The benefit of generating errors during learning.

Rosalind Potts and David R. Shanks

University College London

Author Note

Rosalind Potts, Department of Cognitive, Perceptual and Brain Sciences, University College London; David R. Shanks, Department of Cognitive, Perceptual and Brain Sciences, University College London.

This research was supported by the UK Economic and Social Research Council (studentship ES/H036237/1).

We thank Sevde Inan for assistance with data collection in Experiment 2B, Chris Berry and Joe Devlin for useful suggestions regarding this research, and John Dunlosky and an anonymous reviewer for helpful comments on an earlier version of this article.

Correspondence concerning this article should be addressed to Rosalind Potts, Department of Cognitive, Perceptual and Brain Sciences, University College London, Gower Street, London WC1E 6BT. Email: rosalind.potts@ucl.ac.uk

Abstract

Testing has been found to be a powerful learning tool but educators may be reluctant to make full use of its benefits for fear that any errors made will be harmful to learning. We asked whether testing could be beneficial to memory even during novel learning, when nearly all responses were errors, and where errors were unlikely to be related to either cues or targets. In four experiments participants learned definitions for unfamiliar English words, or translations for foreign vocabulary, either by generating a response and being given corrective feedback, by reading the word and its definition/translation, or by selecting from a choice of definitions/translations followed by feedback. In a final test of all words, generating errors followed by feedback led to significantly better memory for the correct definition/translation than either reading or making incorrect choices, suggesting that the benefits of generation are not restricted to correctly generated items. Even when information to be learned is novel, errorful generation may play a powerful role in potentiating encoding of corrective feedback. Experiments 2A, 2B, and 3 revealed, via metacognitive judgments of learning, that participants are strikingly unaware of this benefit, judging errorful generation to be a less effective encoding method than reading or incorrect choosing when in fact it was better. Predictions reflected participants' subjective experience during learning. If subjective difficulty leads to more effort at encoding, this could at least partly explain the errorful generation advantage.

Keywords: learning, education, errors, generation, metacognition

The benefit of generating errors during learning.

A central question for educators concerns how to maximise students' retention of learned information. One technique which has been shown to be highly effective is the use of testing: A robust and highly replicated finding from both laboratory and classroom studies is that the very act of retrieving items from memory enhances memory for the tested items, the "testing effect" (see Roediger & Karpicke, 2006a, for a review). Simply inserting tests into the learning process therefore has the potential to provide a powerful boost to the amount of information retained. Indeed, the use of testing to promote learning was one of seven recommendations for educational practice made in a recent guide produced for the US Government (Pashler, Bain, Bottge, Graesser, McDaniel, & Metcalfe, 2007), the seven recommendations being based on "the most important, concrete and applicable principles to emerge from research on learning and memory" (p1). Moreover, it has been found that the harder the test, and the greater the effort required for retrieval, the greater the benefit to subsequent memory (e.g., Carpenter & DeLosh, 2006; Pyc & Rawson, 2009). The most benefit is therefore to be gained by setting a difficult test.

However, a difficult test brings with it the risk that the learner may make many errors and educators may be concerned that these errors will be reinforced by the act of testing, with a consequential harmful effect on learning, a concern which may deter them from making optimal use of testing as a learning tool. Such a concern is not unreasonable in the light of evidence that errors are best avoided during learning (e.g., Baddeley & Wilson, 1994). On the other hand, there is also evidence that generating responses can be beneficial even when many errors are produced, as long as corrective feedback is given (e.g., Kornell, Hays & Bjork, 2009). A worthwhile goal, then, is to identify the conditions in which errorful generation may be either helpful or harmful to subsequent retention. The current article seeks to contribute towards achieving this goal.

The prevailing view is that a benefit of errorful generation only occurs when there is a pre-existing semantic association between cue and target. If this is the case, this could limit the usefulness of testing in situations where errors are likely to be made, but this view is based on just a handful of recent studies, all of which have used artificial tasks and materials which are rather different from those likely to be encountered during real world learning, and it remains to be seen whether an errorful generation benefit could occur in a more typical educational scenario in which students are learning novel information. An important issue, therefore, is to understand more fully the effects of generating errors on memory, and to do so using educationally relevant materials such as might be encountered during real world learning. In the current study we examined the effect of generating errors during the learning of previously unfamiliar vocabulary items, where there were no pre-existing relationships between the cues and targets. To foreshadow, we found that generation can be beneficial to memory even when it produces many errors and even when information to be learned is novel: A pre-existing semantic association between cue and target is not necessary for the benefit to occur.

The testing effect is the benefit to memory of taking a test of previously studied material compared with re-reading it. Generating items in response to a cue (e.g., “opposite of hot: c___”), also leads to better memory for the generated items than simply reading them, the generation effect (Slamecka & Graf, 1978). These two highly replicated findings suggest that there is something about the active process of recalling or generating which leads to memory enhancement for the recalled or generated items. But what happens when we generate errors on a test? Are those errors strengthened by generation, leading to impaired memory for correct information? Or can the active process of generation, even when it produces an error, lead to better retention as long as corrective feedback is given?

There are two scenarios in which learners may guess incorrectly in response to a test question. First, they may know the answer but either be temporarily unable to retrieve it or may retrieve the wrong answer. In this case there is a pre-existing association between the question and the correct answer at the time of initial retrieval, and corrective feedback may be used to reinforce this association, to maximise the chance of successful retrieval on future occasions. This scenario we call "unsuccessful retrieval" (following Kornell, Hays, & Bjork, 2009) and is one which has been the focus of some recent studies (Grimaldi & Karpicke, 2012; Hays, Kornell & Bjork, 2013; Huelser & Metcalfe, 2012; Knight, Hunter Ball, Brewer, DeWitt, & Marsh, 2012; Kornell, Hays, & Bjork, 2009). In the second scenario, learners do not know the answer because the test material is completely new to them. In this scenario they may generate a guess which is more or less plausible depending on the constraints provided by the available context, such as the test question itself. In this case, to learn the correct answer, the individual has to make a novel association between the unfamiliar cue material and the corrective feedback provided – there is no pre-existing association to be reinforced. This scenario, which we call "errorful generation", has received less attention and is the focus of the current article. In the four experiments reported here, we examined the effect of making errors in a vocabulary learning task in a situation in which learners make incorrect guesses not because they cannot remember the answer but because they have never learned it in the first place.

Kay (1955) noted the difficulty his participants had in "amending the mistakes which they themselves had introduced into their learning" (p.81). Indeed, a large body of literature on 'errorless' learning has proposed that errors generated during learning can have a detrimental effect on later memory performance. Errorless learning studies typically compare a condition in which participants are encouraged to generate many erroneous responses to a test cue with a condition in which they are presented with the correct answer intact, with the

latter proving more beneficial to later memory. Although the avoidance of errors has been particularly advocated for people with memory impairments (Baddeley & Wilson, 1994), an advantage for errorless over errorful learning has frequently also been observed in healthy young people, with a variety of materials (e.g., Hammer, Kordon, Heldmann, Zurowski, & Munte, 2009, for verbal materials; Haslam, Moss, & Hodder, 2010, for greeble-name associations; Haslam, Hodder, & Yates, 2011, for face-name associations; Kessels, Boekhorst, & Postma, 2005, for spatial locations). Participants often remember their own erroneous responses rather than the correct responses provided by the experimenter. Errorful learning is thought to be detrimental to memory because errors can prove remarkably resistant to correction even when there are multiple opportunities to review the correct information (e.g., Fritz, Morris, Bjork, Gelman, & Wickens, 2000).

On the other hand, when material has been previously studied, there is plentiful evidence that testing can enhance memory relative to re-reading, at least when material is successfully retrieved at initial test (e.g. Allen, Mahler & Estes, 1969; Carrier & Pashler, 1992; Hogan & Kintsch, 1971; Karpicke & Roediger, 2008). The testing effect is observed even when no corrective feedback is given (Allen et al., 1969; Carpenter & DeLosh, 2005, 2006; Kuo & Hirshman, 1996), so it cannot be attributed solely to more efficient processing of feedback in the test condition. However, there is evidence that feedback may enhance the benefits of tests (Butler & Roediger, 2008; Pashler, Cepeda, Wixted & Rohrer, 2005).

There is no consensus as to why the testing benefit occurs but explanations include the idea that testing increases the storage strength of the memory (Bjork & Bjork, 1992; Whitten & Bjork, 1977) or strengthens retrieval routes between the cue and the target (Bjork, 1975). Alternatively, tests may benefit later recall because the act of retrieval involves generating additional cues which elaborate the memory trace and create more routes to retrieval on a subsequent occasion (Carpenter & DeLosh, 2006; Glover, 1989). The

generation effect (Slamecka & Graf, 1978) has been attributed to a similar mechanism, in other words it has been interpreted as a testing effect for semantic memory. When the participant is asked to generate the opposite of *hot*, the association *hot-cold* already exists in memory and the act of generating the target from the cue strengthens the memory in the same way as it does for studied material in a typical testing effect study. The finding that there is no generation benefit when the response terms are nonwords (McElroy & Slamecka, 1982) also suggests that the effect may be the result of enhanced elaborative processing.

Unsuccessful retrieval

Testing typically enhances memory for items which are successfully retrieved on the test but there is also evidence that even tests which yield errors can benefit later retention, as long as corrective feedback is given. Izawa (1970) suggested that tests may potentiate subsequent encoding of the correct response, and Bahrick and Hall (2005) proposed that retrieval failures benefit long term recall by allowing participants to identify items which were inadequately encoded and therefore to focus more time and attention on encoding corrective feedback. Testing can therefore enhance memory directly, by strengthening the generated or retrieved memory, or indirectly, by making the processing of subsequent feedback more effective, or by some combination of the two (see Arnold & McDermott, 2012, for a useful discussion of this point).

Slamecka and Fevreiski (1983) investigated the effect on memory of generation failures by having participants generate relatively unfamiliar opposites, such as *vital* in response to the cue *trivial – v.* Stimuli were designed such that there was only one acceptable response which fitted the cue. Slamecka and Fevreiski observed that, even when many errors were made at study, generation led to better memory on a later test than reading. They proposed that, at study, participants had retrieved semantic attributes of the target but not its

surface features, leading to apparent failure at study but facilitating performance on the subsequent test since the surface features had been supplied as feedback. This conclusion was supported by a third experiment showing that, even without feedback at study, performance on a forced-choice recognition test of the targets was above chance for items which had not been successfully generated at study, suggesting that these items had in fact been partially retrieved.

A study by Kane and Anderson (1978) found a benefit of generating errors over reading when participants were instructed to supply the last word of a sentence or to read the sentence intact. For determined sentences (e.g., "The dove is a symbol of __" [*answer: peace*]), the correct answer was obvious from the sentence, whereas for undetermined sentences (e.g., "The physician asked the patient if he had a ____" [*answer: watch*]), it was not. Even in the undetermined condition, where participants nearly always produced an error, generating led to better final test performance than reading. In this case, it seems less likely that the correct answer was, in fact, partially activated since there was not just one but many possible answers which fitted the cue and the most obvious completions were unrelated to the one designated as correct. Kane and Anderson suggested that the benefit of errorful generation was due to the requirement to process the sentence meaningfully, which was unnecessary in the Read condition.

Whereas in the typical generation effect paradigm there is only one answer which fits the cue, Kane and Anderson's task made it possible to respond with many plausible completions. In their study, therefore, the goal was not to retrieve a sole valid correct answer but, rather, to guess which of many possible responses the experimenter happened to have in mind. This design has been adopted in a handful of recent studies (Grimaldi & Karpicke, 2012; Huelser & Metcalfe, 2012; Kornell, Hays, & Bjork, 2009) as a means of simulating the processes involved in unsuccessful retrieval, that is, the situation where a student has learned

the right answer but retrieves the wrong one. Typically, studies investigating the effect of errors in this type of situation have had participants study the material before being tested on it (e.g., Pashler, Zarow, & Triplett, 2003; S. H. K. Kang, Pashler, Cepeda, Rohrer, Carpenter, & Mozer, 2011). Following study, participants take an initial test with feedback and, later, a final test. Final test performance can then be analysed conditional upon the making of an error at initial test. However, as Pashler et al. (2003) have noted, this design incurs item-selection problems. If an error is made on an initial test, and it is made again on a later test, this could be either because the original error had a deleterious effect on later memory or because the item was intrinsically difficult to learn. While it is possible to examine later test performance for just the subset of items which were incorrect at initial test, it is not possible to compare this with performance for items which were not tested but which would have been incorrect had they been tested, since there is no way of determining which these are.

In order to overcome this item-selection problem, Kornell et al. (2009) eliminated the usual study phase in which to-be-learned associations are studied, starting instead with the initial test, and they selected materials (weakly-associated word pairs, e.g., *pond - frog*) for which the cues would have strong pre-existing associations with items other than the target. This method was designed to encourage participants to attempt retrieval of an existing association while ensuring that many "errors" (i.e., responses different from the target) would be produced. Thus the terms "unsuccessful retrieval" and "retrieval failures" used by Kornell et al. refer not to a failure to retrieve an episodic association between cue and target formed during an earlier study phase, since there was no study phase, but rather to the retrieval by participants of a pre-existing semantic association which differed from the one designated as "correct" by the experimenter. In this way Kornell et al. aimed to simulate a situation in which students retrieve, during a test, an answer which is incorrect but which is related to the

correct one, such as might occur when a student has studied something but has not learned it with sufficient thoroughness.

In the first phase of Kornell, Hays, and Bjork's procedure (2009, Exps 4 - 6), participants were shown a cue word (e.g., *pond)* and were instructed to produce an associate. Typically, participants would produce a strong associate to the cue (e.g., *water*) and were then told the particular associate that the experimenter had in mind (*frog*) and were instructed to remember that item for a later test. Because the correct targets were only weakly associated to the cue, participants typically failed to guess them, thus ensuring that many "errors" were produced. These "*test*" trials were interleaved with "*read-only*" trials in which intact cue-target pairs were presented. At final test participants were again given the cue *pond* but this time their task was to recall the particular associate they had been instructed to study in the first phase (*frog*). Kornell et al. found that the *test* condition led to better final test performance than the *read-only* condition. In their experiments, the instruction to produce an associate constrained guesses to items likely to be highly related to both cue and target. Use of associated pairs ensured there was a pre-existing association between cue and target (*pond – frog)* which could be strengthened by corrective feedback.

Two subsequent studies using the same weak-associate paradigm (Grimaldi & Karpicke, 2012; Huelser & Metcalfe, 2012) came to the conclusion that a benefit of generating errors could only be observed when there was such a pre-existing semantic association between cue and target and not when the cue was unrelated to the target. Grimaldi and Karpicke proposed that this was because, for related pairs, participants not only retrieved, at study, the associate they gave as their guess (e.g., *water*) but also covertly retrieved other associates, including the target (*frog*). Retrieval of the target facilitated its encoding when it was presented as feedback. In other words this was a classic testing effect reinforced by feedback: the relevant cue-target association already existed in the participant's memory and

was retrieved, along with other associations to the same cue, when the participant was prompted with the cue and asked to guess the target. The corrective feedback simply confirmed that this was the cue-target pair required, rather than any of the others retrieved at the same time. Since retrieval strengthens memory more than reading, and feedback enhances the benefits of testing (Butler & Roediger, 2008; Pashler, Cepeda, Wixted, & Rohrer, 2005), this led to an advantage for targets studied in the Generate condition.

Huelser and Metcalfe (2012) similarly proposed that pre-existing semantic relatedness between cue and target was essential for the benefit of generating errors to be observed. They proposed that the benefit occurred because the error generated by the participant would have a pre-existing semantic association with the cue. In a case where the cue and target were related, the error was also likely to be related to the target and could therefore enhance memory by functioning either as an elaborator or as a mediator. The error could function as an elaborator because retrieval of a word that was semantically related to the cue could lead to activation of other concepts associated with the cue which would also be associated with the target, thus creating a more elaborate memory trace, and providing more information which could be used as retrieval cues for the target at final test. Alternatively it could function as a mediator by acting as a link between the cue and the target which could benefit memory as long as participants were able to remember their own incorrect guess and use it to link to the correct target (see Pyc & Rawson, 2010, for an account of the mediator effectiveness hypothesis).

Hays et al. (2013) offered an account similar to the elaboration hypothesis described above, proposing that generating a response primed knowledge related to the cue, activating a network of semantically related information, which facilitated the mapping of the cue to the target. This of course would only apply when the cue was also related to the target. Likewise, Kornell et al. (2009) proposed three “retrieval based” explanations of the effect. These

explanations all assume that material activated during an incorrect guess undergoes memory enhancement and this can benefit the encoding of corrective feedback if the guess is also related to the target.

The prevailing view, then, is that errors can only be beneficial when there is a pre-existing semantic relationship between cue and target. In all of these studies, however, the pre-existing relatedness between cue and target is confounded with the fact that the cues all have strong pre-existing associations, so there is also a pre-existing relationship between the cue and the generated error. Since retrieval confers a direct benefit on the retrieved item, the generation of an error related to the cue may be helpful to memory when it is also related to the target (as in the case of related word pairs), for the reasons proposed above, but it may be detrimental to memory when the error is unrelated to the target (as in the case of unrelated word pairs) since it may interfere, at test, with retrieval of the correct answer. In this case there may still be an indirect benefit of generation to the subsequent feedback (e.g., by causing more attention to be focused on the feedback) but this may be obscured by interference from the error related to the cue which has been strengthened by generation or retrieval. In order to identify whether a failed test can benefit memory purely by potentiating encoding of feedback, it is necessary to examine a situation in which errors are unrelated to either cue or target.

Furthermore, the weak-associate paradigm used in these three studies is rather unlike any real life testing situation, where there is typically only one valid answer to the question that has been set and the task is to recall that answer, not to guess which of many valid answers the experimenter or instructor happens to have in mind. For example, one would not normally expect to be asked “Name one of Marilyn Monroe’s husbands” and then, having given the (valid) answer “Joe DiMaggio”, be told that that was incorrect and the answer required was “Arthur Miller”. Instead, a more realistic question would be, “Name Marilyn

Monroe's third husband". In this instance, retrieving "Joe DiMaggio" would indeed be incorrect, not just for the purposes of this test but always. The weak-associate paradigm is used, however, because of its potential to generate many "errors" – or, at least, responses at study which will be different from those required at test, though it is not clear whether it does in fact involve the same processes as are involved in making and correcting genuine retrieval errors.

Errorful generation

Our focus in the current study is on the rather different scenario, errorful generation, in which students make incorrect guesses because the test material is completely new to them. The learning of novel vocabulary represents a rather more realistic learning scenario in which there is a one-to-one relationship between the cue and the target. Thus, the rare but real English word "menald" means "spotty" and will always mean "spotty"; it does not also mean "brainy", "helpful" or "drowsy", either for the purposes of this experiment or at any time in the future. If a participant says it means "helpful" they are making a genuine error, not simply failing to guess what was in the experimenter's mind at the time. Neither the cue nor the task instructions constrain guessing, so the incorrectly generated item is unlikely to be related to either cue or target. Unlike the Slamecka and Fevreiski design the cue-target pair is not already known and so has no potential to be even partially retrieved. There is no pre-existing association to be reinforced and participants have to learn a completely novel association. Kornell et al. (2009) came closer to this scenario in their first two experiments by using fictional trivia questions, to which participants could not possibly know the answer. They interspersed fictional questions with real ones in order to encourage participants to attempt retrieval, even though there was no memory to be retrieved, and the fictional questions were all based on real ones so participants might have retrieved details related to the real

counterparts, but Kornell et al. found no advantage of generating over reading when total trial time was equated. However, participants tended to produce no answer rather than an incorrect one, so it was not possible to test the hypothesis that producing incorrect answers impairs subsequent memory for correct ones (Baddeley & Wilson, 1994; Roediger & Marsh, 2005).

Grimaldi and Karpicke (2012) argued that the reason they did not observe a generate advantage for unrelated word pairs (pairs with no pre-existing association), e.g., *pillow-leaf,* in their paradigm was because, for a generate advantage to occur, the target had to be retrieved as part of the "search set" along with the incorrect guess. However, this conclusion may be premature. In both Grimaldi and Karpicke's (2012) and Huelser and Metcalfe's (2012) studies, participants generated guesses which were highly related to the cues in both the related and unrelated conditions (e.g., *sleep* for the cue *pillow*). At final test, presentation of the cue (*pillow*) is likely to have brought to mind the same (related) response which was given at study (*sleep*), along with all the other related responses activated at the same time (e.g., *bed, head, feather, cushion* etc). Howard and Kahana (2002) showed that participants tend to recall items which are semantically related to the item just recalled. This would be helpful in the related case, where the target is in fact related to the participant's incorrect guess, but unhelpful in the unrelated case, where retrieval at test of all the related associates of the cue would be likely to interfere with the participant's ability to remember the unrelated target *leaf*.

Moreover, when related and unrelated items appear in a mixed list at study, as in Grimaldi and Karpicke's (2012) study, participants may not remember, at test, which cues were matched with related targets and which with unrelated ones, so might search for the correct answer exclusively among the related associates. Therefore, Grimaldi and Karpicke's proposal that errorful generation is only beneficial for related items (because they are covertly retrieved at study and thus benefit from a classic testing effect) and not for unrelated

items (because they are not retrieved at study), may not be the only possible explanation for their findings. Errorful generation may benefit subsequent encoding even when the target is not retrieved at study and is not therefore among the "search set" and the failure to observe a generation benefit in the unrelated case may be simply because the benefits of generation were outweighed by interference at test from a strong but incorrect associate to the cue. This would also apply to the explanations offered by Huelser and Metcalfe (2012) and Hays et al. (2013).

Indeed, in Huelser and Metcalfe's (2012) study it was significantly more common for an error made at study to be repeated at test in the unrelated than the related condition, whether the design was between subjects (Experiment 1) or within subjects (Experiment 2). Hays, Kornell, and Bjork (2013) also found evidence of interference from incorrect answers generated at study. Furthermore, it is difficult to explain Kane and Anderson's (1978) findings by Grimaldi and Karpicke's account, since in their undetermined condition the targets were unrelated to the most obvious, incorrect, completions so it is highly unlikely that they would have been retrieved along with them. In the current study we ask whether it is, in fact, possible to observe a benefit of incorrect guessing even when the cue is unrelated to the target, in a scenario in which there are no pre-existing associations between cue and target because the materials, novel vocabulary items, have never previously been encountered. The cues in our study are very obscure English words or foreign language words which do not already exist in participants' mental lexicons and which therefore have no pre-existing associations to retrieve. Can making an incorrect guess in response to cues which have never been seen before lead to more effective encoding of subsequent feedback than passively studying correct answers?

There are reasons why errorful generation could be helpful even when the incorrect response is unrelated to the correct one. Making an incorrect guess in response to a test

question may arouse curiosity about the correct answer, leading to more attention being paid to that answer. Berlyne and Normore (1972) found that inducing, then satisfying, curiosity by presenting a blurred picture immediately before a clear picture of the same object, led to better memory for the objects than presenting the clear picture for twice as long. M. J. Kang, Hsu, Krajbich, Loewenstein, McClure, Wang, and Camerer (2009) found that the higher participants rated their curiosity about a question answered incorrectly, the more likely they were to recall the correct answer on a surprise test two weeks later. There is also evidence that greater attention is given to feedback that does not match expectations (e.g., Butterfield & Metcalfe, 2001; Fazio & Marsh, 2009), and much research shows that discrepancies between what is expected and what occurs drive learning (e.g., Rescorla & Wagner, 1972). It is possible, then, that the discrepancy between a generated error and subsequent corrective feedback may capture attention, enhancing encoding of the correct answer. Some support for this notion comes from studies in which students attempt to answer questions before studying a text, leading to better memory for the material than simply reading the questions (e.g., Pressley, Tanenbaum, McDaniel, & Wood, 1990; Richland, Kornell & Kao, 2009).

A handful of previous studies have examined guessing during vocabulary learning but have produced mixed results. Forlano and Hoffman (1937) found that "telling" was better than "guessing" when schoolgirls learned Hebrew-English word pairs. Although total learning time was equated for the two conditions, it is not clear how time was allocated for each item. Berlyne, Carey, Lazare, Parlow, and Tiberius (1968) obtained similar findings for intentional learning of Turkish-English word pairs in adults. However, they did find an advantage of generating under incidental learning conditions, again proposing that memory was reinforced by the satisfaction of curiosity.

In our first experiment participants learned definitions of rare English words (e.g., *roke – mist*) either by reading the word with its definition, by guessing a definition followed

by corrective feedback, or by choosing from two definitions followed by feedback. These three conditions are representative of three common methods of classroom learning: studying by reading, cued recall tests, and multiple choice questions (MCQ). Because we were interested in determining which study method makes optimal use of study time, total trial time was equated for each of the three conditions. Participants then took a final multiple choice test of all the words, again reflecting a typical educational testing scenario. The greater sensitivity of a multiple choice test may reveal differences between conditions which might be harder to detect with a recall test, and it also permits us to examine the effects of different choice alternatives (lures) at test. We were interested in whether the active process of generating a definition for an unknown word, even though it would nearly always produce an error, would lead to better memory for correct definitions than either passive studying or choosing, despite the fact that the cue provided no constraints on guessing and no opportunity for meaningful elaboration. Because we were interested in the effect of making errors, we firmly encouraged participants to guess in the Generate and Choice conditions.

We included the two-alternative Choice condition in order to investigate the effect of giving an error response at study without the component of generation. Participants were shown the cue and two possible definitions, one of which was correct and one a lure, and were instructed to type in the one they thought was correct. When participants are asked to make a choice, no act of generation is required since the correct answer is presented intact. When material has been previously studied, taking an MCQ test followed by feedback has been found to yield comparable final test performance to restudying (e.g. Butler & Roediger, 2007; S. H. K. Kang, McDermott, & Roediger, 2007; McDaniel, Anderson, Derbish, & Morisette, 2007). With repeated testing, superior performance has even been observed (McDaniel, Wildman, & Anderson, 2012). However, in these studies, performance in the initial MCQ test was very high. In the present experiment, with no prior study, we expected

the words to be unfamiliar to participants and that correct guesses at study in the Choice condition would be no higher than chance. Under these conditions we predicted that incorrect choices selected at study would interfere with correct memory at final test such that this condition would produce poorer performance than the Read condition, which involved no interference, and poorer performance than the Generate condition which we expected to benefit from the active process of producing an answer, albeit an incorrect one.

Participants' study decisions, such as how much effort to apply to studying a given item, are likely to be influenced by their perception of how difficult that item will be to remember. In Experiments 2A, 2B and 3 we therefore had participants make a judgment of learning (JOL) after studying each item, predicting their likelihood of remembering it later. People typically believe that studying is more effective than testing for previously studied material, even though the converse is true (e.g., Roediger & Karpicke, 2006b). For unstudied items, generating correct responses has often, though not invariably, been shown to elicit higher JOLs than reading, suggesting that participants are aware of the benefits of generation (e.g., Begg, Vinski, Frankovich, & Holgate, 1991). However, it has also been found that ease of processing, or encoding fluency, influences JOLs (e.g. Castel, McCabe, & Roediger, 2007; Koriat, 2008; Hertzog, Dunlosky, Robinson & Kidder, 2003; Schwartz, Benjamin & Bjork, 1997). We requested JOLs immediately after the learning of each correct definition in order to capture participants' perception of their learning at the very moment they finished studying the item. We wanted to examine whether participants' perception of their learning of correct definitions would be influenced by whether or not that learning had been preceded by the making of an error.

If participants perceive Generate items as more difficult to learn than Read items, because they have generated an error, this might lead them to apply more attention to processing the correct feedback which in turn may lead to better memory for the item. We

predicted that generating errors before encoding correct information would lead to corrective feedback being processed less fluently, and therefore to lower JOLs for Generate than Read items, but that memory would be superior for Generate items. When an error is made in the Generate condition, processing of the feedback may be less fluent because participants have to disengage their attention from the incorrect response they generated, and from any semantically-related concepts activated at the same time, and switch it to the encoding of corrective feedback which may be in an entirely different semantic space. In contrast, for items in the Read condition there is no requirement to switch from processing one definition to another.

We also captured aggregate JOLs at the end of the study phase: We asked participants to estimate, for each of the three study methods, what proportion of definitions they believed they would remember when they took the final test. We were interested in whether these would yield a similar pattern to the item JOLs. In Experiment 2A we again used obscure English words and in Experiment 2B the stimuli were foreign language words from Euskara, the language of the Basque country in Northern Spain, which we chose because it is a "language isolate", a language with no known relations. Foreign language learning is an important real world skill which forms a compulsory part of the high school curriculum in nearly all European countries.

In our first two experiments, test lures were either new items or lures which had been presented at study in the Choice condition. Thus participants were able to select, at test, incorrect definitions they had selected at study in the Choice condition but were not able to select incorrect definitions they had produced at study in the Generate condition. This enabled us to examine separately the effects of interference (by comparing the Choice and Read conditions) and of generation (by comparing the Generate and Read conditions). In Experiment 3 we investigated whether any benefit of generation would be eliminated if

participants had the option to select, at test, the incorrect answer they had themselves generated at study. We also examined the effect, on both memory and JOLs, of allowing participants to choose how much time they spent studying correct answers.

To preview our main conclusions, we observed a benefit of generating errors over reading in all three experiments and a benefit of generating errors over incorrect choosing in Experiments 2A, 2B, and 3. We also found that participants, in their metacognitive judgments of learning, consistently failed to predict this benefit. Indeed they erroneously judged errorful generation to be a less effective encoding method than reading or incorrect choosing.

Experiment 1

Method

Participants. Twenty-four participants, 12 male, average age 28.7 (*SD* = 10.7), were recruited from the University College London (UCL) participant pool which comprises both students and non students. They participated in return for a small payment (£4).

Design. In this and the subsequent experiments reported here, we used a within-subjects design with one independent variable (Study Method) with three levels (Read, Generate, and Choice). The dependent variable was the number of items recalled at final test in each condition.

Materials. For the stimulus materials we created a pool of very unusual English words, each paired with a one-word definition e.g., *hispid - bristly, valinch - tube, frampold – quarrelsome,* from which we selected 60 pairs which were unfamiliar to participants in a pilot study. For the choice condition, a lure was created for each of the English targets. For the final multiple choice test a further two lures were created for each of the words. The set of 60 items was divided into three subsets of 20 items each for counterbalancing purposes. Each subset was matched for average number of letters and syllables per word, and each subset

contained the same number of nouns, verbs and adjectives. Computer software written in Visual Basic 6.0 presented and controlled the experimental task.

Procedure. Participants were told that they would study words presented in three different formats and that they should try to remember the correct definitions for a later memory test. The study phase was preceded by a practice phase to familiarise participants with the task. At study, each word was presented on the computer screen once and one at a time in one of three randomly interleaved formats. Each item appeared equally often in each condition across participants. A Read trial consisted of the cue (the English word) and the target (a one-word definition) being displayed on the screen for 17 s. A Generate trial consisted of an English word being displayed for 10 s while the participant was prompted to type in a one-word definition, followed by presentation of the correct answer for 7 s. During a Choice trial an English word was displayed for 10 s with two possible choices, the true one-word definition and a lure, during which time the participant was prompted to type in the definition they thought was correct. Then the correct answer was displayed for 7 s. Participants were told that if they did not know the word they should guess. Figure 1 depicts the procedure and timings used in all four experiments.

Following the study phase, participants were given 1 min to solve some arithmetic puzzles. The final phase of the experiment was a multiple choice test. All 60 English words were presented, one at a time in random order, with four possible alternatives which included the correct definition and the lure created for the initial choice test, plus two additional lures. The relative position of each alternative on the screen was randomly determined on a trial by trial basis. For each word, participants were prompted to select the correct definition from amongst the four alternatives and type it in. No feedback was given.

Results

At study correct generations were few (M = 6.5%, SD = 12.3) and correct choices were at chance (M = 53.3%, SD = 10.5, $t(23) = 1.6$, $p = .133$), confirming that most definitions were unknown to participants pre-experimentally[1].

Final test performance differed by study method, $F(2,46) = 7.62$, $p = .001$, $\eta_p^2 = .25$, (Figure 2A). When all items were considered, whether the associated response was correct or incorrect at study, Generate items were better remembered than Read ($t(23) = 3.65$, $p = .001$, $d = .47$) and Choice items ($t(23) = 2.80$, $p = .010$, $d = .42$). Fifteen participants remembered more Generate than Read items and only three showed the reverse pattern. The Read and Choice conditions did not differ ($t(23) = .77$, $p = .447$, $d = .09$). Since we were particularly interested in the effect of making errors at study, we analysed final test performance for just those items answered incorrectly at study, which entailed dropping a small number of items from the analysis in the Generate condition and about half the items in the Choice condition (Figure 2A). This analysis revealed a similar pattern, $F(2,46) = 3.33$, $p = .044$, $\eta_p^2 = .13$. Generate scores were higher than Read scores, $t(23) = 2.40$, $p = .025$, $d = .56$, but the difference between Generate and Choice fell short of significance, ($t(23) = 1.91$, $p = .069$, $d = .36$), possibly because there were too few incorrect items to reveal the effect. Again there was no significant difference between the Read and Choice conditions ($t(23) = .01$, $p = .990$, $d = .002$). The benefit of generating over reading is particularly striking because generating nearly always produced an error, and the correct definition was available for much less time - just 7 s, compared with 17 s for Read trials.

Contrary to our expectations, choosing did not lead to poorer performance than reading, even when the analysis was confined to just those items which were incorrect at study. We examined the type of errors made at final test in the Choice condition. When an incorrect response was made at final test, this response was significantly more likely to be the original lure when that lure had also been picked at study (M = 73.4% of incorrect responses

following initial selection of a lure, *SD* = 30.2) than when a correct answer had been given at study (*M* = 35.3% of incorrect responses following an initially correct choice, *SD*= 33.0), $t(15) = 3.44$, $p = .004$. Indeed, when the initial response was correct but the final test response was incorrect, participants picked the original lure from the study phase at a rate no different from the chance rate of 33.3%, $t(17) = .78$, $p = .447$. (Note that 6 participants made no errors at final test following selection of the correct response at study, so they had no data to contribute to this analysis.) Thus, even though the original lure had been seen in the study phase and the other two options had not, participants' incorrect responses were not affected by any additional familiarity associated with the original lure. However, when both the initial and final responses were incorrect, participants picked the same incorrect answer at test as they had done at study at a rate considerably higher than chance, $t(15) = 5.31$, $p < .001$, suggesting that errors made on the initial test can interfere with accurate retrieval at final test. (Note that 8 participants made no incorrect responses at final test following selection of the lure at study, so they had no data to contribute to this analysis.)

Thus, although there was no overall detriment to the Choice condition by comparison with the Read condition, there was some evidence that errors made at study interfered with final test performance. However, any negative effect of interference seems to have been offset by a positive effect of selecting a definition from a choice of two, perhaps because this involved deeper processing than passively reading the word and its definition.

Discussion

Our first experiment revealed a benefit of generating followed by feedback over reading during the learning of unusual English words, even though generation produced many errors at study. Generation was also more beneficial than choosing when all items were considered and there was a marginal benefit when only items incorrect at study were

considered. Our hypothesis that incorrect choosing might lead to poorer final test performance than reading was not supported, though there was some evidence that lures selected at study interfered with selection of the correct answer at test.

Experiments 2A and 2B

In our second experiment we aimed to replicate the benefit of errorful generation over reading for the learning of unusual English words (Experiment 2A) and to examine whether the effect extended to the learning of foreign language vocabulary (Experiment 2B). We also asked whether participants had insight into this benefit by having them give a metacognitive judgment of learning (JOL) after learning each item. Conditions which make learning more effortful often lead to better memory for the learned items (Bjork, 1994). The difficulty experienced during learning, however, may lead people to underestimate this benefit. If generating errors leads to Generate items being perceived as more difficult to learn, participants may apply more effort or attention to encoding corrective feedback for these items, and this could lead to superior memory for Generate items. We therefore predicted that participants would give lower JOLs to Generate items but that final test performance would show the opposite pattern.

Method

Participants. In Experiment 2A there were 30 participants, 12 male, average age 23.9 (*SD* = 5.2). In Experiment 2B there were 24 participants, five male, average age 26.0 (*SD* = 11.4), none of whom reported any prior knowledge of Euskara, the language of the Basque region of Spain.

Materials. In Experiment 2A we used 60 word-definition pairs taken from the same pool of items as in Experiment 1, replacing words for which the definitions had been

correctly generated in Experiment 1. For Experiment 2B we selected 60 Euskara nouns with their English translations (e.g., *igel - frog*, *urmael - pond*, *untxi - rabbit*). In both experiments we created, for each item, three lures derived from the MRC Psycholinguistic Database (Coltheart, 1981) or the English Lexicon Project at http://elexicon.wustl.edu (Balota, Yap, Cortese, Hutchison, Kessler, Loftis, Neely, Nelson, Simpson, & Treiman, 2007). Each lure was matched with the true definition or translation for number of syllables and for approximate word frequency (Kucera & Francis, 1967). These appeared as lures for the Choice condition at study, and for all items at final test. Therefore, for items in the Choice condition, the options presented at test were the same as the options presented at study (i.e., the target and the same three lures). Counterbalancing was as in Experiment 1.

Procedure. The procedure was identical to that of Experiment 1 with the following exceptions. Study time was reduced to 13 s per trial (with 8 s for entry of responses and 5 s for studying of feedback) in order to keep the task to a reasonable length given that participants were also entering JOLs. Four choices were presented at study in the Choice condition instead of two, in order to increase the proportion of items which would be incorrect at study, thereby enabling us to examine the effect of errors more comprehensively. After each trial participants predicted their later likelihood of remembering the item by entering an item JOL, a number from 0 ("No chance I'll remember it") to 100 ("I'll definitely remember it").

Following the study phase, participants gave three aggregate JOLs, predicting the percentage of items they expected to remember from each study method. Entry of item and aggregate JOLs was self-paced. Response time data for making JOLs are given in Appendix A. The procedure was identical for Experiments 2A and 2B except that in Experiment 2A participants were not explicitly told what format the final test would be in, whereas in Experiment 2B they were told to expect a multiple choice test.

Results

Experiment 2A (English words).

At study only 0.3% of Generate responses were correct. Correct responses to Choice items (M = 30.3%, SD = 10.6) were above chance[2] ($t(29) = 2.76$, $p = .010$).

Final test performance: Experiment 2A (English words).

Replicating the findings of Experiment 1, when all items were considered, final test performance differed between study methods, $F(2,58) = 9.85$, $p < .001$, $\eta_p^2 = .25$ (Figure 2B). Generating with feedback was superior to reading ($t(29) = 4.27$, $p < .001$, $d = .40$) and to choosing with feedback ($t(29) = 3.62$, $p = .001$, $d = .33$), while the Read and Choice conditions did not differ ($t(29) = .54$, $p = .596$, $d = .05$). Nineteen participants remembered more Generate than Read items, while only two showed the opposite pattern. For items which were incorrect at study, the difference between study methods remained, $F(2,58) = 10.01$, $p < .001$, $\eta_p^2 = .26$, as did the advantage for Generate over Read items, $t(29) = 4.28$, $p < .001$, $d = .39$. Whereas in Experiment 1 the advantage of generation over choosing fell just short of significance for items incorrect at study, in Experiment 2A there was a clear benefit of generating over choosing incorrect definitions, $t(29) = 3.91$, $p = .001$, $d = .46$. There was no difference between reading and incorrect choosing, $t(29) = .90$, $p = .373$, $d = .10$.

Was there any evidence that making an error at study in the Choice condition interfered with selection of the correct answer at test? When an incorrect choice had been selected at study, and the final test response was also wrong, the same response was selected at test at a rate numerically but not significantly higher than chance (33%, because there are 3 incorrect lures at test), M = 44.0, SD = 37.8, $t(23) = 1.38$, $p = .180$ (not all participants had data to contribute to this analysis.) In Experiment 1, where test lures for Choice items

consisted of the lure which had been present at study and two new lures, participants were much more likely to persist with an incorrect choice than to pick a new lure. By contrast, in Experiment 2A all options at test for Choice items had previously been seen as study lures. In this situation, selecting and typing in an incorrect response at study did not make participants significantly more likely to persist with their own incorrect choice than to select one of the other lures. Put differently, items incorrectly chosen at study were strong enough to lead to perseverative errors at test when the alternatives were new lures (and the correct target), but not strong enough to lead to such errors when the alternative test items were familiar lures (and the correct target).

Experiment 2B (Euskara words).

At study only one response given in the Generate condition was correct across all participants (M = .2%, SD = 1.0). Correct responses to Choice items (M = 31.3%, SD = 14.5) were again above chance[2] ($t(23) = 2.21$, $p = .046$).

Final test performance: Experiment 2B (Euskara words).

Just as with the English version, final test performance differed between study methods, $F(2,46) = 6.05$, $p = .005$, $\eta_p^2 = .21$ (Figure 2C). Generating produced better final test performance than reading, $t(23) = 3.36$, $p = .003$, $d = .28$. Fourteen participants remembered more Generate than Read items, while only 3 showed the opposite pattern. The difference between generating and choosing was close to significant, $t(23) = 1.84$, $p = .079$, $d = .18$. There was no difference between reading and choosing, $t(23) = 1.68$, $p = .106$, $d = .12$. The analysis of most interest, of just those items incorrect at study, revealed an identical pattern of results to the English version of the task in Experiment 2A. There was a difference between study methods, $F(2,46) = 6.81$, $p = .003$, $\eta_p^2 = .23$, and an advantage for generating

errors over reading, $t(23) = 3.36$, $p = .003$, $d = .28$. The advantage of generating errors over choosing incorrectly was also significant in this analysis, $t(23) = 2.77$, $p = .011$, $d = .30$, with no difference between reading and incorrect choosing, $t(23) = .14$, $p = .892$, $d = .011$.

Did making an error at study in the Choice condition interfere with selection of the correct answer at test? Here the results differed from those of Experiment 2A.When an incorrect response was given at study for an item in the Choice condition, and the final test response was also wrong, participants selected the same incorrect response at test at a rate significantly higher than the chance level of 33% ($M = 59.7$, $SD = 41.9$), $t(16) = 2.59$, $p = .020$ (again, not all participants had data to contribute to this analysis.) Just as in Experiment 2A, test lures were the same as study lures in this version of the task but, whereas in Experiment 2A selecting an incorrect answer at study did not make it reliably more likely to be picked at test than any of the other lures also seen at study, in Experiment 2B, when participants made an error, they tended to persist with the same error they had made at study rather than select one of the other lures.

Experiments 2A and 2B therefore replicated the benefit of errorful generation over reading observed in Experiment 1 and also revealed a benefit of errorful generation over incorrect choosing. As in Experiment 1, there was some evidence (in Experiment 2B) that lures selected at study can interfere with selection of the correct answer at test.

Judgments of learning: Experiment 2A (English)

Were participants aware of the benefit of errorful generation during learning? For the English version of the task, item JOLs differed for the three study conditions, $F(2,58) = 20.73$, $p < .001$ (Figure 3A), $\eta_p^2 = .42$. Choice JOLs were higher than both Read ($t(29) = 3.71$, $p = .001$, $d = .27$) and Generate ($t(29) = 6.37$, $p < .001$, $d = .43$) JOLs, and Read JOLs were higher than Generate JOLs ($t(29) = 2.59$, $p = .015$, $d = .17$). Participants' JOLs, then,

were strikingly inaccurate: the Generate condition produced the highest recall scores but the lowest JOLs.

Aggregate JOLs showed a largely similar pattern (Figure 3A). The assumption of sphericity was not met, $\chi^2(2) = 9.63$, $p = .008$, so the Huynh-Feldt correction was applied. JOLs differed by study method ($F(1.62, 46.99) = 7.78$, $p = .002$, $\eta_p^2 = .21$). Choice JOLs were again higher than both Read ($t(29) = 2.58$, $p = .015$, $d = .49$), and Generate JOLs ($t(29) = 3.49$, $p = .002$, $d = .58$) but there was no difference between Read and Generate ($t(29) = .85$, $p = .400$, $d = .10$).

Why were predictions so inaccurate? We examined JOLs made in the Choice condition, the only condition in which participants regularly made both correct and incorrect responses at study, in relation to the accuracy of their responses at study. This revealed three interesting findings. Firstly, Choice JOLs for definitions guessed correctly at study were very substantially higher than for items incorrect at study (Fig. 3A), $t(29) = 7.02$, $p < .001$, $d = 1.11$. Secondly, JOLs for Choice items correct at study were higher than Read JOLs, $t(29) = 6.62$, $p < .001$, and thirdly, JOLs for Choice items incorrect at study were indistinguishable from JOLs for Generate items incorrect at study ($M = 31.3$, $SD = 16.7$), $t(29) = .77$, $p = .450$, and from JOLs for Read items, $t(29) = 1.74$, $p = .093$. These findings suggest that Choice JOLs were largely driven by the fortuitous making of a correct choice at study.

Together with the higher JOLs for Read than Generate items, this suggest that one factor influencing JOLs was fluency of processing at study, which was itself influenced by the outcome of the preceding event. Generating errors (which happens on almost every trial) leads to less fluency and lower JOLs, while reading leads to intermediate fluency and JOLs. Making an incorrect choice also leads to low fluency and JOLs, but correct choice – despite being fortuitous – leads to much greater fluency and JOLs. When an error is made, in either the Generate or Choice conditions, participants have to switch their attention from their own

incorrect response, with all its associations, to the correct response presented as feedback. This is not necessary in the errorless Read condition and, similarly, where the correct selection is made in the Choice condition, processing of this correct answer can continue uninterrupted and unaffected by interference from a previously chosen or generated error. (But see Appendix B for an alternative possibility.)

Appendix C reports the relationship between JOLs and test performance. Although JOLs showed some ability to predict final test scores (Figure 4A), these data should be interpreted with caution since they may be affected by item selection effects.

Judgments of learning: Experiment 2B (Euskara)

For the foreign language version of the task, item JOLs also differed for the three study conditions, $F(2,46) = 12.60$, $p < .001$, $\eta_p^2 = .35$ (Figure 3B). Once again, participants gave lower JOLs to Generate items than to either Read ($t(23) = 3.69$, $p = .001$, $d = .39$) or Choice ($t(23) = 4.31$, $p < .001$, $d = .50$) items, but here there was no difference between JOLs for Read and Choice items, $t(23) = 1.38$, $p = .181$, $d = .12$.

Aggregate JOLs followed a similar pattern to the item JOLs (Figure 3B). There was a main effect of study method, $F(2,46) = 5.09$, $p = .010$, $\eta_p^2 = .18$. Replicating the findings of the English version of the task, Choice JOLs were higher than Generate JOLs, $t(23) = 3.21$, $p = .004$, $d = .67$, and there was no difference between Read and Choice JOLs, $t(23) = .28$, $p = .783$, $d = .07$. This time Read JOLs were also higher than Generate JOLs, $t(23) = 2.73$, $p = .012$, $d = .57$.

Just as for the English version of the task, generating produced the highest final test scores but the lowest JOLs. Again, inspection of Choice JOLs in relation to study performance is illuminating and reveals the same pattern of results as in Experiment 2A. First, JOLs for Choice items guessed correctly at study ($M = 47.4$, $SD = 21.8$) were much

higher than for items guessed incorrectly (M = 31.6, SD = 15.7), $t(23) = 5.52$, $p < .001$, d = .83 (Figure 3B). Second, JOLs for Choice items guessed correctly at study were significantly higher than JOLs for Read items, $t(23) = 4.36$, $p < .001$. Finally, JOLs for Choice items guessed incorrectly at study were significantly lower than for Read items, $t(23) = 2.38$, p = .026, though with the exclusion of two participants who performed above chance at study (see footnote 2), this difference was no longer significant, $t(21) = 1.98$, p = .061 (M = 31.7, SD = 14.8 for Read, M = 29.4, SD = 14.3 for Choice incorrect at study). JOLs for Choice items incorrect at study were higher than JOLs for Generate items, $t(23) = 2.22$, p = .037, but again this difference disappeared ($t(21) = 1.89$, p = .072) (M = 26.5, SD = 14.7 for Generate) when the two participants were excluded (see footnote 2), yielding the same pattern as in Experiment 2A. These results again suggest that participants were strongly influenced by their success or failure at study, and particularly by the fortuitous selection of a correct choice. In Experiment 2B there was no relationship between JOLs and test accuracy (Fig 4B). See Appendix C for these data.

Discussion

Our second experiment replicated the benefit of generating over reading for unusual English words that we observed in Experiment 1 and showed that it extends to the learning of foreign vocabulary. Grimaldi and Karpicke (2012) have proposed that, when guessing produces an error, it will only benefit later memory if the correct answer is, in fact, already known and activated at the time of the guess. Participants in Experiment 2B had no prior knowledge of Euskara, yet they showed better final test performance for items for which they had generated an incorrect guess than for items they had studied in the Read condition. In Experiment 2A the stimuli were obscure English words which were likely to be largely unknown to participants pre-experimentally, yet they showed the same benefit of generating

over reading. These findings are also inconsistent with the other versions of the semantic relatedness hypothesis that we discussed in the introduction (e.g., Huelser & Metcalfe, Hays et al., 2013), and show that a pre-existing relationship between cue and target is not necessary for the errorful generation benefit to be observed.

Perhaps our participants in Exp 2A were all avid crossword-solvers with a passion for Scottish dialect and archaic English and did have some knowledge of the definitions prior to the experiment but, in line with Slamecka and Fevreiski's (1983) proposal, were unable to produce them in the time available? To address this question we investigated to what extent participants' responses were related to the correct answers. We obtained ratings of the similarity between participants' generated words and the correct definitions by using the latent semantic analysis (LSA) tools available at http://lsa.colorado.edu. LSA extracts and represents the similarity in meaning of words by means of statistical computations applied to large bodies of text (Landauer, Foltz, & Laham, 1998). Values close to 1 indicate that items are highly related, while values close to 0 mean they are highly unrelated. (We could not of course compute the similarity of the generated words to the cue words, since the cue words were too obscure to be represented in the corpora used for the LSA.) The mean similarity of the generated items to the correct definitions was .096 (*SD* = .047). As an example of a randomly unrelated set of items, we compared this with the average similarity of the targets to all of the other targets (*M* =.105, *SD* = .045), $t(59) = 1.81$, $p = .076$. Participants' guesses were no more related to the correct definitions than the correct definitions were to each other.

In addition to the benefit of generating over reading, Experiment 2 also revealed a clear benefit of generating errors over incorrect choosing, both for rare English words and for foreign language words. Participants' JOLs, however, showed a very different pattern from their actual test performance, with the lowest item JOLs being given to Generate items and the highest to Choice items. These high average JOLs given to Choice items were largely due

to much higher JOLs for Choice items guessed correctly at study, while JOLs for items incorrect at study were no higher than JOLs for Read items, a pattern which is suggestive of participants using their own performance at study as a basis for predicting their future memory performance. Participants' aggregate JOLs were also highest in the Choice condition, though memory performance showed no advantage of choosing over either reading or generating. This suggests that participants' correct answers in the Choice condition, which were likely to occur around 25% of the time simply by chance, gave them a sense of having been successful with this mode of learning and led them to give a higher aggregate JOL to this condition than to the Generate condition, where they did not experience such success, consistent with previous research showing that ease of processing at encoding can lead to relative overestimation of future memory performance (e.g., Castel et al., 2007).

The JOLs data may shed some light on the reason for the benefit of generating over reading observed here. The higher JOLs given to Read than Generate items in both these experiments are consistent with our hypothesis that Generate items are experienced as more difficult than Read items. This may lead to greater attention being paid to the corrective feedback for Generate items. The Read condition may even give rise to an illusion of knowing, or "knew it all along" effect (Fischhoff, 1977), making it difficult for the participant to imagine producing an error when tested later. The illusion is exaggerated in the Choice condition when the participant happens to select the correct response. Our proposal is that higher JOLs in the Read and Choice conditions reflect an illusion of knowing which leads to less effort being applied to encoding, or to a less efficient encoding strategy, than for Generate items, where initial responses are always incorrect. Generating errors therefore leads both to lower JOLs and to higher final test performance. This is consistent with other research showing that greater effort during study leads both to better recall and to lower JOLs (Zaromb, Karpicke, & Roediger, 2010).

What we can conclude from both these experiments is that participants' JOLs seemed to reflect their subjective experience at study. For Generate items, participants predicted poorest test performance while achieving highest performance. If metacognitive beliefs influence study decisions, such as the decision as to how much effort needs to be applied to encoding a given item, participants may make more effort to learn definitions for items subjectively experienced as more difficult and this could be one way that even errorful generation potentiates subsequent encoding.

Experiment 3

In our first three experiments, we separated the effects of generation and interference by giving participants the opportunity to select, at test, Choice lures they had selected at study but not responses they had generated themselves in the Generate condition. Experiment 3 was designed to examine whether the Generate over Read advantage observed in our first three experiments would persist when participants had the opportunity to select, at test, their own incorrect generated responses as well as their own incorrect Choice responses. We had hypothesised that incorrect choosing might lead to poorer performance than reading, because the Choice condition involves potential interference from errors without any benefit from the act of generation, but in fact our experiments revealed no detriment of incorrect choosing by comparison with reading. Memory for Choice items incorrect at study was, however, poorer than for Generate items. If participants had the opportunity to select the same erroneous response they had made at study in the Generate condition, just as they could in the Choice condition, this might eliminate the advantage of generating over both reading and incorrect choosing. Furthermore, in our first three experiments test options for Choice items included familiar lures from the study phase whereas test options for the other conditions had not been seen at study. This allowed us to examine the effect of interference from study lures, but it

may have disadvantaged the Choice condition relative to the other two. In Experiment 3 we equated familiarity of the lures between the three conditions.

A further aim of this experiment was to examine whether we would observe the same pattern of results, in terms of both memory performance and JOLs, when participants were allowed to choose how long to study correct definitions. One possible explanation for the higher JOLs given to Read than Generate items is simply that participants had less time to study correct definitions in the Generate condition and for this reason were less confident about their ability to remember them. In the Choice condition the correct answer is on screen for the total trial time, even though participants do not know it is the correct answer until the last few seconds. Participants may give higher JOLs to Choice items than to Generate items because of this additional exposure. If we observe the same pattern of JOLs when participants are free to study targets for as long as they choose, this would strengthen the argument that the process of generating an error leads participants to perceive Generate items as harder to learn than Read or Choice items and that their JOLs are a reflection of this perception.

Similarly, if participants felt there was insufficient time to learn Generate items in the previous experiments, it is possible that they used time allocated to Read items for the rehearsal of Generate items and this could have led to the Generate over Read advantage in final test performance (cf. Slamecka & Katsaiti, 1987). Allowing participants to study each item for as long as they choose will obviate the need for displaced rehearsal of Generate items. If the Generate over Read advantage remains under these conditions, this will add further support to the notion that there is something about the act of generation which potentiates encoding of the correct answer, even when generation produces an error which is unrelated to the correct answer. In Experiment 3 we therefore had two groups: a self-paced (SP) group and an experimenter-paced (EP) group.

Method

Materials and procedure were as for Experiment 2A with three exceptions. The first concerned the timing of trials. For both groups, we allowed 10 s for entry of responses at study in the Generate and Choice conditions, in order to maximise the chance that participants would enter a complete and valid word. For the EP group, feedback time for these two conditions remained at 5 s, with 15 s to study Read items. For the SP group, correct definitions in the Generate and Choice conditions, and the word plus definition in the Read condition, were displayed until the participant clicked on a button labelled "Finished studying", at which point the JOLs screen was displayed just as in Experiment 2. We kept time to enter a response equal for the Generate and Choice conditions in both groups since we were interested in how participants allocated time to study correct answers, not time to generate or choose a response. In addition, always having 10 s to respond prevented participants from simply skipping over items as they might have done if they assumed there would be no benefit in making incorrect guesses. We wanted to ensure participants had time to go through the process of generating a response. Second, as in Experiment 2B, we told participants to expect a test in multiple choice format.

Third, we altered the format of the final multiple choice test in order to examine the effect on test performance of having available, at test, definitions that participants had generated themselves. Table 1 illustrates these test options. At test there were five options for each cue word: the correct definition; two previously-studied definitions from other items, one taken from each of the other two conditions; an incorrect definition generated by the participant (either for that very item, if it had appeared in the Generate condition, or for another item); and an incorrect definition chosen by the participant (either as the definition for that very item, if it had appeared in the Choice condition, or for another item). Each option presented therefore appeared in the test three times, as options for three different cue

words. For example, for a given participant, imagine that the word *carcanet* was studied in the Generate condition and the participant generated the definition *trumpet*. At test, *carcanet* appeared with its true definition, *necklace*; with *cup* (the definition for the word *hanap*, presented at study as a Read item); with *beggar* (the definition for the word *gaberlunzie*, presented at study as a Choice item); with *trumpet*, the definition generated by the participant at study; and with *dove*, the Choice option incorrectly chosen when the participant studied *peridot* (whose true definition is *gem*). Each of these options would also appear as options for words studied in the Read and Choice conditions. For example, the true definition for *carcanet*, *necklace*, would also appear as a lure definition for two other items (one Read, one Choice), e.g., for *rapparee* (*bandit*) and *barbet* (*bird*), while the generated definition *trumpet* would appear as the generated option for one Read and one Choice word, since these of course had no generated response of their own, e.g., for *mechlin* (*lace*) and *bistoury* (*knife*).

In cases where the participant either failed to generate a response for a Generate item at study, or entered a definition which was a true definition for another item in the experiment, this was replaced at test by a new lure for all three affected items. When participants selected the correct item at study in the Choice condition, this was replaced at test by one of the studied Choice lures for that cue word. These measures were taken to ensure that the final test did not include options which were simply blank (where no response had been generated) or repeated (e.g., to avoid the correct definition appearing both as the target and as the chosen option, in cases where the participant made the correct choice at study).

If the participant ran out of time to enter their chosen definition at study in the Choice condition, the program checked which of the study options corresponded to the partial answer and replaced it with this one at final test. For example, if the participant entered *versa* in response to the cue word *levisomnous*, the program compared this entry with the beginning of

all the available options, i.e. *observant*, *nocturnal*, *expectant* and *versatile*, and *versatile* would appear as an option at final test. If, on the other hand, the participant gave part of the correct answer, e.g. *observ*, the program would recognise this as correct and replace it with one of the other lures, e.g. *nocturnal.* For Generate items, however, it was impossible to program the task to check whether the response entered was a real and complete word and therefore impossible to prevent partial words and nonwords sometimes appearing as options at final test. We accepted that this would be likely to occur fairly frequently, given the time constraints on entry of the generated definition. Since the appearance of partial words and nonwords at test was likely to alert the participant to the fact that their own responses were being shown at test, we excluded from the analysis any participant whose final test options included any such items. This left 24 participants (20 female), average age 18.6, SD = .8, in the self-paced group and 16 (14 female), average age 18.3, SD = .6, in the experimenter-paced group, out of an original sample of 56 and 51, respectively. Participants were first year Psychology undergraduates who took part in fulfilment of a course requirement.

Results

At study, just one participant gave the correct response to just one Generate item, and the percentage of correct answers given for Choice items (M = 27.9, SD = 10.5) did not differ from the chance level of 25% ($t(39) = 1.73$, $p = .091$).

Final test performance

A primary aim of this experiment was to examine the effect on memory of being able to select items participants had generated or chosen themselves. We therefore included, in the analysis of the final test data, only those trials where none of the final test options had been replaced by new items. Every trial included in the final analysis, therefore, was one for which

the five options at test were the correct definition, two incorrect definitions from other items, one response generated by the participant and one incorrectly chosen by the participant or, in a case where the correct choice was made at study, an incorrect lure for that item from the study phase.

Test accuracy was evaluated by means of a 3 (Study Method) * 2 (Group) ANOVA. When all items were considered, whether correct or incorrect at study, there was a main effect of Study Method, $F(2,76) = 9.17$, $p < .001$, $\eta_p^2 = .19$, no effect of Group, $F(1,38) = .27$, $p = .605$, and no interaction, $F(2,76) = 1.57$, $p = .216$. Figure 2D shows the means collapsed across groups. Table 2 shows the means for each group. Generating produced better final test performance than reading, $t(39) = 5.22$, $p < .001$, $d = .56$. Twenty-seven participants (17 in the SP group, 10 in the EP group) showed a benefit of generating over reading, with only five (2 in the SP group, 3 in the EP group) showing the opposite pattern. Choosing also led to higher test scores than reading, $t(39) = 2.49$, $p = .017$ $d = .28$, with no difference between choosing and generating, $t(39) = 1.75$, $p = .089$, $d = .24$.

Next we repeated the analysis for just those items incorrect at study (Figure 2D). The assumption of sphericity was not met, so the Huynh-Feldt correction was applied. There was a main effect of Study Method, $F(1.70, 64.44) = 7.35$, $p = .002$, $\eta_p^2 = .19$, no main effect of Group, $F(1,38) = .19$, $p = .662$, and no interaction, $F(1.70, 64.44) = 1.53$, $p = .224$. Generating produced better performance than reading ($t(39) = 5.22$, $p < .001$, $d = .56$) , but incorrect choosing was no better than reading, $t(39) = 1.36$, $p = .181$, $d = .17$, and was less beneficial than generating, $t(39) = 2.08$, $p = .044$, $d = .32$, replicating the findings of Experiments 2A and 2B. Despite the availability, at test, of definitions participants had generated themselves, generating errors followed by studying corrective feedback led to better final test performance than either reading or incorrect choosing.

Were participants less susceptible to interference from definitions they had generated themselves than from definitions they had chosen themselves? We computed how often participants selected, at test, the very same definition that they had generated themselves for that item at study as a percentage of all the Generate items (M = 4.2, SD = 6.7), and how often they selected at test the same incorrect choice they had made at study as a percentage of all Choice items (M = 7.7, SD = 7.3), and compared these two figures. Participants were more inclined to persist with incorrectly chosen than with incorrectly generated definitions, $t(39) = 2.49$, $p = .017$. When we calculated the number of times a generated or chosen item was selected at test as a percentage of just the incorrect responses at test, the comparison remained significant, $t(39) = 2.21$, $p = .033$. The response a participant had generated at study was only selected at test 13.2% of the time (SD = 21.5), which was significantly lower than the chance level of 25% (since there were 4 incorrect options), $t(39) = 3.49$, $p < .001$, whereas an incorrectly chosen response was selected 24.0% of the time (SD = 22.6), which was no different from chance, $t(39) = .27$, $p = .786$.

This pattern suggests that, when a participant failed to remember the correct definition at test, an incorrect choice at study did not interfere with memory any more than other familiar definitions appearing as lures at test, being selected with the same frequency as other incorrect lures. This is in line with the findings of Experiment 2A, where participants selected their own incorrect choices with no greater frequency than other incorrect options which had appeared as lures at study. For Generate items, on the other hand, participants seemed to be able to recognise and reject their own incorrect study responses, selecting them less often than the other incorrect options at test.

We also conducted the main analysis (final test performance) for all 107 original participants. These data, which showed a similar pattern, can be found in Appendix D.

Judgments of learning

Did participants' JOLs show any awareness of the benefits of generating over reading and choosing? A 3 (Study Method) * 2 (Group) ANOVA revealed a main effect of study method for the item JOLs. The assumption of sphericity was not met, $\chi^2(2) = 11.10$, $p = .004$, so the Huynh-Feldt correction was applied, $F(1.69, 64.22) = 24.43$, $p < .001$, $\eta_p^2 = .39$. Choice JOLs were higher than both Generate, $t(39) = 6.95$, $p < .001$, $d = .80$, and Read JOLs, $t(39) = 3.24$, $p < .002$, $d = .41$, and Read JOLs were higher than Generate JOLs, $t(39) = 4.75$, $p < .001$, $d = .40$. These data, collapsed across groups, are shown in Figure 3C. There was no main effect of Group, $F(1,38)$ .48, $p = .491$, $\eta_p^2 = .013$, but there was an interaction between Study Method and Group, $F(1.69, 64.22) = 3.76$, $p = .035$, $\eta_p^2 = .09$. For the SP group, Choice JOLs were higher than Generate JOLs, $t(23) = 7.71$, $p < .001$, $d = .95$, and higher than Read JOLs, $t(23) = 3.92$, $p = .001$, $d = .64$, which were higher than Generate JOLs, $t(23) = 2.77$, $p = .011$, $d = .28$, replicating the pattern observed in Experiment 2A. Participants gave the lowest JOLs to Generate items even when they could choose how long to study correct definitions, suggesting that their low JOLs did not stem from a perception of having insufficient time to process the corrective feedback in the Generate condition. For the EP group, Read JOLs were again higher than Generate JOLs, $t(15) = 4.25$, $p = .001$, $d = .59$, and Choice JOLs were higher than Generate JOLs, $t(15) = 2.72$, $p = .016$, $d = .59$, but there was no difference between Read and Choice JOLs, $t(15) = .39$, $p = .701$, $d = .07$.

We repeated the ANOVA on JOLs for just those items which were incorrect at study. There was a main effect of Study Method, $F(1.62, 61.54) = 6.84$, $p = .004$, $\eta_p^2 = .15$ (Huynh-Feldt correction). Again, the lowest JOLs were given to Generate items. These JOLs were lower than both Read JOLs, $t(39) = 4.75$, $p < .001$, $d = .40$, and JOLs for Choice items incorrect at study, $t(39) = 3.05$, $p = .004$, $d = .38$, but now there was no difference between JOLs for Read items and those for Choice items incorrect at study, $t(39) = .11$, $p = .910$, $d =$

.02 (Figure 3C). Just as in Experiment 2, errorful generation produced better final test performance than either reading or incorrect choosing but participants failed to predict this benefit, giving the lowest JOLs to Generate items. There was no effect of Group, F(1,38) = .53, p = .469, η_p^2 = .01, but again there was an interaction between Study method and Group, $F(1.62, 61.54) = 3.78$, p = .037, η_p^2 = .09. In the SP group, Generate items were given lower JOLs than both Choice items incorrect at study, $t(23) = 3.04$, p = .006, d = .49, and Read JOLs, $t(23) = 2.77$, p = .011, d = .29, with no difference between Read and Choice, $t(23) = 1.34$, p = .193, d = .24. In the EP group, JOLs for Generate items were no different from JOLs for incorrect choices, $t(15) = .95$, p = .355, d = .17, but were lower than Read JOLs, $t(15) = 4.25$, p = .001, d = .59, and JOLs for Read items were higher than those for Choice items incorrect at study, t (15) = 2.66, p = .018, d = .40.

The finding that the advantage for Choice JOLs over Read JOLs disappeared (and, indeed, was reversed for the EP group) when we considered only Choice items incorrect at study, suggests that, as in Experiment 2, the high JOLs given to Choice items were driven by very high JOLs elicited by items correctly selected at study. Confirming this impression, JOLs for Choice items correct at study were significantly higher than JOLs for Choice items incorrect at study $F(1,38) = 57.6$, $p < .001$, η_p^2 = .60, with no difference between the groups, $F(1,38) = 1.83$, p = .185, η_p^2 = .05, and no interaction, $F(1,38) = .02$, p = .901, η_p^2 = 0 (Figure 3C). Replicating the findings of Experiment 2A and 2B, JOLs for Choice items correct at study were significantly higher than JOLs for Read items, $t(39) = 7.83$, $p < .001$. This confidence that Choice items correct at study would be better remembered than items incorrect at study was misplaced: In fact, the percentage of items correct at study which persisted to be correct at test (M = 77.81, SD = 26.18) was not significantly greater than the percentage of items incorrect at study which were converted to being correct at test (M = 73.05, SD = 22.99), $t(39) = 1.06$, p = .296, d = .19. As for the other experiments, we also

report the relationship between JOLs and final test accuracy in Appendix C. Figure 4C shows the means.

For the aggregate JOLs there was a main effect of Study method, $F(2,76) = 4.10$, $p = .020$, $\eta_p^2 = .10$ (Figure 3C). Choice JOLs were higher than Generate JOLs, $t(39) = 3.25$, $p = .002$, $d = .55$, with no difference between Generate and Read, $t(39) = 1.78$, $p = .083$, $d = .32$, or between Read and Choice, $t(39) = 1.27$, $p = .211$, $d = .25$. There was no difference between the groups, $F(1,38) = .37$, $p = .545$, $\eta_p^2 = .01$, but there was an interaction between Study method and Group, $F(2,76) = 3.81$, $p = .027$, $\eta_p^2 = .09$. The EP group gave higher aggregate JOLs to the Read condition ($M = 40.3$, $SD = 4.3$) than to the Choice ($M = 34.3$, $SD = 5.2$) or Generate ($M = 29.8$, $SD = 3.6$) conditions but the difference between the study methods fell short of significance, $F(2,30) = 2.78$, $p = .078$, $\eta_p^2 = .16$. For the SP group, the difference was significant, $F(2, 46) = 5.80$, $p = .006$, $\eta_p^2 = .20$. Aggregate JOLs for the Choice condition ($M = 40.8$, $SD = 21.1$) were higher than JOLs for the Generate condition ($M = 26.8$, $SD = 17.6$), $t(23) = 3.46$, $p = .002$, $d = .72$, and for the Read condition ($M = 28.7$, $SD = 15.9$), $t(23) = 2.32$, p = .030, $d = .65$, with no difference between Read and Generate aggregate JOLs, $t(23) = .49$, $p = .632$, $d = .12$.

Study time

Did participants in the SP group spend any longer studying correct definitions in one study condition than another? There was a significant difference between conditions in study time, $F(2, 46) = 35.44$, $p < .001$, $\eta_p^2 = .60$. Participants spent longer studying definitions for Read items ($M = 8.02$ s, $SD = 4.35$) than for both Generate items ($M = 6.68$, $SD = 4.14$), t(23) = 3.75, $p = .001$, $d = .32$, and Choice items ($M = 5.35$, $SD = 3.90$), t(23) = 9.25, $p < .001$, $d = .65$, and longer for Generate than for Choice items, $t(23) = 4.41$, $p < .001$, $d = .33$.

Did accuracy at study affect how long participants spent studying correct definitions? In the Choice condition participants spent significantly longer studying items which they had got wrong at study (M = 5.75 sec, SD = 3.81) than items they got right (M = 4.44, SD = 4.11), t (23) = 3.64, $p < .001$, d = .33. Together with the finding that JOLs were also higher for Choice items participants got right at study than for those they got wrong, these results suggest that participants' experience of success in the Choice condition led them to perceive correctly guessed items as easier to learn, both devoting less time to their study and giving them higher JOLs. Interestingly, although participants spent longer studying Choice items which were incorrect at study than they spent on correct items, they spent even longer on Generate items incorrect at study, $t(23) = 3.13$, $p = .005$. The fact that study time, like JOLs, was influenced by the accuracy of the participant's guess at study in the Choice condition, adds further support to our proposal that participants' perception of the difficulty of remembering an item is affected by the outcome of an event preceding study.

The relationship between study time and final test performance is subject to item selection effects, but these data are included in Appendix E for completeness.

Discussion

Experiment 3 replicated the benefit of generating errors over both reading and incorrect choosing observed in our previous experiments and confirmed that this benefit persisted even though participants had the opportunity to select, at test, an incorrect guess they had generated at study, and even though test lures were equally familiar across the three study conditions. Furthermore, the benefit of generating over reading in the SP group suggests that the superiority of the Generate condition was not due to displaced rehearsal of Generate items during study of Read items, since allowing participants to choose how long to study should also eliminate any need for displaced rehearsal. Participants' JOLs also

replicated the pattern observed previously, in that item JOLs were significantly lower in the Generate condition than in either the Read or Choice conditions. This was true even when participants were allowed to study correct definitions for as long as they liked, suggesting that the low JOLs for Generate items observed in Experiment 2 did not stem from a perception that there was insufficient time to process correct feedback.

As in Experiments 2A and 2B, participants gave much higher JOLs to Choice items they guessed correctly at study than to Choice items they guessed incorrectly. JOLs for Choice items incorrect at study were higher than JOLs for Generate items. Even when they made errors, participants believed they would learn better by the Choice than the Generate method, though in reality the reverse was true. The amount of time participants in the SP group allocated to studying correct definitions also reflected this misconception. Although they spent significantly more time studying Choice items that were incorrect at study than correct ones, they spent even longer on Generate items, suggesting they believed the Choice items would be more easily learned even when they had made an error. In the Choice condition the correct answer is present for the whole of the trial time, although it is only revealed as the correct one in the last few seconds, after the participant's choice has been made. Therefore even when an incorrect choice is made, the answer, when it appears, is already familiar and may, as a result, be processed more fluently, leading to higher JOLs and shorter study times. However, this same fluency may also mean that these items are processed less deeply, leading to poorer subsequent memory. Indeed, at final test, participants were more likely to select the same incorrect choice they had made at study than an incorrect response they had generated.

It is interesting that the advantage for generating over reading was observed even though participants spent longer on Read than Generate items, suggesting that there was something about the process of generating a guess which enabled more efficient processing

of the correct answer. Against this interpretation, however, study times for Read items include time spent processing the cue word, whereas in the Generate and Choice conditions this has already taken place before timing begins, a factor which may at least partly account for the longer study times for Read items.

General Discussion

In all four experiments reported here, we observed a benefit of errorful generation over reading when participants learned definitions of previously unfamiliar English words or translations of novel foreign vocabulary, even though the correct answer was displayed for a much shorter time in the Generate than in the Read condition. This effect was observed even when participants regulated their own study time. Generating errors followed by feedback was also more beneficial than incorrect choice in all experiments except the first, in which the difference fell just short of significance. Participants' JOLs, however, showed a very different pattern, with the lowest item JOLs consistently being given to Generate items and the highest to Choice items. These high Choice JOLs were largely driven by high JOLs for items guessed correctly at study, suggesting that participants used their own performance at study as a basis for predicting their future test performance. To assist in conveying the main data pattern, and because all experiments employed a within-subjects design with the same study conditions, Table 3 summarises test accuracy and JOLs (and their 95% confidence intervals) by aggregating data from all four experiments.

Theoretical accounts of the benefits of errorful generation.

Why did generating errors followed by feedback lead to better memory performance than reading or choosing? Grimaldi and Karpicke (2012) proposed that generating errors can only benefit memory to a greater extent than passive reading when the correct answer is

already known and is activated during the initial generation attempt. Others (Huelser & Metcalfe, 2012; Hays et al., 2013) have also argued that an existing cue-target relationship is necessary for this benefit to be observed. Our findings are inconsistent with these proposals. Whether participants learned definitions for obscure English words or translations for previously unstudied foreign language words, we consistently observed a benefit of generating over reading, despite the fact that the stimulus materials were previously unknown to participants so there could be no pre-existing association to activate and reinforce. The benefit of errorful generation cannot, therefore, be solely due to a strengthening, through testing, of items related to the cue and the target. Our findings suggest that generation can benefit memory even when it yields many errors, by making the encoding of subsequent feedback more effective than when it is not preceded by a generation attempt.

The errorful generation benefit was observed whether or not the participant's initial guess appeared as one of the final test options. When, in Experiment 3, participants had the option, at final test, of selecting their own generated responses and their own chosen responses, they were more likely to pick a definition they had incorrectly selected at study in the Choice condition than one they had incorrectly generated in the Generate condition, which may have contributed to the advantage for generating over choosing. However, when participants made an error at test in the Choice condition, they chose their own original error at a rate no higher than chance, suggesting that incorrect choosing caused no particular detriment to memory. On the other hand they were very good at rejecting their own incorrect generations, choosing these at test at a rate significantly below chance.

Although incorrect choosing led to lower test performance than errorful generation, it never led to worse performance than reading. This was true even in Experiment 2B, when participants showed some tendency to persist with their own incorrect choices when compared with making a new error. However, in all experiments, when participants made

incorrect choices at study they were more likely to answer those items correctly at test than to make any type of error. Making incorrect choices was neither harmful nor helpful to memory compared with reading.

Grimaldi and Karpicke (2012) and Huelser and Metcalfe (2012) found no advantage for generating over reading when cue and target were unrelated to one another. However, unlike in our study, their paradigm involved cues which would have been very familiar to participants, and it encouraged generation of incorrect responses with a strong pre-existing association to the cue. If generating a guess activates not only the guess but many other associated concepts, then these pre-existing associations may well have interfered, at test, with retrieval of the designated "correct" but unrelated answer, counterbalancing and masking a beneficial effect of generation on the encoding of feedback. In our study the cue words were unfamiliar to participants pre-experimentally so had no pre-existing associations with participants' guesses and there was therefore less potential for interference from the incorrect guess at final test. Thus generation can benefit memory both by strengthening existing associations and by potentiating the encoding of feedback. Where the generated guess is related to the cue but not to the target, the strengthening of this interfering cue-guess association may cancel out the benefit generation confers on the processing of feedback.

Of course in a real world situation it is highly unlikely that someone would produce an answer to a question which had a strong pre-existing relationship to the question but was not at all related to the correct answer. Our design allowed us to observe a benefit to the encoding of feedback which was uncontaminated by an opposite, interfering effect of generation on the generated item itself. Thus we were able to show that the errorful generation benefit does not depend on a strengthening of pre-existing associations that enhance memory as a result of also being associated to the target. Instead, the act of generation per se, regardless of what is generated, makes the encoding of subsequent

feedback more effective. This is interesting from a theoretical perspective but it is also important from a practical point of view, since our task is much closer to the kind of learning situations people are likely to encounter in educational settings. In addition, this line of research is relevant to any real world situation where novel information is to be learned, for example when learning concepts in science, economics, politics, philosophy, literary theory or art. An understanding of the effect of errors is also particularly important in a world in which technological innovations mean that students are increasingly creating their own online content, using tools such as discussion boards, wikis and self-assessment software packages, creating ample opportunity for the generation of erroneous material.

Generating errors, then, was beneficial to memory even when there was no pre-existing association which could be reinforced by corrective feedback. Incorrect guessing in the Choice condition, however, was less effective, producing equivalent memory performance to reading. This suggests that there was something about the active process of generating a response, rather than merely selecting one, which facilitated encoding of corrective feedback, even when the generated response was incorrect. Our proposal is that participants focused more attention on correct feedback in the Generate condition than in the Read or Choice conditions. M. J. Kang et al. (2009) found that curiosity to know the correct answer, following the making of an error, enhanced subsequent memory. A possible explanation for our findings is that participants' curiosity about the correct answer was aroused to a greater extent in the Generate than in the Choice condition, perhaps because searching for an answer in the Generate condition involves more active engagement and effort than simply selecting one in the Choice condition or perhaps because, in the Choice condition, the participant knows that the correct answer will be one of the limited number of possible options displayed for selection, and the answer when it comes is therefore less surprising.

Butterfield and Metcalfe (2001; 2006) and Fazio and Marsh (2009) showed that participants processed corrective feedback more effectively when it did not match their expectations. This occurred when they found they had made an error after being highly confident that they were right, and also when they got an answer right despite having low confidence in that answer. Butterfield and Metcalfe (2001) proposed that surprising feedback captured participants' attention which, in turn, led to more effective encoding. In our study, where participants had to learn completely novel material, they would have known that their generated guesses were almost certain to be wrong but they may still have been highly curious to learn the correct answer, and the discrepancy between their own response and the actual answer may have induced a similar sense of surprise and led to greater attention being applied to encoding that answer, consistent with other research showing that such discrepancies drive learning (Rescorla & Wagner, 1972).

There are other possible explanations for the benefit of generating over reading and incorrect choosing. Kornell et al. (2009) suggested that the advantage for generating over reading observed in their weak-associate task may have occurred because searching for an answer encourages deep processing of the question, activating related concepts and facilitating integration of the correct answer. This could be the case in a situation where the cue is a familiar word and the correct association is already present in participants' memories, as it was in their experiments and those of Grimaldi and Karpicke (2012) and Huelser and Metcalfe (2012), but it cannot explain our results, where the cue was completely novel. However, generating a response is likely to activate many concepts which could, even if unrelated to the correct answer, create a distinctive context for the learning of the answer and serve as retrieval cues on a subsequent recall attempt. This may be particularly effective if, as we are suggesting, difficulty experienced during learning leads to enhanced attention to

corrective feedback. In this case, the pattern of results we observed could be due to a combination of factors.

Metacognition and errorful generation

Ours is the first errorful generation study to ask participants to make a judgment of learning after studying each item. These judgments of learning are interesting for both theoretical and practical reasons. Our finding that JOLs were heavily influenced by success at study is theoretically interesting because it suggests that not only are people's perceptions of their learning strongly affected by fluency of processing at encoding but that this fluency is itself influenced by the production of either a correct or incorrect answer immediately prior to encoding of corrective feedback. JOLs are important from a practical point of view because of the effect they may have on the study strategies people are likely to adopt in everyday learning situations. However, they are also informative with regard to helping to understand the errorful generation benefit we have observed in the current set of experiments. Participants consistently gave lower JOLs to Generate items than to Read or Choice items, even when, in Experiment 3, they had the opportunity to study correct answers for as long as they liked. This perception that Generate items were harder to learn than items studied under the other two methods may have led participants to apply more effort to the learning of corrective feedback for Generate items, consistent with research showing that studying difficult items requires greater cognitive effort which in turn enhances memory (Ellis, Thomas, & Rodriguez, 1984; Tyler, Hertel, McCallum, & Ellis, 1979; Zaromb, Karpicke, & Roediger, 2010).

Interestingly, Huelser and Metcalfe (2012) asked participants, after their final test, to rank the study methods according to their effectiveness. Participants ranked reading as more effective than generating, even though their own test performance had produced the opposite

result. This suggests that they did not remember which items had been studied under which method and were making their rankings on the basis of their expectations about the relative efficacy of the two methods. In our study, aggregate JOLs given at the end of the study phase may also have reflected participants' expectations rather than their actual experience, and these too showed that participants expected memory to be poorest for items studied in the Generate condition.

Whereas these post-study or post-test summary ratings may be driven by a pre-existing expectation or heuristic regarding the different methods (what Matvey, Dunlosky & Gutentag, 2001, call an analytic inference), item JOLs may give us a measure of participants' perception of the ease or difficulty of learning for each individual item at the moment of study. The fact that these item JOLs were also lower for Generate items than for the other study methods even when participants could control how long they studied feedback, suggests that the effort involved in coming up with a response, followed by the making of an error, led to a perception that these items were more difficult to learn than items in the other conditions. Furthermore, JOLs for Choice items were heavily influenced by whether or not the correct definition had been chosen at study. In particular, Choice JOLs for items correct at study were always higher than JOLs for Read items, even when participants gave similar aggregate JOLs to Choice and Read items (as in Exp 2B and Exp 3), suggesting that success at study for Choice items gave participants high confidence that they would also remember those items at final test. As can be seen in Table 3, Choice items correct at study were indeed better remembered than Read items, in line with participants' JOLs. However, final test performance for these items was similar to performance for Generate items while JOLs were substantially higher, the highest JOLs being given to Choice items correct at study and the lowest to Generate items. Ours is also the first errorful generation study to include a Choice condition. By creating a situation in which participants make both correct and incorrect

guesses, and by collecting item by item JOLs, we have been able to show that participants' perception of their learning of correct answers is strongly affected by whether or not they happened to guess correctly at study.

Experiment 3 showed that participants were more likely to reproduce, at final test, an incorrect guess made at study in the Choice condition than one made in the Generate condition. However, Choice JOLs for items incorrect at study were often higher (Exp 2B and Exp 3) than JOLs for Generate items, even though both involved making an error at study, suggesting that the making of an error was not the only factor affecting JOLs. In the Choice condition, the options available at study constitute the "search set" for the given item. Here, far from facilitating encoding of corrective feedback, options in the search set interfered with correct recognition at test to a greater extent than generated guesses did. The greater familiarity of the correct answer when it was presented as feedback in the Choice condition may have led to higher JOLs but also to less effort applied to encoding, leading to a dissociation between JOLs and final test performance. This is consistent with our proposal that fluency at encoding influences participants' metacognitive judgments which in turn affect the degree of effort or attention applied to the processing of corrective feedback. However, participants do not realise that the extra effort involved in processing Generate feedback will lead to better memory, so they still give low JOLs to these items. Of course, we have not measured effort or attention directly but this proposal is consistent with other research suggesting that ease of processing leads to higher JOLs (Castel et al., 2007; Koriat, 2008; Hertzog et al., 2003; Schwartz, Benjamin & Bjork, 1997) and that both curiosity (Berlyne and Normore, 1972, M. J. Kang et al., 2009) and effort (e.g., Bjork, 1994; Ellis, Thomas, & Rodriguez, 1984; Tyler et al, 1979; Zaromb, Karpicke, & Roediger, 2010) lead to enhanced memory. It will be interesting for future research to examine effort more directly.

Conclusion

To generalise the pedagogical relevance of this work, it will be important for future research to examine whether the same benefit occurs when memory is tested under different conditions, such as with a recall final test format or after a delay. Also, as we mentioned in the introduction, a benefit from errors at encoding is striking in view of the large "errorless learning" literature (e.g., Baddeley & Wilson, 1994), which employs a very different method from the current study. Seeking to determine the factors that yield a benefit of error generation over reading in our procedure but a benefit of reading over error generation in some "errorless learning" situations is an important question for future research. Much of the "errorless learning" literature focuses on the design of interventions for memory-impaired populations. The avoidance of errors frequently advocated in this literature means that, often, generation too is avoided, with the result that these interventions fail to make optimal use of the benefits of generation. A better understanding of how generation can enhance learning even when it produces errors is therefore critical for designing the most effective interventions, an issue which is increasingly important in a world with an ageing population (see also Middleton & Schwartz, 2012). Much of the errorless learning literature was inspired by principles from studies of animal learning during the behaviorist era which emphasised the importance of minimizing errors (see Clare & Jones, 2008, for a review). Our findings demonstrate that generating responses followed by feedback is helpful to memory even when many errors are generated, compared with errorless studying without generation.

In conclusion, generating errors benefitted vocabulary learning even when the items to be learned had no pre-existing associations, but participants did not predict this benefit. In fact their JOLs were heavily influenced by success or failure at study. If participants apply more effort, or more effective encoding strategies, to items they find difficult to learn this could account, at least in part, for the surprising benefit of errorful generation. Previous

studies (Grimaldi & Karpicke, 2012; Huelser & Metcalfe, 2012) observed no benefit of generating over reading when there was no pre-existing association between cue and target. However, this was in a task in which participants were encouraged to produce guesses which were highly related to a familiar cue and then to learn a new association which bore no relation to it, a scenario which is rarely likely to occur in everyday life. In a more educationally relevant learning scenario, we found that generating errors could be helpful to memory even during the learning of novel material but that participants were strikingly unaware of this benefit.

References

Allen, G. A., Mahler, W. A., & Estes, W. K. (1969). Effects of recall tests on long-term retention of paired associates. *Journal of Verbal Learning and Verbal Behavior, 8,* 463 – 470.

Arnold, K. M., & McDermott, K. B. (2012, July 9). Test-potentiated learning: Distinguishing between direct and indirect effects of tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition.* Advance online publication. doi: 10.1037/a0029199

Baddeley, A., & Wilson, B. A. (1994). When implicit memory fails: amnesia and the problem of error elimination. *Neuropsychologia, 32*(1), 53 – 68.

Bahrick, H. P., & Hall, L. K. (2005). The importance of retrieval failures to long-term retention: A metacognitive explanation of the spacing effect. *Journal of Memory and Language, 52,* 566 – 577.

Balota, D.A., Yap, M.J., Cortese, M.J., Hutchison, K.A., Kessler, B., Loftis, B., Neely, J.H., Nelson, D.L., Simpson, G.B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods, 39,* 445-459.

Begg, I., Vinski, E., Frankovich, L., & Holgate, B. (1991). Generating makes words memorable, but so does effective reading. *Memory and Cognition, 19*(5), 487 - 497.

Berlyne, D. E., Carey, S. T., Lazare, S. A., Parlow, J., & Tiberius, R. (1968). Effects of prior guessing on intentional and incidental paired-associate learning. *Journal of verbal learning and verbal behaviour,* 750 – 759.

Berlyne, D.E., & Normore, L.F. (1972). Effects of prior uncertainty on incidental free recall. *Journal of Experimental Psychology, 96*(1), 43-48.

Bjork, R. A. (1975). Retrieval as a memory modifier. In R. Solso (Ed.), Information processing and cognition: The Loyola Symposium (pp. 123-144). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes* (Vol. 2, pp. 35–67). Hillsdale, NJ: Erlbaum.

Butterfield, B., & Metcalfe, J. (2001). Errors committed with high confidence are hypercorrected. *Journal of Experimental Psychology, Learning, Memory, and Cognition, 27* (6), 1491 – 1494.

Butterfield, B., & Metcalfe, J. (2006). The correction of errors committed with high confidence. *Metacognition Learning, 1,* 69 – 84.

Butler, A. C., & Roediger, H. L. III. (2007).Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology, 19* (4-5), 514 – 527.

Butler, A. C., & Roediger, H. L. III. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple choice testing. *Memory and Cognition, 36* (3), 604 – 616.

Carpenter, S. K., & DeLosh, E. L. (2005). Application of the testing and spacing effects to name learning. *Applied Cognitive Psychology, 19,* 619-636.

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory and Cognition, 34* (2), 268 – 276.

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition,35*(6), 1563 – 1569.

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory and Cognition, 20*(6), 633 – 642.

Castel, A. D., McCabe, D. P., & Roediger, H. L. III. (2007). Illusions of competence and overestimation of associative memory for identical items: Evidence from judgments of learning. *Psychonomic Bulletin and Review, 14*(1), 107 – 111.

Clare, L., & Jones, R. S. P. (2008). Errorless learning in the rehabilitation of memory impairment: A critical review. *Neuropsychological Review, 18*, 1 – 23.

Coltheart, M. (1981). The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology*, 33A, 497-505.

Dunlosky, J., Serra, M. J., Matvey, G., & Rawson, K. A. (2005). Second order judgments about judgments of learning. *Journal of General Psychology, 132*(4), 335 – 346.

Ellis, H.C., Thomas, R. L., & Rodriguez, I. A. (1984). Emotional mood states and memory: elaborative encoding, semantic processing, and cognitive effort. *Journal of Experimental Psychology: Learning, Memory, and Cognition,10*(3), 470 – 482.

Fazio, L. K., & Marsh, E. J. (2009). Surprising feedback improves later memory. *Psychonomic Bulletin and Review, 16* (1), 88 – 92.

Fischhoff, B. (1977). Perceived informativeness of facts. *Journal of Experimental Psychology: Human Perception and Performance, 3*(2), 349 – 358.

Forlano, G., & Hoffman, N. M. H. (1937). Guessing and telling methods in learning words of a foreign language. *Journal of Educational Psychology, 28*(8), 632 – 636.

Fritz, C. O., Morris, P. E., Bjork, R. A., Gelman, R., & Wickens, T. D. (2000). When further learning fails: Stability and change following repeated presentation of text. *British Journal of Psychology, 91,* 493 – 511.

Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Experimental Psychology, 81* (3), 392 – 399.

Grimaldi, P. J., & Karpicke, J. D. (2012). When and why do retrieval attempts enhance subsequent encoding? *Memory and Cognition, 40*(4), 505 – 513.

Hammer, A., Kordon, A., Heldmann, M., Zurowski, B., & Munte, T.F. (2009). Brain potentials of conflict and error-likelihood following errorful and errorless learning in obsessive-compulsive disorder. *PLoS One, 4,* e6553.

Haslam, C., Hodder, K.I., & Yates, P.J. (2011). Errorless learning and spaced retrieval: How do these methods fare in healthy and clinical populations? *Journal of Clinical and Experimental Neuropsychology, 33*(4), 432 – 447.

Haslam, C., Moss, Z., & Hodder, K. (2010). Are two methods better than one? Evaluating the effectiveness of combining errorless learning with vanishing cues. *Journal of Clinical and Experimental Neuropsychology, 32*, 973 – 985.

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*(1), 290 – 296.

Hertzog, C., Dunlosky, J., Robinson, A.E., & Kidder, D. P. (2003). Encoding fluency is a cue used for judgments about learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*(1), 22 – 34.

Hogan, R. M., & Kintsch, W. (1971). Differential effects of study and test trials on long-term recognition and recall. *Journal of Verbal Learning and Verbal Behavior, 10,* 562 – 567.

Howard, M. C., & Kahana, M. J. (2002). When does semantic similarity aid episodic retrieval? *Journal of Memory and Language, 46,* 85 – 98.

Huelser, B., & Metcalfe, J. (2012). Making related errors facilitates learning, but learners do not know it. *Memory and Cognition, 40*(4), 514 – 527.

Izawa, C. (1970). Optimal potentiating effects and forgetting-prevention effects of tests in paired-associate learning. *Journal of Experimental Psychology, 83* (2), 340 – 344.

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*(4), 626 – 635.

Kang, M.J., Hsu, M., Krajbich, I.M., Loewenstein, G., McClure, S.M., Wang, J.T., & Camerer, C.F. (2009). The wick in the candle of learning: Epistemic curiosity activates reward circuitry and enhances memory. *Psychological Science, 20*(8), 963 – 973.

Kang, S. H. K., McDermott, K.B., & Roediger, H. L. III. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology, 19* (4-5), 528 – 558.

Kang, S. H. K., Pashler, H., Cepeda, N. J., Rohrer, D., Carpenter, S. K., & Mozer, M. C. (2011). Does incorrect guessing impair fact learning? *Journal of Educational Psychology, 103*(1), 48 – 59.

Karpicke. J. D., & Roediger, H. L. III. (2008). The critical importance of retrieval for learning. *Science, 31,* 966 – 968.

Kay, H. (1955). Learning and retaining verbal material. *British Journal of Psychology,46(2),* 81 – 100.

Kessels, R.P.C., Boekhorst, S.T., & Postma, A. (2005). The contribution of implicit and explicit memory to the effects of errorless learning: A comparison between young and older adults. *Journal of the International Neuropsychological Society, 11,* 144–151.

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.

Koriat, A. (2008). Easy comes, easy goes? The link between learning and remembering and its exploitation in metacognition. *Memory and Cognition, 36*(2), 416 – 428.

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35* (4), 989 – 998.

Kucera, H., & Francis, W.N. (1967). *Computational Analysis of Present-Day American English*. Providence, RI: Brown University Press.

Kuo, T., & Hirshman, E. (1996). Investigations of the testing effect. *American Journal of Psychology, 109* (3), 451 – 464.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes, 25,* 259-284.

Matvey, G., Dunlosky, J., & Guttentag, R. (2001). Fluency of retrieval at study affects judgments of learning (JOLs): An analytic or nonanalytic basis for JOLs? *Memory & Cognition, 29(2),* 229 – 233.

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology, 19* (4-5), 494 – 513.

McDaniel, M. A., Wildman, K. M., & Anderson, J. L. (2012). Using quizzes to enhance summative-assessment performance in a web-based class: An experimental study. *Journal of Applied Research in Memory and Cognition, 1*(1), 18 – 26.

McElroy, L. A., & Slamecka, N. J. (1982). Memorial consequences of generating non-words – implications of semantic memory interpretations of the generation effect. *Journal of Verbal Learning and Verbal Behavior, 21* (3), 249 – 259.

Middleton, E. L., & Schwartz, M. F. (2012). Errorless learning in cognitive rehabilitation: a critical review. *Neuropsychological Rehabilitation: An International Journal, 22*(2), 138 – 168.

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95,* 109–133.

Nelson. T. O. (1993). Judgments of learning and the allocation of study time. *Journal of Experimental Psychology:General, 122*(2), 269 – 273.

Pashler, H., Bain, P. M., Bottge, B. A., Graesser, A., McDaniel, M. A., & Metcalfe, J. (2007). Organizing instruction and study to improve student learning (NCER 2007–2004).

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When Does Feedback Facilitate Learning of Words? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31* (1), 3–8.

Pashler, H., Zarow, G., & Triplett, B. (2003). Is temporal spacing of tests helpful even when it inflates error rates? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*(6), 1051 – 1057.

Pressley, M., Tanenbaum, R., McDaniel , M. A., & Wood, E. (1990). What happens when university students try to answer prequestions that accompany textbook material? *Contemporary Educational Psychology, 15,* 27 – 35.

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language, 60* (4), 437 – 447.

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: mediator effectiveness hypothesis. *Science, 330,* 335.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (eds), Classical conditioning II: current research and theory (pp64 – 99). New York: Appleton-Century-Crofts.

Richland, l. E., Kornell, N., & Kao, L. S. (2009). The pretesting effect: do unsuccessful tests enhance learning? *Journal of Experimental Psychology: Applied, 15*(3), 243 – 257.

Roediger, H.L., III, & Karpicke, J.D. (2006a). The power of testing memory: Basic research and implications for educational practice. *Perspectives of Psychological Science, 1,* 181-210.

Roediger, H. L. III., & Karpicke, J.D. (2006b). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249 – 255.

Roediger, H. L. III., & Marsh, E. J. (2005). The positive and negative effects of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31* (5), 1155 – 1159.

Schwartz, B. L., Benjamin, A.S., & Bjork, R.A. (1997). The inferential and experiential bass of memory. *Current Directions in Psychological Science, 6*(5), 132 – 137.

Slamecka, N. J., & Graf, P. (1978). The generation effect: delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory, 4* (6), 592 - 604.

Slamecka, N.J., & Fevreiski, J. (1983). The generation effect when generation fails. *Journal of Verbal Learning and Verbal Behavior,22,* 153 – 163.

Slamecka, N.J., & Katsaiti, L.T. (1987). The generation effect as an artifact of selective displaced rehearsal. *Journal of Memory and Language, 26*(6), 589 - 607.

Tyler, S. W., Hertel, P. T, McCallum, M. C., & Ellis, H. C. (1979). Cognitive effort and memory. *Journal of Experimental Psychology: Human Learning and Memory, 5,* 607-617.

Whitten, W. B., & Bjork, R. A. (1977). Learning from tests: The effects of spacing. *Journal of Verbal Learning and Verbal Behavior, 16,* 465 - 478.

Zaromb, F. M., Karpicke, J. D., & Roediger, H. L. III. (2010). Comprehension as a basis for metacognitive judgments: effects of effort after meaning on recall and metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(2), 552 – 557.

Footnotes

1. In Experiment 1, the number of correct generations was largely driven by the responses of one participant who generated correct guesses for over half the items in the Generate condition and three quarters in the Choice condition. We reran the analysis with this participant excluded and the results were unaffected. In Experiment 2 the stimulus materials were changed so as not to include any of the items which were correctly guessed in Experiment 1.
2. Somewhat surprisingly, and unlike in Experiments 1 and 3, participants in Experiments 2A and 2B selected the correct definition/translation at study in the Choice condition at a rate higher than chance. In case this reflected some existing familiarity with the words and their definitions (though this seems especially unlikely in the case of Experiment 2B) which might have affected the subsequent analyses, we removed four participants from Experiment 2A and two from Experiment 2B who achieved particularly high scores at study in the Choice condition. Without these participants, Choice performance at study was no longer significantly greater than chance in either experiment. We recomputed all the subsequent analyses with these participants excluded and none of the conclusions was changed, except (as noted in the relevant Results section) for two comparisons involving JOLs for Choice items in Experiment 2B, which come into line with the findings of Experiment 2A when these two participants are excluded.

			Options appearing at final test				
Word	**Study condition**		**Correct definition**	**Definition from another item**	**Definition from another item**	**Incorrect generation**	**Incorrect choice**
carcanet	**Generate**		**necklace**	**cup**	**beggar**	***trumpet***	**dove**
hanap	**Read**		**cup**	**knife**	**post**	**shaman**	**bomb**
gaberlunzie	**Choice**		**beggar**	**post**	**lace**	**pray**	***healer***
peridot	**Choice**		**gem**	**fool**	**bandit**	**shaman**	***dove***
rapparee	**Read**		**bandit**	**necklace**	**bird**	**pray**	**judge**
barbet	**Choice**		**bird**	**anchor**	**necklace**	**sweet**	***fence***
mechlin	**Read**		**lace**	**fool**	**gossip**	**trumpet**	**fence**
bistoury	**Choice**		**knife**	**fish**	**mask**	**trumpet**	***judge***

Table 1. Example final test options in Experiment 3. Definitions generated or chosen at study for the given item are italicized.

	Final test performance		Item JOLs		Aggregate JOLs	
	SP	EP	SP	EP	SP	EP
Read	62.84 (18.09)	69.06 (19.75)	34.35 (15.03)	35.45 (14.30)	28.71 (15.90)	40.31 (17.29)
Generate	73.65 (17.70)	77.31 (12.84)	30.25 (13.62)	27.77 (12.77)	26.75 (17.61)	29.81 (14.21)
Choice (All)	71.40 (18.16)	69.89 (22.99)	44.29 (15.82)	36.51 (16.65)	40.83 (21.09)	34.25 (20.82)
Choice (incorrect at study)	69.84 (21.85)	67.34 (25.35)				

Table 2. Mean final test performance and item and aggregate judgments of learning (JOLs) in Experiment 3 (SD in brackets).

	Read	Generate	Choice	
			Correct at study	Incorrect at study
Test accuracy	76.30 (74.9 – 77.7)	84.36 (82.9 – 85.9)	84.29 (82.0 – 86.6)	78.40 (76.3 – 80.5)
JOLs	34.48 (33.2 – 35.8)	29.63 (28.3 – 30.9)	54.24 (51.7 – 56.8)	33.20 (31.8 – 34.6)

Table 3. Mean final test accuracy and item judgments of learning (JOLs) for participants across all experiments (95% within-subjects confidence intervals in brackets).

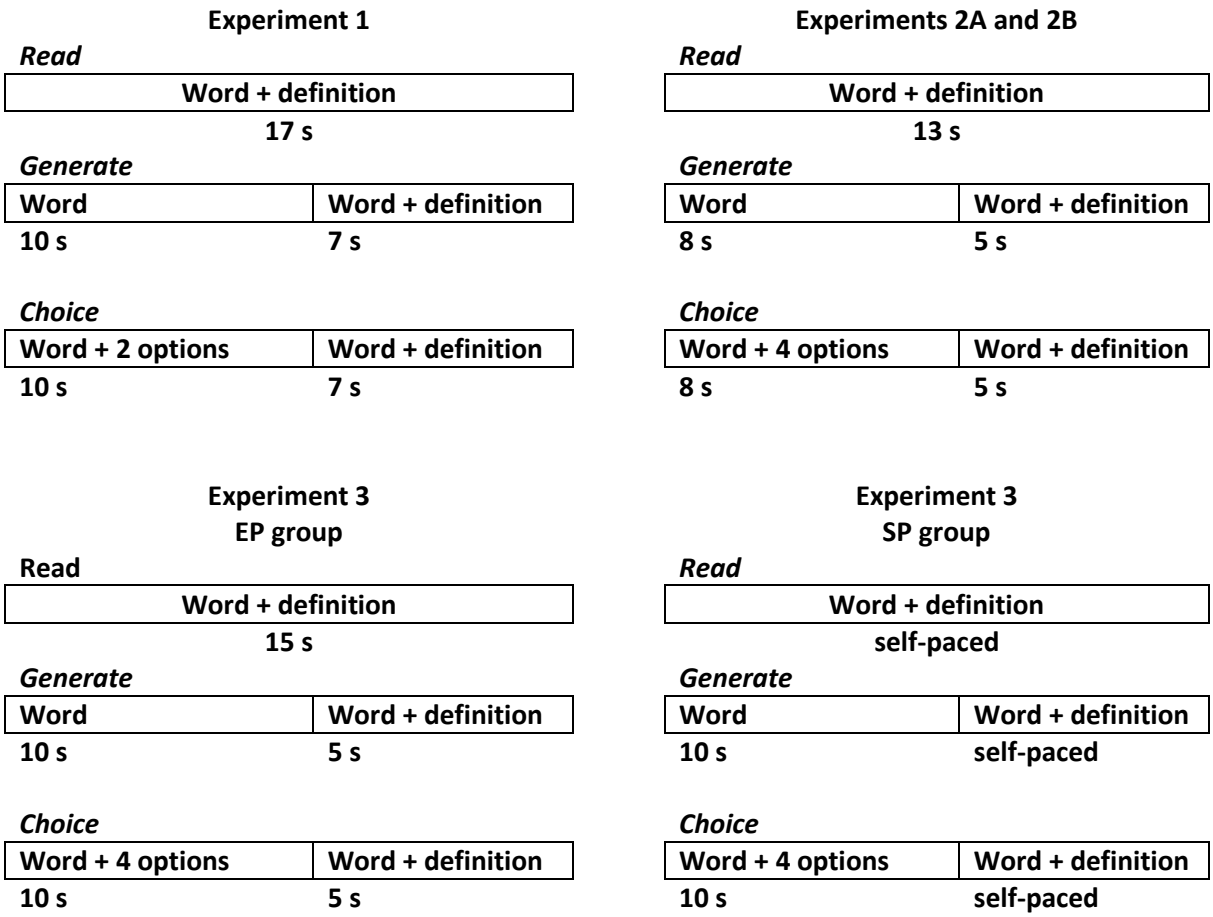

Figure 1. Study trial procedure for Experiments 1 – 3. EP = Experimenter-paced, SP = Self-paced.

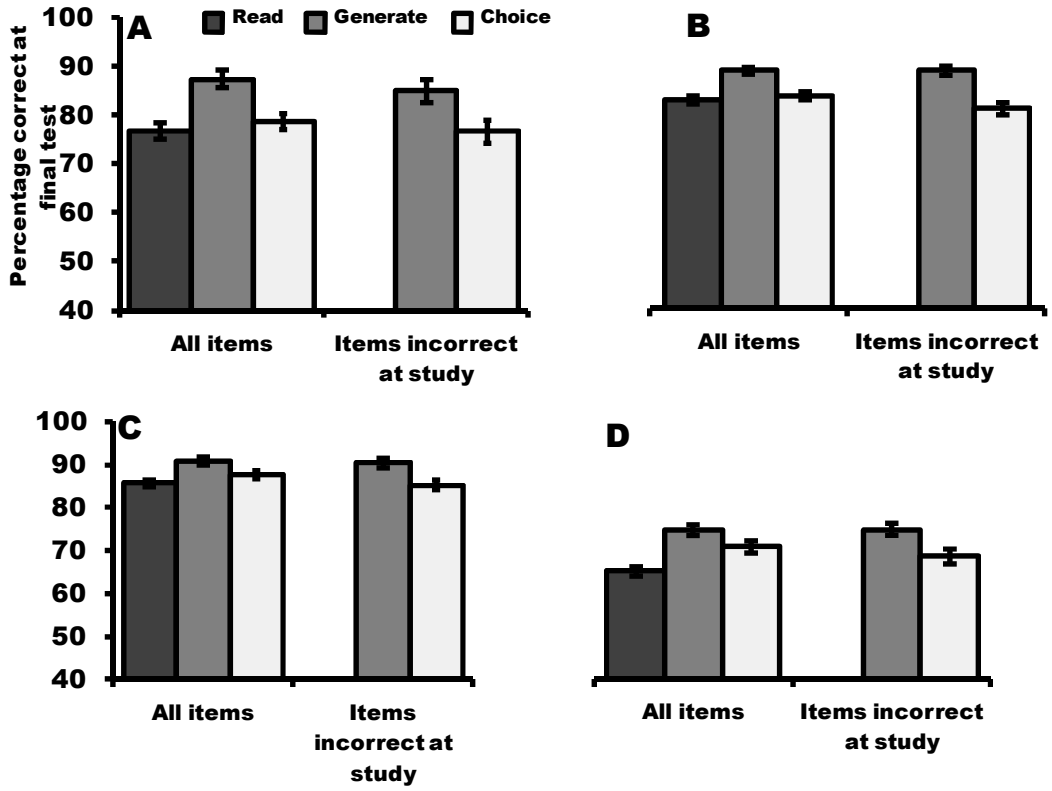

Figure 2. Mean percentage correct at final memory test in (A) Experiment 1, (B) Experiment 2A, (C) Experiment 2B, and (D) Experiment 3. Error bars indicate standard errors.

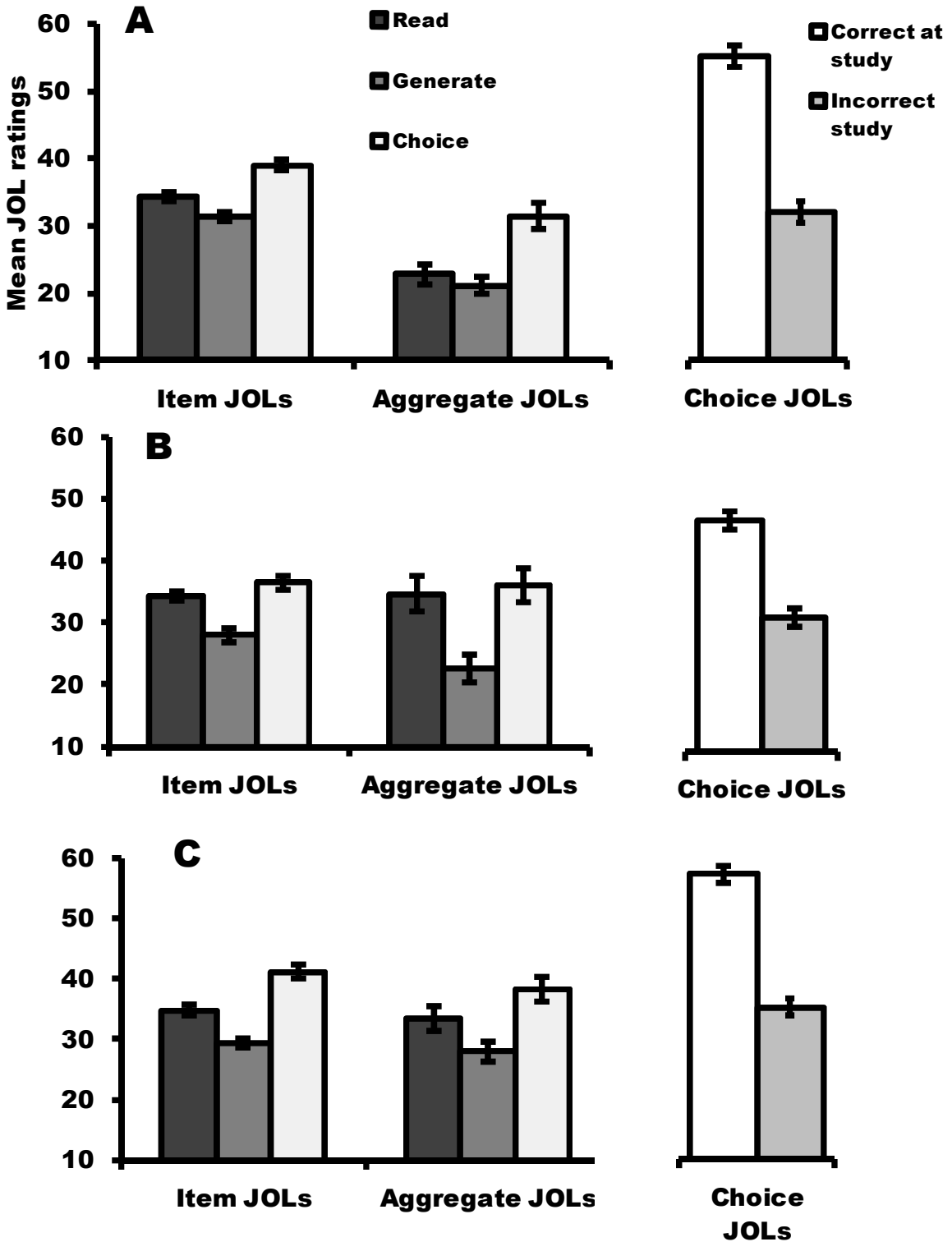

Figure 3: Mean item and aggregate judgments of learning (JOLs), and JOLs for items correct and incorrect at study, in (A) Experiment 2A, (B) Experiment 2B, and (C) Experiment 3. Error bars indicate standard errors.

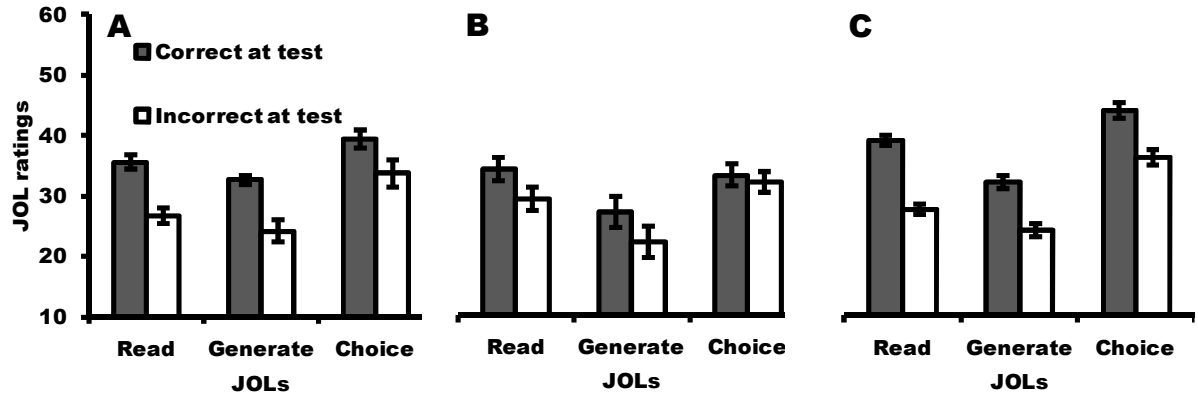

Figure 4: Mean judgments of learning (JOLs) for items correct and incorrect at test in (A) Experiment 2A, (B) Experiment 2B, and (C) Experiment 3. Error bars indicate standard errors.

Appendix A – Response times for item judgments of learning.

In Experiment 2A there was no difference in JOL response times between the three study conditions, $F(1.29, 37.37) = .22$, $p = .704$ (Greenhouse-Geisser correction applied). Mean response times were 3.51s (*SD* = 1.5) for the Read condition; 3.69s (*SD* = 2.0) for the Generate condition; and 3.59s (*SD* = 3.1) for the Choice condition. This was also the case in Exp 3, for the SP group: There was no difference between the three study methods, $F(2,46) = .43$, $p = .651$. The mean time to make JOLs was 2.54s (*SD* = .8) in the Read condition; 2.57s (*SD* = .6) in the Generate condition; and 2.49s (*SD* = .5) in the Choice condition. These data were not captured for the EP group.

In Experiment 2B, however, there was a significant difference in times to make JOLs between the three conditions, $F(2,46) = 5.85$, $p = .005$). Participants spent longer making JOLs for Generate items (*M* = 3.66s, *SD* = 1.6) than they did for Read items (*M* = 3.27s, *SD* = 1.4), $t(23) = 2.84$, $p = .009$, and longer for Generate than for Choice JOLs (*M* = 3.40s, *SD* = 1.6), $t(23) = 2.74$, $p = .012$.

Appendix B - Interpretation of JOLs data

An alternative account of our JOLs data is that, for Read items, participants anchor their judgments near the middle of the scale, indicating that they do not know whether or not they will remember the items, whereas for Generate items they are more confident. Their lower JOLs for Generate than Read items would therefore reflect higher confidence that they would not remember these items. To test this possibility, we collected data from 9 additional participants who took the standard version of the task, as in Experiment 2A, with the modification that following each JOL they gave a second rating indicating how accurate they thought their JOL was. As shown in Figure B1, we obtained a similar curve to that obtained by Dunlosky, Serra, Matvey and Rawson (2005) in that, perhaps unsurprisingly, higher confidence ratings (what Dunlosky et al. called “second order JOLs”) were given to the lowest and the highest JOLs. However, there was no difference between conditions, suggesting that participants were not using a different basis for their JOLs in the Read and the Generate conditions.

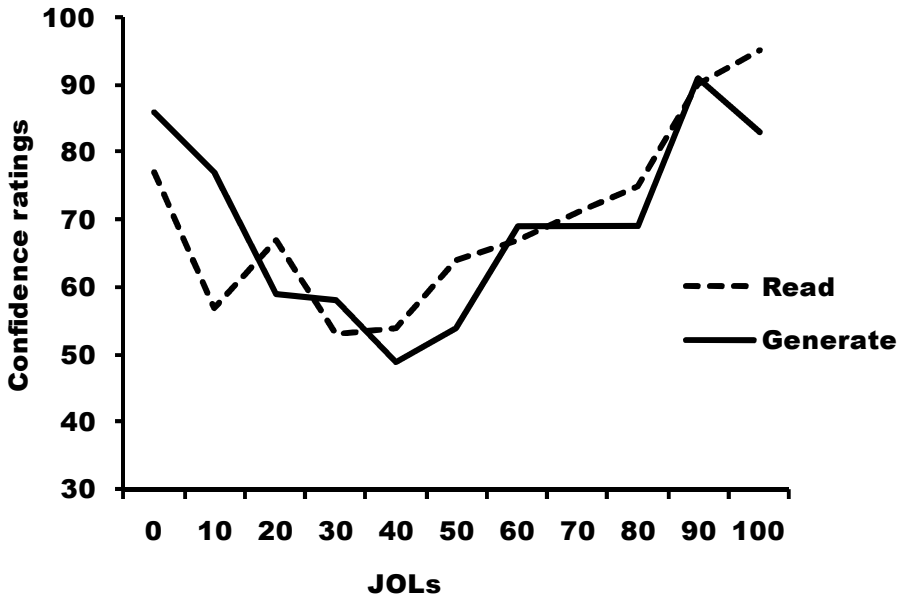

Figure B1: Mean confidence ratings as a function of judgments of learning (JOLs).

Appendix C - JOLs for items correct and incorrect at test

For completeness, we report data concerning the relationship between JOLs and test accuracy, though item selection effects mean these data should be interpreted with caution: Items which are very difficult to learn may elicit both lower JOLs and greater effort but the greater effort may not be sufficient to result in correct test performance. Figure 4 shows that, in all experiments and within each condition, JOLs for items subsequently answered correctly at test were numerically higher than for items answered incorrectly.

We carried out within-participant gamma rank correlations between JOLs and final test accuracy to examine this relationship further (see Nelson, 1984). The means are given in Table C1. In Experiment 2A the mean gamma correlation was significantly higher than zero in all conditions ($t(24) = 4.43$, $p < .001$ for Read; $t(19) = 3.50$, $p = .002$ for Generate; and $t(24) = 2.96$, $p = .007$ for Choice), showing some ability for JOLs to predict test performance. (Note that gamma cannot be computed in cases where a participant has no incorrect responses in a given condition, hence the differences in degrees of freedom.) There was no difference in resolution (i.e., accuracy at monitoring the relative recallability of items) between conditions, $F(2,36) = .112$, $p = .894$, indicating that participants were no more accurate in one condition than in another.

In Experiment 2B, the mean gamma correlation between JOLs and test accuracy did not differ significantly from zero in any condition ($t(16) = .69$, $p = .501$ for Read; $t(10) = 1.28$, $p = .230$ for Generate; $t(16) = .99$, $p = .336$ for Choice), and did not differ between conditions, $F(2,18) = .35$, $p = .710$. Resolution was generally poor with this version of the task.

In Experiment 3, the mean gamma correlation between JOLs and test accuracy was significantly higher than zero in all conditions, ($t(38) = 7.36$, $p < .001$ for Read; $t(36) = 5.31$,

$p < .001$ for Generate; and $t(37) = 2.44$, $p = .020$ for Choice), with no difference in resolution between conditions, $F(2,70) = 2.67$, $p = .076$.

Overall these results demonstrate that, within each condition, participants showed some ability to predict their true likelihood of recalling each item, although this was not statistically significant in Experiment 2B. Resolution did not differ across conditions. Of course, this within-condition relationship between memorability and JOLs is distinct from the between-conditions influence of study condition on JOLs, on which the main text focuses.

	Read	Generate	Choice
Exp 2A	.34 (.37)	.34 (.43)	.27(.45)
Exp 2B	.09 (.54)	.20 (.53)	.13 (.53)
Exp 3	.39 (.33)	.33 (.38)	.18 (.46)

Table C1: Mean (*SD*) gamma values for the correlation between JOLs and final test performance.

Appendix D - Final test performance for all participants in Experiment 3

We ran the main analysis (test accuracy) for all 107 participants in Experiment 3, excluding test trials containing partial words and nonwords. In the case of one participant this left only one usable item so this participant's data were excluded. This analysis yielded a similar pattern to that of the subset. There was a main effect of Study method, $F(2, 208) = 7.36$, $p = .001$, $\eta_p^2 = .066$, no effect of Group, $F(1,104) = 2.72$, $p = .102$, and no interaction, $F(2,208) = 2.32$, $p = .101$. Generating ($M = 75.38$, $SD = 17.6$) led to better recall than reading ($M = 67.48$, $SD = 20.66$), $t(105) = 4.18$, $p < .001$, $d = .41$. Choosing ($M = 72.79$, $SD = 21.31$) was better than reading, $t(105) = 2.45$, $p = .015$, $d = .25$, with no difference between choosing and generating, $t(105) = 1.20$, $p = .234$, $d = .13$.

Appendix E - Study time and test performance in Experiment 3

For completeness we report the relationship between study time and final test performance in Experiment 3, but these results should be treated with caution since item selection artefacts make it impossible to draw any conclusions about them. For example, participants may spend longer on items which are most difficult, but the extra time spent may not be sufficient to compensate for the difficulty and result in correct test performance (see Nelson, 1993, for a useful discussion of this point). For each study method, we compared average study times for items which were ultimately correct versus incorrect at test. A 2 (Accuracy: correct vs. incorrect) x 3 (Study method) ANOVA showed a main effect of Study method, $F(2,40) = 30.68$, $p < .001$, but no effect of Accuracy, $F(1,20) = .02$, $p = .884$, and no interaction, $F(2,40) = 1.46$, $p = .245$. Participants spent the same amount of time studying items they would later get right at test ($M = 7.58$ s, $SD = 3.77$, for Read; $M = 6.32$, $SD = 4.26$ for Generate; $M = 5.24$, $SD = 3.86$ for Choice) as they did studying items they would later get wrong ($M = 8.31$, $SD = 5.52$, for Read; $M = 6.13$, $SD = 3.54$ for Generate; $M = 4.81$, $SD = 3.50$ for Choice).