 3) support the experimental hypothesis. Values smaller than one third (BF_{10} < 1/3)

[4] Here and in all subsequent cases, a Greenhouse-Geisser correction was applied to adjust for violations of sphericity.

support the null hypothesis. Bayes factors between one third and three suggest that the data are insensitive to distinguish the theories (Jeffreys, 1961).

A Bayes factor was calculated to determine whether the cued recall data support the null hypothesis. Potts and Shanks (2014) reported that generating errors improved final test performance, relative to studying, by approximately 5%. We used this figure (5%) as the mean, and half that figure (2.5%) as the standard deviation, of our Gaussian prior distribution (see also Dienes, 2014; Edmunds, Milton, & Wills, 2015; Edmunds, Wills, & Milton, 2018). The observed mean difference score in the cued recall group of our experiment was -1.67% (*SEM* = 2.39%), with the Study condition producing better numerical performance than the Meaning condition. This calculation produced a Bayes factor of 0.14. A comparable analysis with the liberal scoring approach (*mean difference* = -2.92%, *SEM* = 2.40%) produced a Bayes factor of 0.11. Both scoring approaches continue to support the null with more conservative priors (the Bayes factor remains less than 1/3 for a prior of a mean difference of 2.6% for strict coding, and of 2% for lenient coding). Thus, the data support the notion that the Generate condition did not produce better cued recall performance than the Study condition.

Discussion

In Experiment 1, incorrectly guessing the meaning of archaic English words improved recognition of the definitions compared with studying. This benefit of generating errors was observed even though the incorrect guesses were at the expense of time that was otherwise spent studying. Thus, Experiment 1 demonstrated that Potts and Shanks' (2014) effect generalises to target recognition. The cued recall test, by contrast, revealed no significant effect of encoding condition, and errorful generation produced numerically worse performance than studying. We recognise here that cued recall performance was close to floor in all three conditions, which might have masked any benefit of generating errors over studying. Experiment 2 therefore tested whether Potts and Shanks' (2014) effect would generalise to a more sensitive cued recall test.

Interestingly, incorrectly guessing the target during encoding improved both cue and target recognition compared with studying. We therefore demonstrated that Potts and Shanks' (2014) effect generalises to cue recognition as well as target recognition. Clark (2016) observed a similar result with unrelated common English word pairs (e.g. *pond-spanner*); she found that generating errors improved subsequent free recall of the cues (which does not require associative knowledge) compared with studying (unpublished doctoral thesis). These findings are notable, because they demonstrate that generating errors does not improve target recognition at the *expense* of cue recognition. This is important from an applied perspective, since students often need to have good memory of both the cues and the targets (Carpenter, Pashler, & Vul, 2006).

Finally, stating the first word that came to mind improved cue recognition compared with simply studying the word pairs. The effect did not, however, extend to target recognition. This is perhaps to be expected, given that the participants were not instructed to relate the generated response to the target in the First Word condition. The Meaning condition, on the other hand, asked participants to guess a word that was about to appear as feedback. This perhaps led to deeper processing of the target.

Experiment 2

The aim of Experiment 2 was to make the cued recall task from Experiment 1 easier, to improve overall performance. This would allow us to determine whether the failure to observe a benefit of errorful generation in the cued recall test of Experiment 1 was because performance was subject to a floor effect, or because generating errors does not improve the learning of cue-target associations (relative to studying).

As in Experiment 1, the participants first attempted to learn the definitions of rare English words by studying them, guessing the definitions before they were revealed, or stating the first word that the cues brought to mind. These encoding conditions were blocked and the order was counterbalanced between-subjects. In contrast to Experiment 1, a cued recall test was administered

immediately after each encoding condition. This meant that the participants were only tested on 15 word pairs at any time, as opposed to the 45 word pairs that were tested at once in Experiment 1. We expected this modification to boost overall performance, thereby providing a better test of the effect of generating errors on cued recall. Our focus in Experiment 2 was on cued recall performance, so an item recognition test was not administered. No foils were presented during the cued recall tests, because there was no item recognition test to match the total number of word pairs with.

Method

The method was the same as that used for the cued recall group in Experiment 1, except in the following respects.

Participants. Forty-eight Plymouth University students (11 males and 37 females, aged between 18 and 43, *M* = 20.98 years, *SEM* = 0.68 years) took part for course credit. The sample size was chosen before data collection because it provided good power to detect a medium-sized effect (>90% power at d_z = 0.49).

Procedure. Fifteen word pairs from a pool of 45 word pairs were randomly allocated to the Study, Meaning and First Word encoding conditions for each participant. A cued recall test was administered for the 15 word pairs immediately after each encoding condition. Three practice trials were completed before each encoding condition. After the first three practice trials of the experiment (i.e., immediately before the first encoding condition), three practice cued recall trials were administered.

Results

The trial-level raw data for this experiment, including the complete list of word pairs, are publicly available at https://osf.io/q7jwf/.

One participant generated one target in the First Word condition during encoding; this word pair was excluded from further analysis for that participant. Figure 2 shows the cued recall data with a conservative scoring approach (see Experiment 1 for further details). On average, responses were submitted on 95.83% of trials. Thus, incorrect responses were usually errors of commission rather than omission. Overall mean accuracy on the recall test was 33.11% (*SEM* = 2.27%). As in Experiment 1, we adopted first a conservative and then a liberal scoring approach. The graph suggests that the Study condition promoted numerically better subsequent cued recall than the Meaning and First Word conditions. However, the main effect of encoding condition was not significant, F (2, 94) = 2.71, MSE = 181.77, p = .08, η_g^2 = .02. This pattern did not change with the liberal scoring approach, F (2, 94) = 3.14, MSE = 181.40, p = .06, η_g^2 = .02[5].

A Bayes factor was calculated to determine whether the cued recall data provide evidence for the null. We suspected that the null cued recall result in Experiment 1 was due to a floor effect. The cued recall results in Experiment 1 are not, therefore, an appropriate prior. Instead, we used the target recognition difference observed in Experiment 1 as our prior. Here, target recognition was 11.67% more accurate in the Meaning condition than in the Study condition. Our Gaussian prior therefore had a mean of 11.67% and a standard deviation of half that (5.84). With the conservative scoring approach, the mean difference score in the current experiment was -5.69% (*SEM* = 3.13%), with the Study condition producing numerically better cued recall performance than the Meaning condition. This resulted in a Bayes factor of 0.08. A comparable analysis with the liberal scoring approach (*mean difference* = -6.67%, *SEM* = 3.12%) produced a Bayes factor of 0.10. Both values are below one third, and therefore support the null hypothesis. Both scoring approaches still supported

[5] It might be argued that the use of multiple cued recall tests could have improved the encoding of subsequently presented material (e.g., Szpunar, McDermott, & Roediger, 2008). Of course, this possibility cannot explain differences between encoding conditions, because the order of conditions was counterbalanced. An encoding condition × counterbalancing condition mixed ANOVA revealed no significant main effects (Fs < 2.67, ps > .09) and no significant interaction, F < 1.

the null with more conservative priors (the Bayes factor remained below 1/3 for a prior of a mean difference of 1.98% for strict coding, and of 1.75% for lenient coding). Thus, the data suggest that the Meaning condition did not produce better cued recall performance than the Study condition.

Discussion

Experiment 2 successfully improved cued recall performance, allowing a more thorough examination of the effect of generating errors on cued recall (average recall accuracy was 33.11% in Experiment 2 versus 12.92% in Experiment 1). Numerically, the Study condition improved cued recall performance compared to the Meaning and First Word conditions. This non-significant trend is consistent with the cued recall result in Experiment 1. Together, the results suggest that errorful generation improves item recognition but not cued recall performance.

Experiment 3

Experiment 3 first aimed to replicate the benefit of errorful generation over studying that was observed for the targets in the item recognition test of Experiment 1. We also used an associative recognition test to more cleanly examine whether errorful generation potentiates the learning of *associations* (see e.g., Clark, 1992; Glenberg & Bradley, 1979 for examples of associative recognition tests). Associative recognition tasks involve initially presenting a series of cues (C) and targets (T) together (e.g., C1-T1, C2-T2, C3-T3, C4-T4). In a subsequent test, half of the pairs remain intact (C1-T1, C2-T2), and the rest are re-paired (C3-T4, C4-T3). The participants have to determine whether the word pairs were presented *together* at encoding. Both cued recall and associative recognition tests assess knowledge of the cue-target associations. Associative recognition is a purer test of associative strength than cued recall, though, because it is not dependent on participants' ability to actively generate the target. Experiment 3 therefore tested whether Potts and Shanks' (2014) effect would extend to associative as well as item recognition.

As in the previous experiments, the participants first attempted to learn the definitions of obscure English words by guessing the definition and receiving corrective feedback, or by studying

the words for an equivalent duration. The associative recognition task required more word pairs than the previous experiments, because we needed a reasonable number of paired and re-paired items in each encoding condition. We therefore omitted the First Word condition, to reduce the potential for chance performance (because of an overload of items) in the Meaning and Study conditions. After completing both encoding conditions, the participants completed either a target or an associative recognition test.

For the target recognition group, the participants were presented with the targets from the encoding phase (plus foils), and were asked to state whether each word was presented previously. For the associative recognition group, half of the word pairs (randomly determined for each participant) in each encoding condition were allocated to a paired list; the rest were allocated to a re-paired list. In the subsequent associative recognition test, the items from the paired list were presented intact (e.g., the cue *roke* was presented alongside the correct target *mist*). For the re-paired items, the target was presented with *another* cue from that same encoding condition (e.g., the cue *contumacious* might have been presented alongside the target *disgrace*, which is actually the target for the cue *opprobrium*). The participants had to determine whether the word pairs were presented *together* at encoding.

We expected the Meaning condition to boost target recognition relative to the Study condition (as in Experiment 1). The question was with respect to the associative recognition test. If the benefit of generating errors over studying is seen in the associative recognition test, it would suggest that errorful generation improves associative knowledge as well as item memory strength. If generating errors does not improve associative recognition performance, however, it would suggest that errorful generation does not strengthen cue-target associations. This would explain why generating errors did not improve cued recall in Experiments 1 and 2, since cued recall also requires associative knowledge.

Method

Participants. Fifty-six participants (16 males and 40 females, aged between 19 and 73, *M* = 29.21 years, *SEM* = 1.93 years) were recruited from Plymouth University for £4 each. The participants were randomly allocated to the target (*N* = 28) or associative (*N* = 28) recognition group beforehand. The sample size was chosen before data collection because it provided adequate power to detect an advantage of Meaning over Study at the effect size observed in Experiment 1 (> 80% power at d_z = 0.51).

Apparatus and materials. Eighty-four word pairs were used for the main task, plus eight additional practice pairs. They were taken from Potts (2013) and were randomly assigned to the conditions for each participant. All other materials were identical to those of Experiment 1.

Procedure. The participants completed an encoding phase, then a test phase. The encoding conditions (Study, Meaning) were blocked and counterbalanced for order. The target/associative recognition test occurred immediately after completion of *both* encoding conditions. There were 32 trials in each encoding condition, with four additional practice trials before each. The encoding phase was otherwise the same as in Experiments 1 and 2.

The target recognition test followed the format of the item recognition test of Experiment 1, except that only target recognition was assessed (rather than cue and target recognition). The 64 targets from the encoding phase were presented, plus 20 foils. Eight practice trials preceded the main test, using the encoding practice targets. Other aspects were as in Experiment 1.

In the associative recognition test, word pairs (e.g., *roke = mist*) were presented above the statement, “Were these words presented together?” The participants chose between “YES” and “NO” options using the mouse (responding was not time-limited). Half of the word pairs were from each encoding condition. Furthermore, half of the word pairs in each encoding condition (randomly chosen for each participant) were paired (i.e., they were presented together during encoding), and

the rest were re-paired (they were not presented together at encoding). Thus, there were four trial types: Study paired, Study re-paired, Meaning paired, and Meaning re-paired. For the re-paired trials, a random subset of the word pairs from each encoding condition (e.g., Meaning) were allocated to the re-paired condition during encoding for each participant. These cues (C) and targets (T) were labelled C1-C16 and T1-T16, respectively, with the numbers denoting the order of presentation during encoding. Reassignment of the re-paired cues and targets for presentation on test was achieved by swapping adjacently presented targets. Hence, within each re-paired encoding condition on test (e.g., Meaning re-paired), cue C1 was presented with target T2, and C2 was presented with T1 (likewise, C3 and T4, and C4 and T3 were presented together, and so forth). Note that the re-paired cues and targets were not necessarily from word pairs that were *immediately* adjacent during encoding. Rather, the re-paired cues and targets were from the adjacent word pairs that were (randomly) allocated the re-paired condition from the same encoding condition (e.g., Meaning). On test, there were 16 trials of each of the four trial types; these were randomly ordered. Each cue and target appeared only once each on test.

Eight practice test trials (using words from the encoding practice phases) preceded the main task. The practice phase included two trials from each of the four trial types. Feedback was provided on practice trials to emphasise that the task was to determine whether the cue and the target were presented *together*, not simply whether they were presented at all during encoding. No feedback was provided during the main test.

Results

The trial-level raw data for this experiment, including the complete list of word pairs, are publicly available at https://osf.io/mf7s5/.

Nine participants correctly guessed the definition of one cue each during encoding. These items were removed from further analysis for those participants. Two of these items were from the re-paired condition in the associative recognition group. This meant that four test trials needed to be

removed to eliminate trials containing either a correctly-guessed target or an associated cue. Thus, 11 test trials were removed in total.

In the target recognition test, the foils were correctly identified as novel, on average, on 90.00% of trials (*SEM* = 1.80%). Figure 3 shows the associative and target recognition test data. A mixed ANOVA on the test format (target recognition, associative recognition) and encoding condition (Study, Meaning) variables revealed a significant main effect of encoding condition, F (1, 54) = 6.55, MSE = 117.56, p = .01, η_g^2 = .03, but not of test format, F < 1. There was a significant test format × encoding condition interaction, F (1, 54) = 11.51, MSE = 117.56, p = .001, η_g^2 = .05. Errorful generation significantly boosted target recognition relative to studying, t (27) = 3.74, p < .001, d_z = 0.71, but not associative recognition, t < 1.

A Bayes factor was calculated to determine whether the associative recognition data support the null hypothesis. We based our prior on the target recognition effect from Experiment 1 (and hence our Gaussian prior had a mean of 11.67% and a standard deviation of 5.84%). The mean difference score in the current associative recognition group was -1.71% (*SEM* = 2.49%), with the Study condition producing numerically better performance than the Meaning condition. These values produced a Bayes factor of 0.05, which supports the null. There was still support for the null (BF < 1/3) with a more conservative prior of a mean difference of 2.8%. Thus, the data support the conclusion that the Meaning condition did not improve associative recognition performance relative to the Study condition.

We also calculated *d'* and response bias (*c*) scores for each recognition group. For the target recognition group, the Meaning condition (M = 2.16, *SEM* = 0.15) produced larger *d'* scores than the Study condition (M = 1.79, *SEM* = 0.18), t (27) = 3.58, p = .001, d_z = 0.68. Response bias scores also differed significantly between encoding conditions; the participants adopted a more liberal strategy for word pairs from the Meaning condition (M = 0.30, *SEM* = 0.66) than word pairs from the Study condition (M = 0.49, *SEM* = 0.06), t (27) = 3.58, p = .001, d_z = 0.68. For the associative recognition

group, *d'* scores did not significantly differ for the Meaning (*M* = 1.14, *SEM* = 0.14) and Study (*M* = 1.17, *SEM* = 0.13) conditions, $t < 1$. Response bias scores, by contrast, did significantly differ between encoding conditions; the participants adopted a more liberal strategy for word pairs from the Meaning condition (*M* = -0.43, *SEM* = 0.08) than word pairs from the Study condition (*M* = -0.11, *SEM* = 0.05), $t(27) = 3.90$, $p < .001$, $d_z = 0.74$.

Discussion

In the target recognition task, participants performed significantly better for targets that they had incorrectly guessed than those that they had studied. This result replicates the target recognition result of Experiment 1, and is broadly consistent with Potts and Shanks' (2014) results. However, the associative recognition test revealed no significant benefit of errorful generation. This result is consistent with the failure to detect benefits of errorful generation in cued recall (Experiments 1 and 2). Together, the results suggest that, relative to an equivalent period of study time, generating errors strengthens the cues and targets in isolation, without simultaneously strengthening the cue-target associations.

The associative recognition group also showed a more liberal response bias for word pairs from the Meaning condition than word pairs from the Study condition. That is, they were more likely to say that the Meaning test pairs were presented together during the encoding phase (compared with the Study condition), regardless of whether those word pairs were paired or re-paired. This result might reflect the fact that the targets from the Meaning condition were better recognised than the targets from the Study condition. That is, the associative recognition group might have tended to say that the Meaning cues and targets were presented together at encoding because they were more familiar than those from the Study condition.

Experiment 4

In Experiment 3, errorful generation produced no significant benefit relative to studying in the associative recognition test. Importantly, the cues and targets were all equally familiar in the

associative recognition test, because they were all presented once during encoding. Potts and Shanks' (2014) multiple-choice tests, by contrast, typically involved asking participants to identify the correct target for a cue from novel foils. Experiment 4 tested whether this difference in foil familiarity is crucial.

Experiment 4 manipulated the familiarity of the foils in a two-alternative forced choice test that was very similar to Potts and Shanks' (2014) multiple-choice tests. Table 1 shows some example trials. Half of the participants were allocated to a Familiar Foils group; the remainder were allocated to an Unfamiliar Foils group. As in the previous experiments, the participants were first presented with rare English words and either guessed the definition before receiving corrective feedback, or studied the word pairs for an equivalent duration. Half of the word pairs in each encoding condition were then allocated to the “test” list – to be presented on test. The remaining word pairs were allocated to a “non-test” list, and were used to provide the foils for the Familiar Foils group on test. In addition, in both groups, a further subset of the overall pool of word pairs (randomly chosen for each participant) was not presented during the encoding phase. The targets from these word pairs would later serve as foils for the Unfamiliar Foils group.

In the two-alternative forced choice test, the cues from the test list were presented with the correct target and a (familiar or unfamiliar) foil. Participants had to select the correct target. For the Unfamiliar Foils group, the foils were targets that were not presented during the encoding phase. For the Familiar Foils group, the foils were targets that were paired at encoding with cues from the *non-test* list. We only tested memory for a subset of the cues (those from the test list) to avoid presenting the targets more than once in the Familiar Foils test. We also used a two-alternative forced choice test (rather than the four-choice test that Potts and Shanks, 2014, employed) so that we were able to test half (rather than one quarter) of the cues. Thus, the test procedure was the same as Potts and Shanks', except that we manipulated the familiarity of the foils, and the participants had to select the correct target for each cue from two rather than four options.

We expected the Unfamiliar Foils group to show a benefit of generating errors over studying. This result would be consistent with both Potts and Shanks' (2014) results and our item recognition results (Experiments 1 and 3). The question of most interest related to the Familiar Foils group. If errorful generation strengthens the cue-target *associations*, then an advantage of generating errors over studying should also be seen in the Familiar Foils group. If, however, errorful generation boosts item but not associative strength (relative to studying), then no such benefit should be seen in the Familiar Foils group. This result would be consistent with the absence of benefits of errorful generation in our cued recall (Experiments 1 and 2) and associative recognition (Experiment 3) tests.

Method

Participants. Ninety-two Plymouth University Psychology undergraduates (21 males and 71 females, aged between 18 and 42, *M* = 20.74 years, *SEM* = 0.53 years) took part for course credit. The participants were randomly allocated to the Familiar Foils (*N* = 46) or Unfamiliar Foils (*N* = 46) group before the experiment. The sample size was determined before data collection, and provides good power (>90%) at the effect sizes observed our Experiments 1 and 3.

Apparatus and materials. One hundred and twenty word pairs were used for the main task, plus 12 additional practice pairs. Potts (2013) provided 118 unique word pairs; 14 pairs were therefore added to the list. All other aspects were identical to that of Experiment 1.

Procedure. As in Experiment 3, each participant completed two encoding conditions (Study and Meaning, counterbalanced for order), then a final test. Each encoding condition consisted of 40 trials, with four practice trials before each. Forty word pairs (randomly chosen for each participant) were not presented at encoding; these words were allocated to the Unfamiliar Foil list. Likewise, a random four practice word pairs were allocated as Unfamiliar foils for the practice test trials.

As described above, half of the cues in each encoding condition (including practice cues) were randomly allocated to a test list. The rest were allocated to a non-test list. The cues from the test list were scheduled to be presented during the subsequent test. All other aspects of the encoding phase were identical to that of the previous experiments.

A two-alternative forced choice test immediately followed the encoding phase. On each trial, a cue from the test list was presented with the question, “What is the correct definition?”, and the target plus a foil. The cue was presented in the top-centre of the screen, the question was presented centrally, and the correct target and a foil were presented beneath the question (side by side). For the Familiar Foils group, the foil was randomly selected from the non-test list (i.e., the list of targets that were presented during encoding but whose cues were not allocated to the test list). Importantly, the foil and target were always from the same encoding condition (Study/Meaning). For the Unfamiliar Foils group, the foil was randomly selected from the targets that were not presented at encoding. The location of the target and foil was counterbalanced across trials. Responses were not time-limited and were made using the mouse. There were 40 test trials, with 20 trials from each encoding condition. The trial order was randomly determined for each participant. Four practice trials, consisting of two test cues from each encoding condition, preceded the main test.

Results

The trial-level raw data for this experiment, including the complete list of word pairs, are publicly available at https://osf.io/q6agj/.

Eight participants guessed the correct target for one cue each at encoding. These items were removed from further analysis on an individual basis. Two of the items were allocated to the non-test list in the Unfamiliar Foils group, and so did not appear in the subsequent test anyway. Hence, only six test trials were removed.

Figure 4 shows the final test data. A mixed ANOVA on the foil familiarity (Familiar Foils, Unfamiliar Foils) and encoding condition (Study, Meaning) factors revealed a significant main effect of foil familiarity, F (1, 90) = 29.59, MSE = 265.22, $p < .001$, $\eta_g^2 = .20$, but not of encoding condition, F (1, 90) = 1.32, MSE = 76.92, $p = .25$, $\eta_g^2 = .003$. There was a significant foil familiarity × encoding condition interaction, F (1, 90) = 5.60, MSE = 76.92, $p = .02$, $\eta_g^2 = .01$. Errorful generation significantly improved final test performance with unfamiliar foils, t (45) = 2.87, $p = .006$, $d_z = 0.42$, but not with familiar foils, $t < 1$.

A Bayes factor was calculated to determine whether the Familiar Foils test data support the null hypothesis. The mean difference score in the Familiar Foils group was -1.57% (*SEM* = 2.04%), with the Study condition producing numerically better performance than the Meaning condition. Using the same prior as Experiment 3, there was substantial support for the null (BF = 0.04). There was still evidence for the null with a more conservative prior of a 2.22% difference (BF < 1/3). Thus, the data support the conclusion that the Meaning condition did not improve final test performance compared to the Study condition in the Familiar Foils group.

Discussion

When participants were asked to select the correct target for a cue from a novel foil (Unfamiliar Foils group), they performed better when they had incorrectly guessed the target during encoding (Meaning condition) than when they had simply studied the word pairs (Study condition). This result is consistent with Potts and Shanks' (2014) results, where errorful generation improved performance on multiple-choice tests relative to studying. It is also consistent with our item recognition results (Experiments 1 and 3), where generating errors improved cue and target recognition compared with simply studying the word pairs. However, errorful generation did not significantly improve performance (relative to studying) when the target and the foil were equally familiar (Familiar Foils group). This result is consistent with our cued recall (Experiments 1 and 2) and associative recognition (Experiment 3) results.

Experiment 5

Experiments 1-4 collectively suggest that Potts and Shanks' (2014) effect will only be seen in tests that reflect item strength. That is, errorful generation does not appear to strengthen associative knowledge. There were, however, some procedural differences between our experiments and Potts and Shanks', which might render this conclusion premature. Perhaps most importantly, we blocked and counterbalanced the order of our encoding conditions. Potts and Shanks, by contrast, interleaved the encoding conditions. It is possible, then, that our failure to see any benefit of errorful generation on tests of associative memory (e.g., cued recall) is simply a consequence of our encoding procedure. Experiment 5 tested whether interleaving the trials from the Meaning and Study conditions at encoding would reveal a benefit of errorful generation on associative memory.

The justification for Experiment 5 assumes that there must be some interaction between the different trials at encoding that impacts on cue-target associative strength. Specifically, perhaps Study trials benefit from appearing among other Study trials in the blocked design (relative to Meaning trials). This is not an unreasonable suggestion; each Study trial is 17 seconds long, and participants are not required to respond at all. On Meaning trials, by contrast, 10 of the 17 seconds is spent guessing the target. It is possible that participants spend a large part of their “spare” time on Study trials rehearsing (and perhaps testing themselves on) the word pairs from the previous few trials. This would result in lots of extra rehearsal time for Study items in our blocked design. In Potts and Shanks’ (2014) interleaved procedure, by contrast, the Meaning and Study items would benefit equally from rehearsal during subsequent Study trials. Of course, if our blocked encoding procedure benefitted the Study condition in terms of cue-target associations, one might wonder why it did not also benefit the encoding of the individual cues and targets, as measured on the item recognition tests of Experiments 1 and 3. Nevertheless, an interleaved encoding procedure, such as that used by Potts and Shanks, might still reveal a benefit for errorful generation on associative memory.

To test the idea above, Meaning and Study trials were interleaved during encoding in Experiment 5. Half of the word pairs in each encoding condition were then tested in a cued recall task. The remainder were tested on a multiple-choice task with novel foils, to replicate Potts and Shanks' (2014) original experiment. We used the procedure of Potts and Shanks' Experiment 2b because it used Euskara-English word pairs. These materials minimise the possibility that participants could have had any prior knowledge of the cue-target pairings; although only a tiny minority of the rare English words were correctly guessed in Experiments 1-4, the correct definitions (targets) may have been partially activated as a consequence of guessing. To be consistent with Potts and Shanks (2014), participants also completed Judgements of Learning (JOLs) after each encoding trial, where they were asked to measure their likelihood of remembering the word pair.

Method

Participants. One hundred and five psychology undergraduates from the University of Bristol completed the experiment for course credit. This sample size has good power to detect within-subject effects of the size seen by Potts and Shanks (2014) in the Euskara-word version of the task used here (89% power at $d_z = 0.28$). Most of the participants reported that they were proficient English language users (self-reported as: 89 native/fluent, 12 excellent/very good, two good, two absent or uncodeable responses). One participant reported being familiar with the Euskara words. This participant was removed from all further analyses. The experiment was approved by the Faculty of Science Research Ethics Board at the University of Bristol.

Apparatus and materials. The experiment was programmed in Blitz3D (Blitz Research, https://blitzresearch.itch.io/blitz3d) and was presented on a 21-inch monitor. The 60 Euskara-English word pairs and 120 English foils that Potts and Shanks (2014) used in their Experiment 2b were randomly assigned to the encoding and test conditions for each participant. Two additional practice word pairs were included. The foils were matched to the targets for frequency and number of syllables. All text was presented in white Arial font in size 24 on a grey background.

Procedure. The experiment followed the same basic procedure as the previous experiments, but we made a number of minor procedural changes to allow maximum compatibility with Potts and Shanks' (2014) Experiment 2b. As in the previous experiments, the participants completed an encoding phase, then a test phase. The encoding phase consisted of 30 Meaning trials and 30 Study trials, which were randomly interleaved. On Study trials, a cue (Euskara noun) and a target (English translation) were centrally presented (e.g., *gazta = cheese*) for 13 seconds (as in Potts and Shanks' Experiment 2b). On Meaning trials, the cue was initially presented with a question mark (e.g., *sagu = ?*). A blinking underscore appeared beneath the question mark, and the participants had eight seconds to guess the target by typing it in. The participants were told that all of the targets were nouns. After the eight seconds, the target was presented with the cue (e.g., *sagu = mouse*) for a further five seconds. The word pair was then replaced by a prompt asking the participants to provide a JOL by rating their likelihood of remembering the item from zero ("No chance I'll remember it") to 100 ("I'll definitely remember it"). The name of the condition was presented above the word pair throughout the encoding phase[6]. The trials were separated by one-second intervals. The encoding phase was preceded by two practice trials, one from each encoding condition.

After the encoding phase, the participants completed a final JOL task, where they gave the percentage of word pairs they thought they would remember from each encoding condition. They also completed a distractor task that lasted for approximately one minute, where they answered randomly generated addition and subtraction calculations, using integers that ranged from 1-100.

The final test then began. Memory for half of the targets in each encoding condition were tested via a multiple-choice test. The rest were tested via a cued recall test. In the multiple-choice test, the cues were presented individually, along with an equality sign and a question mark (e.g.,

[6] The Meaning and Study conditions were referred to as "Generate" and "Read" conditions for the participants, respectively, throughout Experiment 5.

sagu = ?). The target plus the three foils that were created for that target were then presented below the equality sign in a random order. A blinking underscore was also presented below the target/foils and was centred beneath the question mark. The participants were asked to type in the target and press the Enter key to submit their answer. Responding was not time-limited, and the next trial began immediately after completion of the previous trial. The cued recall test was the same as the multiple-choice test, except that the target and foils were not presented. The trials were randomly ordered for each participant, and the two encoding conditions were randomly interleaved. Upon completion of the test phase, the participants completed self-report measures on their language ability and knowledge of Euskara.

Results

The trial-level raw data for this experiment, including the complete list of word pairs, are publicly available at https://osf.io/cbtmk/.

Three participants guessed the correct target for one cue each at encoding. These items were removed from further analysis on an individual basis. For the JOLs, six impossible answers (JOLs > 100) were removed. Trial-by-trial JOLs were higher for word pairs from the Study condition (M = 30.00, SEM = 1.37) than word pairs from the Meaning condition (M = 26.80, SEM = 1.30), t (103) = 4.33, $p < .001$, $d_z = 0.42$. The Meaning (M = 19.86%, SEM = 1.61%) and Study (M = 21.81%, SEM = 1.60%) conditions did not, however, significantly differ on the aggregate, post-encoding JOLs, t (103) = 1.75, $p = .08$, $d_z = 0.17$. We conducted a Bayes analysis on these post-encoding JOLs. In Potts and Shanks' (2014) Experiment 2b, post-encoding JOLs were approximately 13% higher for Study word pairs than the Meaning word pairs. We used this figure (13%) as the mean, and half that figure (6.5%) as the standard deviation, of our Gaussian prior distribution. In our experiment, the mean difference score was 1.95% (SEM = 1.11%), with Study items receiving higher post-encoding JOLs than Meaning items. This calculation produced a Bayes factor of 0.19, which supports the null.

Figure 5 shows the multiple-choice and cued recall test data with a conservative scoring approach (see Experiment 1). On average, responses were submitted on 96.33% of trials. Thus, incorrect responses were rarely errors of omission. A repeated-measures ANOVA on the encoding condition (Meaning, Study) and test format (multiple-choice, cued recall) factors revealed a main effect of test format, F (1, 103) = 2890.78, MSE = 184.09, $p < .001$, $\eta_g^2 = .85$, with participants performing better on the multiple-choice test than the cued recall test. The main effect of encoding condition was not significant, $F < 1$, but there was an encoding condition × test format interaction, F (1, 103) = 22.02, MSE = 86.92, $p < .001$, $\eta_g^2 = .02$. The Meaning condition led to better performance than the Study condition on the multiple-choice test, t (103) = 3.51, $p < .001$, $d_z = 0.34$. Performance on the cued recall test, by contrast, was better for word pairs that were allocated to the Study condition than those that were allocated to the Meaning condition, t (103) = 3.35, $p = .001$, $d_z = 0.33$.

Discussion

In Experiment 5, incorrectly guessing the English translations of Euskara nouns (Meaning condition) improved recognition of those translations in a multiple-choice test with unfamiliar foils, relative to studying the Euskara-English pairs for an equivalent duration (Study condition). Thus, we replicated Potts and Shanks' (2014) key finding from Experiment 2b. The cued recall test, however, revealed the opposite pattern; the Meaning condition produced significantly *worse* cued recall performance than the Study condition. Together, the results suggest that, relative to an equivalent period of time spent studying, errorful generation improves item memory but impairs associative memory.

We also replicated Potts and Shanks' (2014) results with respect to the JOLs given to each word pair during the encoding phase. Like Potts and Shanks' participants, our participants gave higher item JOLs to the word pairs from the Study condition than those from the Meaning condition. In Potts and Shanks' experiments, these JOLs were contradictory to their final test performance, which led the authors to conclude that participants were unaware of the benefits of errorful

generation. In the current experiment, participants' JOLs were inconsistent with their multiple-choice test performance, but were *consistent* with their cued recall performance. We therefore suggest that participants' item JOLs reflect judgements of their knowledge of the *association* between each cue and target, rather than judgements on how well they will remember each cue and target individually.

General Discussion

Five experiments examined the effect of generating errors versus studying when learning the definitions of rare English words and unfamiliar Euskara nouns. Errorful generation followed by corrective feedback boosted recognition of both the rare words (Experiment 1) and the definitions (Experiments 1 and 3). A similar benefit was seen in a two-alternative forced choice test and a multiple-choice test with novel foils (Experiments 4 and 5). However, errorful generation did not improve performance over studying in tests of cued recall (Experiments 1, 2 and 5), associative recognition (Experiment 3), or a two-alternative forced choice test with familiar foils (Experiment 4). Indeed, in Experiment 5, errorful generation produced significantly *worse* cued recall performance than studying. Together, the results clearly demonstrate that generating errors improves recognition of cues and targets in isolation. However, the data provide no support for the notion that errorful generation improves the learning of *associations* between cues and targets, relative to an equivalent period of time spent studying. If anything, errorful generation impairs associative learning[7].

[7] It might be argued that the benefits of generating errors over studying seen in the item recognition (Experiment 1), target recognition (Experiment 3), Unfamiliar Foil choice (Experiment 4) and multiple-choice (Experiment 5) tests might reflect a speed-accuracy trade-off. When we explored participants' reaction times in the final test of each experiment, we found either no significant difference in reaction times to the items in the Meaning and Study conditions (Experiments 3 and 4), or that participants responded more quickly to items in the Meaning condition than items in the Study condition (Experiments 1 and 5). In the cued recall test of Experiment 5, participants responded more quickly to items in the Study condition than items in the Meaning condition. These reaction time results suggest that the accuracy data do not reflect speed-accuracy trade-offs. Analyses of the reaction time data in all experiments are reported in Supplementary Materials C.

Given the failure to observe benefits of errorful generation in cued recall, associative recognition and familiar foil choice tests, it is important to compare Potts and Shanks' (2014) procedure to ours. Importantly, their multiple-choice test foils were novel in their first three experiments. Similar to our item recognition, familiar foil choice, and multiple-choice tests, then, their observed benefit of generating errors over studying is likely to have reflected a boost in item strength rather than associative strength.

In Potts and Shanks' (2014) final experiment, the multiple-choice test foils included the participant's erroneous response from the encoding phase, plus targets from the other encoding conditions. The foils were not, therefore, novel. Crucially, however, foils from the same encoding condition as the correct target were never included. When a cue from the Meaning condition was presented, for example, foils from the Meaning condition were never presented. Our results show that generating errors improves target recognition compared with studying. The targets from the Meaning condition should, then, have been stronger than the other foils in Potts and Shanks' final experiment. If participants' answers were driven by target strength, they would tend to choose the targets from the Meaning condition. This would lead to better performance on the Meaning trials, and worse performance on other trials. Hence, an increase in target strength might also explain why generating errors improved final test performance over studying in Potts and Shanks' last experiment.

It might be argued that our experiments disadvantaged the Meaning condition, because the cues and targets were presented together for much less time than in the Study condition (to be consistent with Potts and Shanks', 2014, experiments). Kornell et al. (2009), for example, found that generating errors was beneficial when the target duration was equated between the Study and Meaning conditions, but not when the total trial duration of the two conditions was equated. Perhaps, then, we would also have seen an associative benefit of errorful generation in our experiments if we had equated the target presentation duration rather than the total trial duration.

Of course, Kornell et al.'s (2009) effects were seen when participants attempted to learn complex fictional trivia questions, while our experiments examined the learning of simple word pairs. There may be important differences in the effects of errorful learning with complex facts and simple word pairs (see Kornell, 2014, for evidence of this). Indeed, when Kornell et al. (2009) used simple, related word pairs such as *pond-frog*, they found that generating errors improved subsequent cued recall over studying, regardless of whether the target presentation or total trial time was equated. In other studies that are more comparable with the present experiments, researchers have, again, consistently found that generating errors does not improve subsequent cued recall for unrelated word pairs, even when the target presentation time was equated (Grimaldi & Karpicke, 2012; Huelser & Metcalfe, 2012; Knight et al., 2012). Thus, while we cannot say with certainty that our effects would remain when target duration is equated, there is little empirical evidence at present to suggest that they would not.

We mentioned in the Introduction that three previous studies failed to detect a benefit of generating errors on cued recall for unrelated word pairs. Those studies led to a consensus that generating errors only benefits the learning of word pairs that have a pre-existing semantic association (Grimaldi & Karpicke, 2012; Huelser & Metcalfe, 2012; Knight et al., 2012). Potts and Shanks' (2014) results spoke against this view, but our results are broadly consistent with it. In our experiments, generating errors boosted performance on tests that reflected item strength (item recognition and unfamiliar foil choice tests), but not on tests that required associative knowledge (cued recall, associative recognition, and familiar foil choice tests). This is exactly the pattern of results seen in the existing literature. Potts and Shanks used a non-associative test (unfamiliar foil choice test) and found a benefit of guessing with novel word pairs. Grimaldi and Karpicke (2012), Huelser and Metcalfe (2012) and Knight et al. (2012), by contrast, used associative tests (cued recall) and found no such benefit for unrelated word pairs.

Given what we now know about the effects of errorful generation, what can we say about the underlying mechanisms? We begin with our recognition data. Search set theory (e.g., Grimaldi & Karpicke, 2012) suggests that generating errors can strengthen targets in memory. It therefore predicts that errorful generation might improve item recognition (as in Experiments 1 and 3). However, search set theory also predicts that generating errors should only strengthen target memory when the cue and target have a pre-existing semantic association, because the correct target should only be available for activation in the search set under these circumstances. Our recognition results with unrelated word pairs therefore speak against search set theory; errorful generation increased target strength even though the cue and the target were unrelated.

Our data also cast doubt on Potts and Shanks' (2014) analysis of the effects of errorful generation. It is clear that Potts and Shanks' (2014) effect did not relate to the learning of word *pairs* at all, but rather just to memory of the individual components of each pair. Nevertheless, the mechanisms they proposed to explain (what they implied was) the associative effects of errorful generation apply equally well to the effects of errorful generation on item strength. They first suggested that incorrect guesses might produce surprise when the correct target is revealed. This prediction error should then boost learning (following Rescorla & Wagner, 1972). They also suggested that guessing might foster more interest in the correct answer. In light of our data (see also Potts, Davies, & Shanks, 2018), the "general interest" explanation seems preferable to the "error correction" account. It is not obvious why an error correction mechanism (which produces surprise and greater attention to the target), would also improve recognition of the cue (as in Experiment 1)[8]. In contrast, a generalised increase in interest could have a more widespread effect, thereby improving both cue and target recognition. In general, a boost in attention to, or interest in,

[8] Speculatively, one might appeal to an attentional-associative process, such as that described by Pearce and Hall (1980), where prediction errors increase the salience of the cue. However, the emerging consensus is that the opposite (such as that described by Mackintosh, 1975) more typically happens, at least in adult humans (see Le Pelley, Mitchell, Beesley, George, & Wills, 2016, for a review).

the cue and the target provides a straightforward explanation of the effects of generating errors on recognition performance.

What remains to be explained are the effects of unsuccessful retrieval on cued recall (Grimaldi & Karpicke, 2012; Huelser & Metcalfe, 2012; Knight et al., 2012). Grimaldi and Karpicke (2012) found that generating errors improved subsequent cued recall for related, but not unrelated, word pairs. Search set theory was the favoured explanation of those data. However, search set theory is not supported by the current findings because it predicts that item recognition for unrelated word pairs (like those used in the current experiments) should not benefit from errorful generation.

Why else might “pond-frog” benefit from errorful generation in cued recall, while “pond-spanner” does not? At this point we can only speculate. Our data suggest, however, that both of the target items “frog” and “spanner” are boosted at encoding from guessing, perhaps due to greater interest on the part of the participant. Grimaldi and Karpicke's (2012) effect might then rely on an interaction between item strength and semantic relatedness in recall. Taking the example above, the cue “pond” will activate (from semantic memory) a range of potential targets such as “frog”, “lily” and “weed”. It is much less likely, however, to activate “spanner”. If one of the activated items (e.g., “frog”) has been especially well strengthened – because it was presented on a Meaning trial at encoding – then it is more likely to be chosen from among the semantically activated candidates at test. The unrelated target “spanner”, however, benefits less from this process, because it is not semantically activated in the first instance. This analysis suggests that the effect of generating errors on cued recall for related pairs is not simply associative, but rather a combination of both item and associative strength. This account provides a unified explanation of performance on both non-associative (e.g., item recognition) and associative (e.g., cued recall) tests; incorrect guesses boost item strength in both cases.

Another possibility is that participants' guesses serve as mediators (Pyc & Rawson, 2010). Participants given the cue "pond" might guess "water" before being given the correct target "frog". When presented with "pond" on test, "frog" might be easier to recall because "water" is semantically related to frog". In the case of "pond-spanner", however, the guess "water" is unlikely to aid retrieval of "spanner" because these words are unrelated. In fact, the unrelated guess might interfere with retrieval of the target. Associative mediation of this kind would not be expected to play a role in the item recognition tests used in our experiments, and so semantic relatedness would not be expected to be crucial to our effects. This second analysis therefore implies that the mechanisms underlying cued recall (associative mediation) and recognition (item strength) are quite different. Both of the explanations of the effects of errorful generation outlined here (item-strength and guess-mediation) are consistent with all of the currently available data.

To conclude, the current experiments tested the effect of generating errors when learning novel word pairs. Errorful generation boosted both cue and target recognition relative to studying alone. Similar benefits were observed in a two-alternative forced choice test with novel foils. However, no such benefits were observed in cued recall, associative recognition or familiar foil choice tests. Together, the results demonstrate that errorful generation strengthens the cues and targets in isolation, but not their associations.

Funding

This work was supported by an Economic and Social Research Council grant (ES/N018702/1) to Timothy J. Hollins, Chris J. Mitchell and Andy J. Wills.

References

Ariel, R., & Karpicke, J. D. (2017). Improving self-regulated learning with a retrieval practice intervention. *Journal of Experimental Psychology: Applied*, *24*, 43–56. https://doi.org/10.1037/xap0000133

Baddeley, A., & Wilson, B. A. (1994). When implicit learning fails: Amnesia and the problem of error elimination. *Neuropsychologia*, *32*, 53–68. https://doi.org/10.1016/0028-3932(94)90068-X

Baguley, T. (2012). Calculating and graphing within-subject confidence intervals for ANOVA. *Behavior Research Methods*, *44*, 158–175. https://doi.org/10.3758/s13428-011-0123-7

Carpenter, S. K., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued recall test? *Psychonomic Bulletin & Review*, *13*, 826–830. https://doi.org/10.3758/BF03194004

Clark, C. M. (2016). *When and why does learning profit from the introduction of errors?* University of California.

Clark, S. E. (1992). Word frequency effects in associative and item recognition. *Memory & Cognition*, *20*, 231–243. https://doi.org/10.3758/BF03199660

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*, 781. https://doi.org/10.3389/fpsyg.2014.00781

Edmunds, C. E. R., Milton, F., & Wills, A. J. (2015). Feedback can be superior to observational training for both rule-based and information-integration category structures . *The Quarterly Journal of Experimental Psychology*, 1203–1222. https://doi.org/10.1080/17470218.2014.978875

Edmunds, C. E. R., Wills, A. J., & Milton, F. (2018). Initial training with difficult items does not facilitate category learning. *Quarterly Journal of Experimental Psychology*. https://doi.org/https://doi.org/10.1080/17470218.2017.1370477

Glenberg, A. M., & Bradley, M. M. (1979). Mental contiguity. *Journal of Experimental Psychology:*

Human Learning & Memory, *5*, 88–97. https://doi.org/10.1037//0278-7393.5.2.88

Grimaldi, P. J., & Karpicke, J. D. (2012). When and why do retrieval attempts enhance subsequent encoding? *Memory & Cognition*, *40*, 505–513. https://doi.org/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. https://doi.org/10.1037/a0028468

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

Jeffreys, H. (1961). *The theory of probability* (3rd ed.). Oxford: Oxford University Press.

Kane, J. H., & Anderson, R. C. (1978). Depth of processing and interference effects in the learning and remembering of sentences. *Journal of Educational Psychology*, *70*, 626–635. https://doi.org/http://dx.doi.org/10.1037/0022-0663.70.4.626

Karpicke, J. D., Butler, A. C., & Roediger III, H. L. (2009). Metacognitive strategies in student learning: Do students practise retrieval when they study on their own? *Memory*, *17*, 471–479. https://doi.org/10.1080/09658210802647009

Knight, J. B., Ball, B. H., 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. https://doi.org/10.1016/j.jml.2011.12.008

Kornell, N. (2014). Attempting to answer a meaningful question enhances subsequent learning even when feedback is delayed. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*, 106–114. https://doi.org/10.1037/a0033699

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

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

Kornell, N., & Son, L. K. (2009). Learners' choices and beliefs about self-testing. *Memory*, *17*, 493–501. https://doi.org/10.1080/09658210902832915

Kornell, N., & Vaughn, K. E. (2016). How retrieval attempts affect learning: A review and synthesis. *Psychology of Learning and Motivation*, *65*, 183–215. https://doi.org/10.1016/bs.plm.2016.03.003

Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review*, *82*, 276–298. https://doi.org/10.1037/h0076778

McCabe, J. (2011). Metacognitive awareness of learning strategies in undergraduates. *Memory & Cognition*, *39*, 462–476. https://doi.org/10.3758/s13421-010-0035-2

Metcalfe, J. (2017). Learning from errors. *Annual Review of Psychology*, *68*, 465–489. https://doi.org/10.1007/BF01457248

Middleton, E., & Schwartz, M. (2012). Errorless learning in cognitive rehabilitation: A critical review. *Neuropsychological Rehabilitation*, *22*, 37–41. https://doi.org/https://doi.org/10.1080/09602011.2011.639619

Pashler, H., Rohrer, D., Cepeda, N. J., & Carpenter, S. K. (2007). Enhancing learning and retarding forgetting: Choices and consequences. *Psychonomic Bulletin & Review*, *14*, 187–193. https://doi.org/10.3758/BF03194050

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, *87*, 532–552. https://doi.org/10.1037/0033-295X.87.6.532

Potts, R. (2013). *Memory interference and the benefits and costs of testing*. University College London.

Potts, R., Davies, G., & Shanks, D. R. (2018). The benefit of generating errors during learning: What is the locus of the effect? *Journal of Experimental Psychology: Learning Memory and Cognition*. https://doi.org/dx.doi.org/10.1037/xlm0000637

Potts, R., & Shanks, D. R. (2014). The benefit of generating errors during learning. *Journal of Experimental Psychology: General*, *143*, 644–667. https://doi.org/10.1017/CBO9781107415324.004

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science*, *330*, 335. https://doi.org/10.1126/science.1191465

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical Conditioning II Current Research and Theory*, *21*, 64–99. https://doi.org/10.1101/gr.110528.110

Richland, L. E., Kornell, N., & Kao, L. S. (2009). The pretesting effect: do unsuccessful retrieval attempts enhance learning? *Journal of Experimental Psychology: Applied*, *15*, 243–257. https://doi.org/10.1037/a0016496

Roediger III, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*, 181–210. https://doi.org/10.1111/j.1467-8721.2008.00612.x

Skinner, B. F. (1958). Teaching machines. *Science*, *128*, 969–977. https://doi.org/10.1109/TE.1959.4322064

Slamecka, N. J., & Fevreiski, J. (1983). The generation effect when generation fails. *Journal of Verbal Learning and Verbal Behavior*, *22*, 153–163. https://doi.org/10.1016/S0022-5371(83)90112-3

Szpunar, K. K., McDermott, K. B., & Roediger, H. L. I. (2008). Testing during study insulates against the buildup of proactive inteference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 1392–1399. https://doi.org/10.1037/a0014896

Terrace, H. S. (1963). Discrimination learning with and without "errors." *Journal of the Experimental Analysis of Behavior*, *6*, 1–27. https://doi.org/10.1901/jeab.1963.6-1

Vaughn, K. E., & Rawson, K. A. (2012). When is guessing incorrectly better than studying for enhancing memory? *Psychonomic Bulletin & Review*, *19*, 899–905. https://doi.org/10.3758/s13423-012-0276-0

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. https://doi.org/10.3758/s13421-014-0454-6

Yang, C., Potts, R., & Shanks, D. R. (2017). Metacognitive unawareness of the errorful generation benefit and its effects on self-regulated learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*, 1073–1092. https://doi.org/10.1037/xlm0000363

Tables

Table 1

Example trials from Experiment 4.

Encoding	Test	
	Familiar foils	**Unfamiliar foils**
Meaning: roke – mist	roke: mist or fussy?	roke: mist or flimsy?
Meaning: spoffish – fussy	picaroon: cheat or fruit?	picaroon: cheat or fish?
Study: picaroon – cheat		
Study: achene – fruit		

Note. During encoding, participants first attempted to learn the definitions of rare English words by either studying them (Study condition) or attempting to guess the meaning before the true definition was revealed (Meaning condition). In a subsequent two-alternative forced choice test, the participants had to select the correct target for each cue from two options. In the Familiar Foils group, the correct target was placed among another target from the same encoding condition. In the Unfamiliar Foils group, the correct target was placed among a foil that was not presenting during the encoding phase.

Figure 1

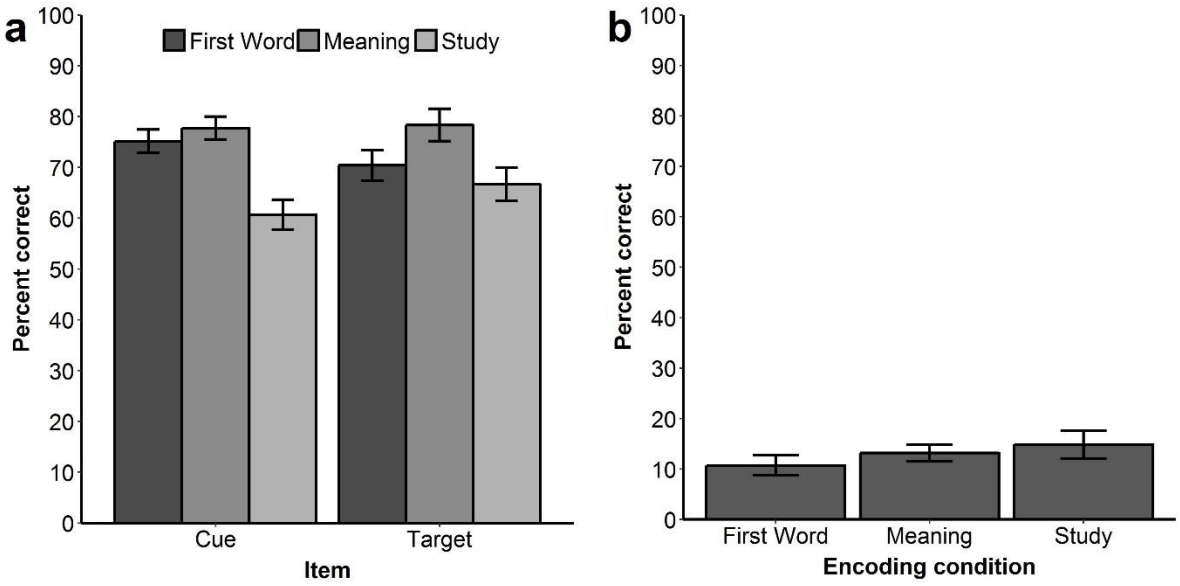

Figure 1. Mean percentage correct in the **(a)** item recognition and **(b)** cued recall test of Experiment 1. Error bars are difference-adjusted within-subject 95% confidence intervals (Baguley, 2012).

Figure 2

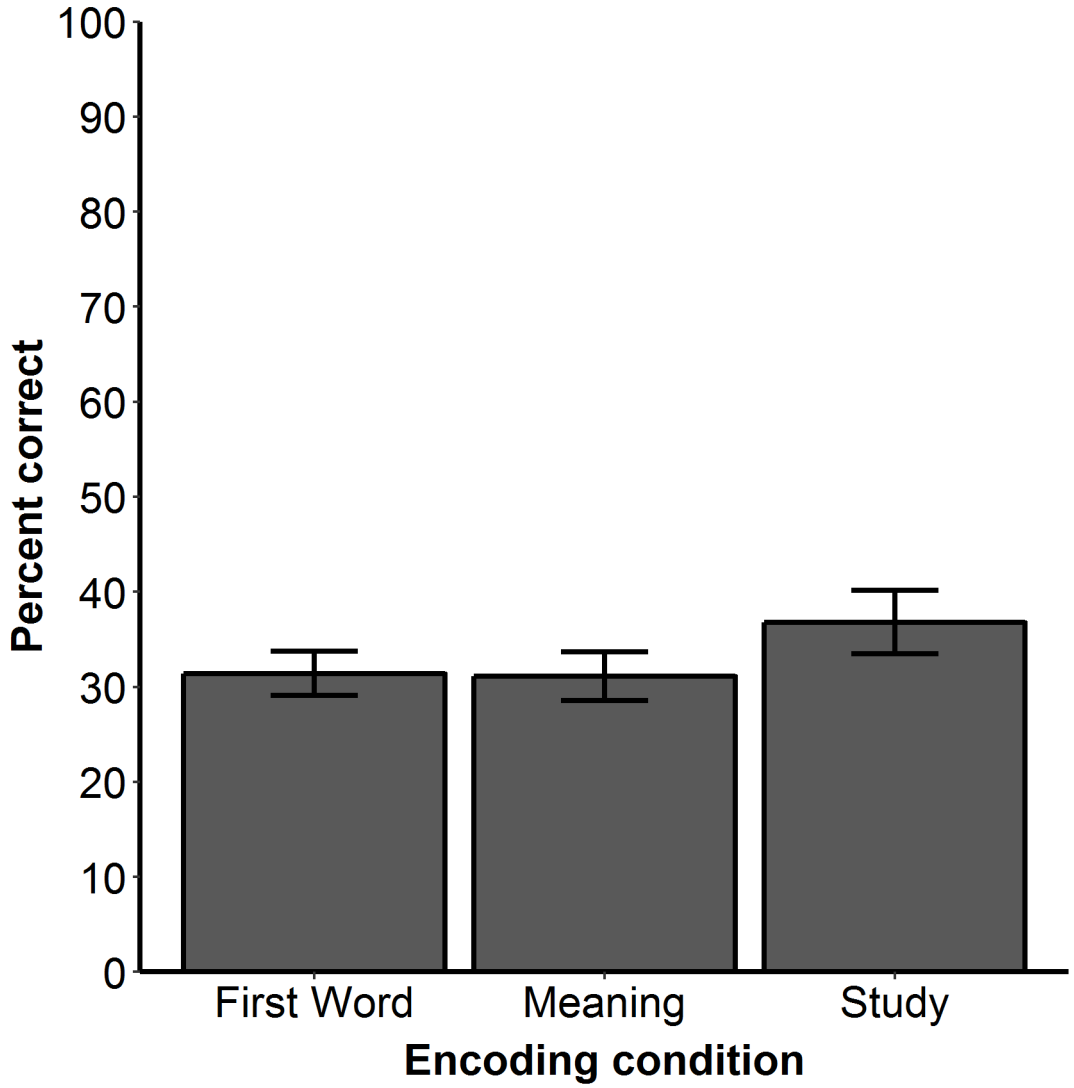

Figure 2. Mean percentage correct in the recall test of Experiment 2. Error bars are difference-adjusted within-subject 95% confidence intervals (Baguley, 2012).

Figure 3

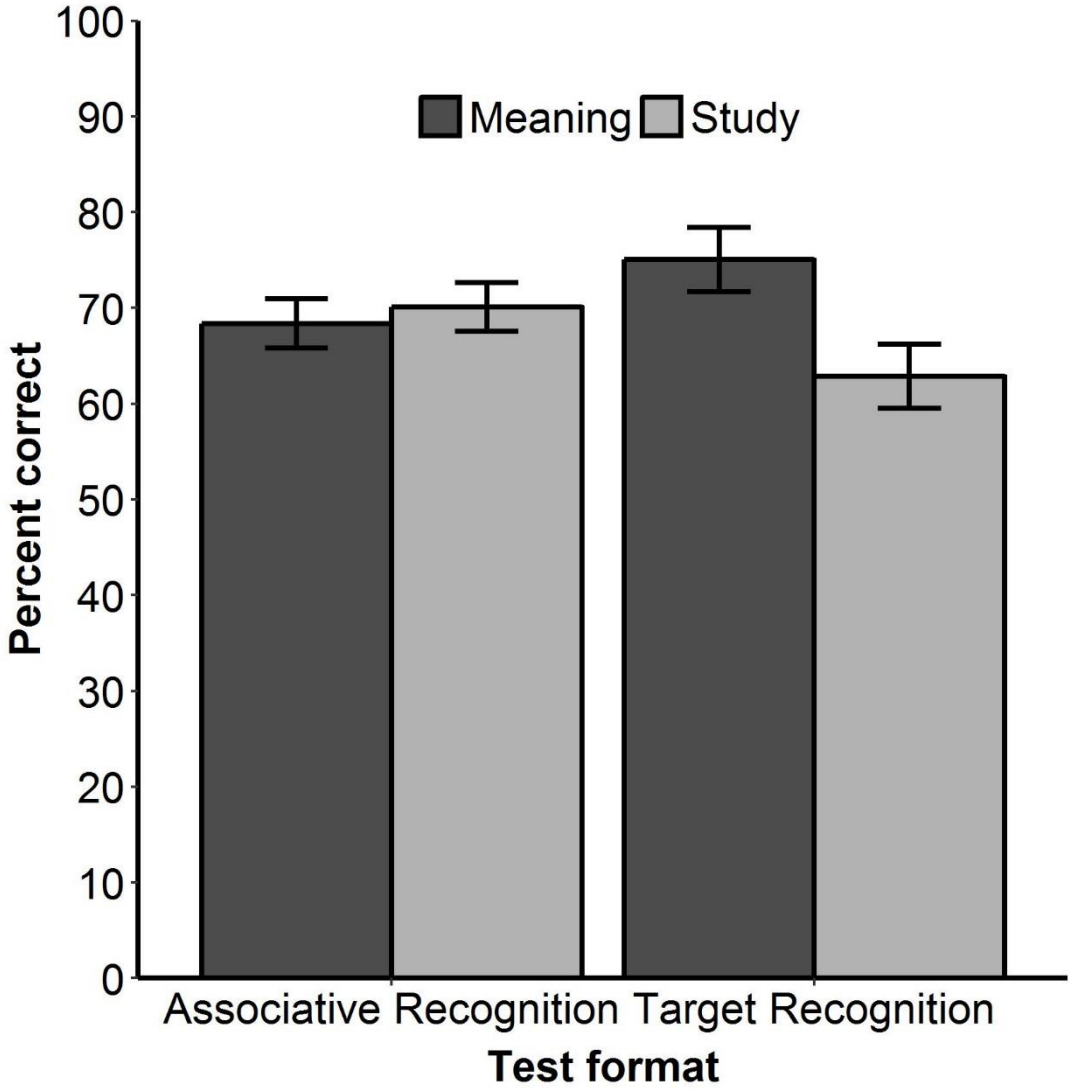

Figure 3. Mean percentage correct in the associative and target recognition tests of Experiment 3. Error bars are difference-adjusted within-subject 95% confidence intervals (Baguley, 2012).

Figure 4

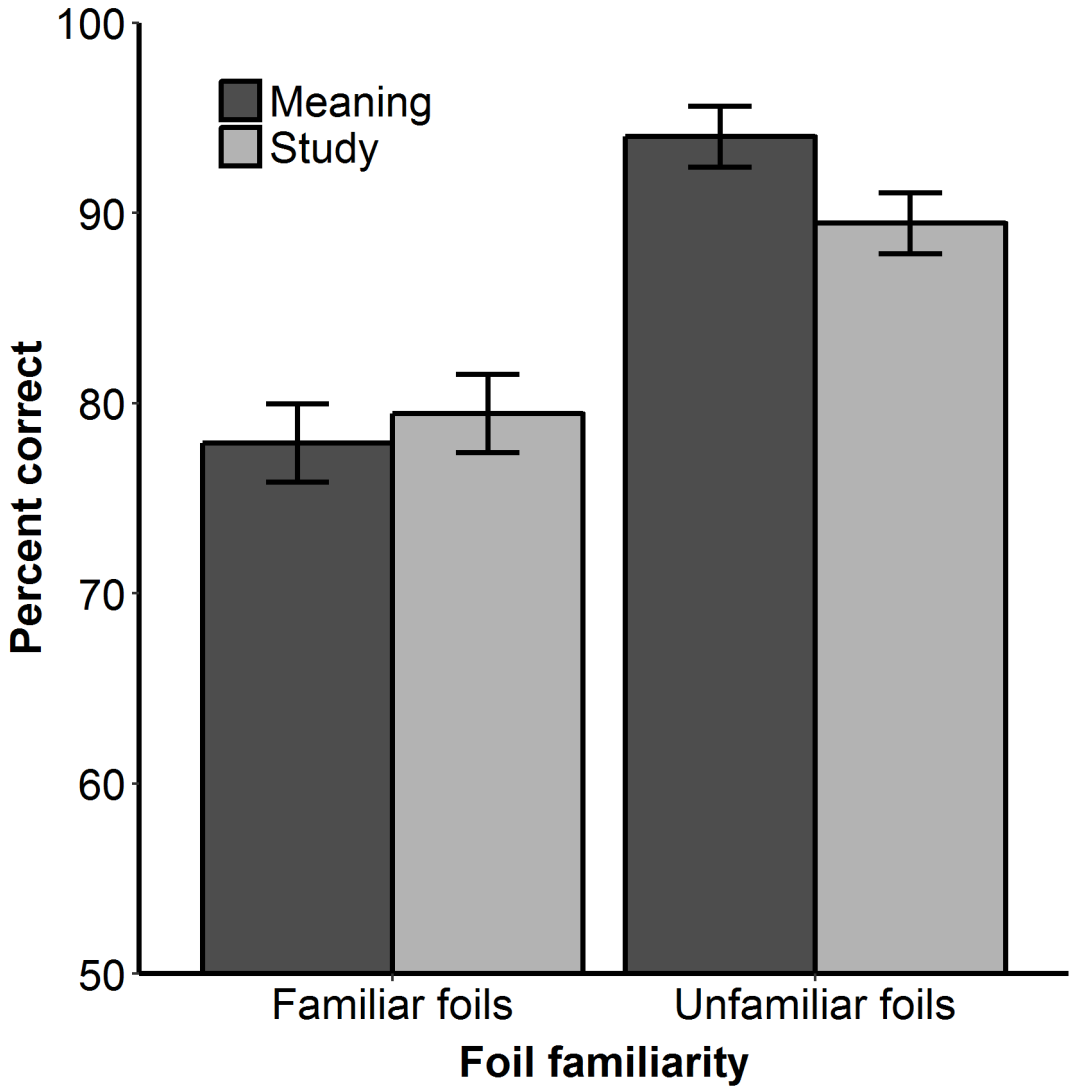

Figure 4. Mean percentage correct in the Familiar and Unfamiliar foils groups of Experiment 4. Error bars are difference-adjusted within-subject 95% confidence intervals (Baguley, 2012).

Figure 5

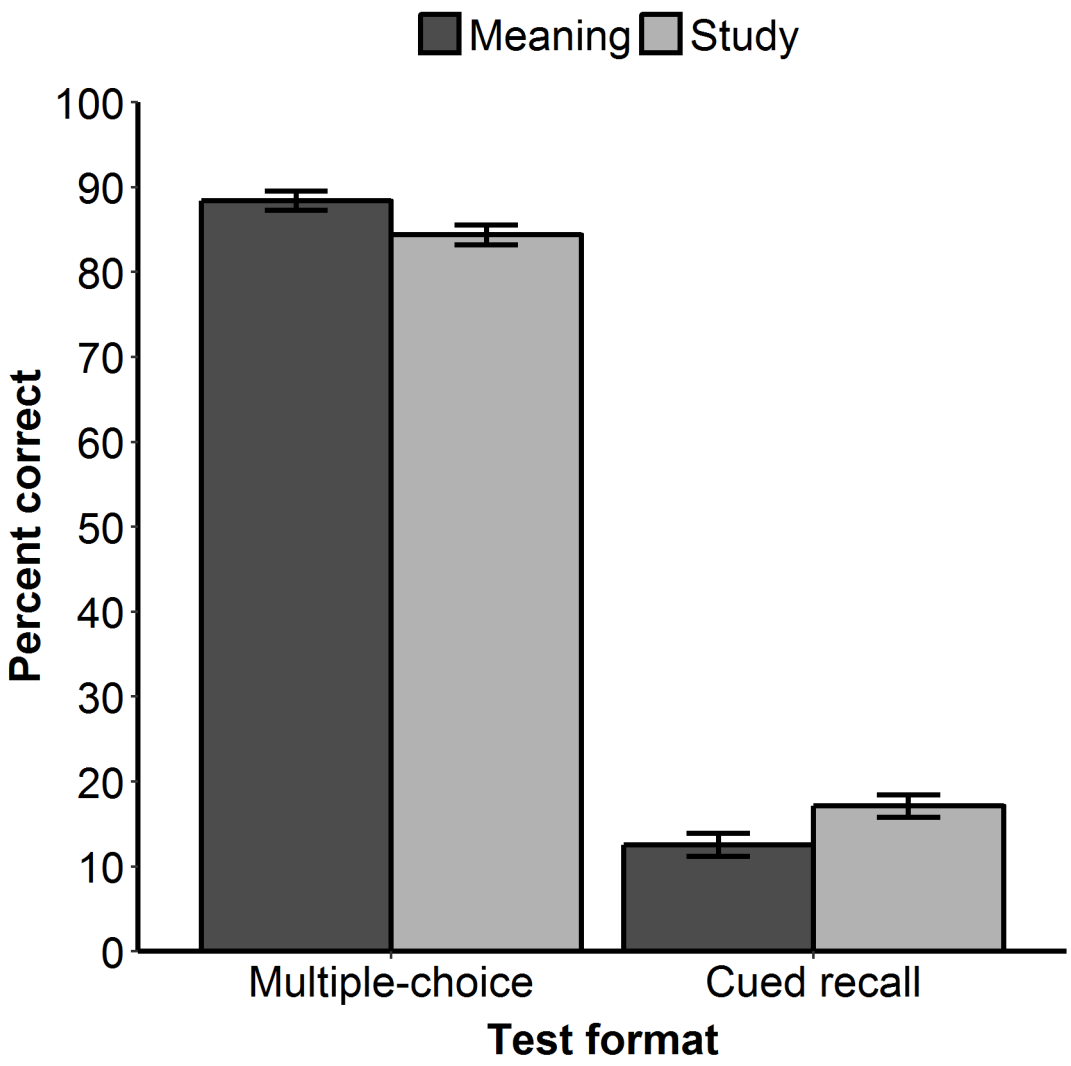

Figure 5. Mean percentage correct on the multiple-choice and cued recall tests of Experiment 5. Error bars are difference-adjusted within-subject 95% confidence intervals (Baguley, 2012).