Memory interference and the benefits and costs of testing

Rosalind Potts

University College London

Submitted for the degree of Doctor of Philosophy

September 2013

Declaration

I, Rosalind Potts, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Experiments 1, 2A, 2B and 3 appear in Potts, R., and Shanks, D. R. (2013, July 1). The benefit of generating errors during learning. *Journal of Experimental Psychology: General.* Advance online publication. doi:10.1037/a0033194.

Experiment 10 appears in Potts, R., and Shanks, D. R. (2012). Can testing immunize memories against interference? *Journal of Experimental Psychology: Learning, Memory and Cognition, 38,* 1780-1785.

Some of the data collected for Experiment 9 were reported in my thesis submitted for the degree of Master of Science at University College London.

I thank Sevde Inan and Yasemin Olcusenler for assistance with data collection in Experiments 2B and 4 respectively.

Acknowledgments

It has been a privilege to carry out my doctoral work under the supervision of David Shanks, who has been an excellent mentor to me. I have learned a great deal under his expert guidance, and it has been an enormously stimulating and enjoyable experience. I am very grateful to David for his unwavering support and generosity, his incisive thinking, and the sharing of his immense knowledge and wisdom.

I’d also like to thank my second supervisor, Leun Otten, and lab members Tom Beesley, Chris Berry, Maarten Speekenbrink and Emma Ward for their helpful support and suggestions. I am grateful to the Economic and Social Research Council for funding my studies, and to the UCL Graduate School Conference Fund for financial assistance to present some of my work at the Psychonomic Society in Minneapolis last year.

Considerable thanks are due to my husband, Richard, and our daughter Sacha. Without their support, both moral and practical, my return to academia would not have been possible. This thesis is dedicated to them and to my parents, Denys and Doraine Potts.

Abstract

Testing often enhances memory, but memory can be harmed by interference from similar or competing items. This thesis examines two situations in which it has been proposed that testing can be harmful to memory because the test itself increases susceptibility to interference. Experiments 1-8 investigate the effect of generating errors during new learning. 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, or by selecting from a choice of definitions followed by feedback. In a final test of all words, generating errors followed by feedback led to significantly better memory for the correct definition 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. Metacognitive judgments of learning revealed that participants were 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.

A second series of experiments (Experiments 9-10) examines the claim that reactivating a consolidated memory destabilizes it, making it more susceptible to interference from new learning. Participants learned English-Swahili word pairs (List 1) on Day 1 with a final test on Day 3. When memory of List 1 was reactivated in the form of a reminder test immediately before learning Finnish words (List 2) on Day 2, testing, far from impairing List 1 memory, enhanced it, revealing a testing effect. Furthermore, List 2 learning disrupted List 1 memory when there was no reminder test, but reminder testing immunized the memory against interference.

TABLE OF CONTENTS

FIGURES

TABLES

CHAPTER ONE: TESTING AND INTERFERENCE

It is well established that taking a memory test can lead to enhanced memory for the tested items, a phenomenon known as the testing effect (see Roediger & Karpicke, 2006a, for a review). On the other hand, incorrect or irrelevant information can often be detrimental to memory, causing interference. This thesis examines whether the benefits of testing are sufficient to outweigh the negative effects of interference, in two situations in which it has been proposed that taking a test can be harmful to memory precisely because the test itself increases the opportunity for incorrect information to interfere with memory for correct information.

The first of these is a situation in which the test involves the production of many errors. If errors are made on a test, are these errors reinforced, making it more difficult to remember correct answers? A large body of literature advocating “errorless learning” claims that this is so, but there is also evidence contradicting this claim. In Experiments 1–8 I examine whether generating errors during learning is helpful or harmful to memory.

In the second situation, it has been proposed that reactivating memories which have already been consolidated can return them to a labile or malleable state, similar to the state they are in when they are first acquired, and that they then need to go through a reconsolidation process in order to persist in their original form. During the reconsolidation period, however, memories are particularly susceptible to interference, which may cause the original memory to be modified or even unlearned. Experiments 9 and 10 examine this claim and ask whether the well-known benefits of testing are counterbalanced by a tendency for testing to make memory more susceptible to interference.

The testing effect

A central question for educators concerns how to maximize 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" (e.g., Allen, Mahler, & Estes, 1969; Karpicke & Roediger, 2008). In a typical testing effect experiment, participants may study a list of, for example, paired associates, and then take an initial test of half of the items. Later, on a final test of all of the items, recall is often higher for the tested items than the untested ones. The effect remains when testing is compared with a restudy condition (e.g., Carpenter, Pashler, Wixted, & Vul, 2008; Cull, 2000) so it cannot be attributed to additional exposure to the materials occurring as a result of retrieval on the initial test, and it occurs when no corrective feedback is given on the initial test (Allen et al., 1969; Carpenter & DeLosh, 2005, 2006; Kuo & Hirshman, 1996; Roediger & Karpicke, 2006b), so it cannot be solely due to more efficient processing of feedback in the test condition. However, feedback may enhance the benefits of tests, both by enabling errors to be corrected and also by confirming the accuracy of correct responses (Butler, Karpicke, & Roediger, 2008; Pashler, Cepeda, Wixted, & Rohrer, 2005). 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).

Theoretical accounts of the testing effect

Most of the testing effect literature has been empirical rather than theoretical and there is no consensus as to why testing is beneficial. Among the explanations offered for the effect are that testing strengthens retrieval routes between the cue and the target (Bjork, 1975), or that it increases the retrieval strength, or accessibility, of successfully retrieved items more than studying increases the strength of studied items. The more difficult the retrieval, and the lower retrieval strength is at the time of retrieval, the greater the boost to retrieval strength (Bjork & Bjork, 1992). Bjork and colleagues (Halamish & Bjork, 2011; Kornell, Bjork, & Garcia, 2011) have also recently proposed a distribution-based model to explain why testing is sometimes beneficial to memory and sometimes not. The model assumes that testing increases memory strength more than studying does. When there is no feedback at initial test, testing results in a bifurcated distribution of items, in that a subset of tested items (those that were successfully retrieved) undergoes a large increase in strength while the strength of unretrieved tested items is left unchanged. Meanwhile, all restudied items are strengthened, but to a lesser degree than the successfully retrieved items. Since final test performance is measured in terms of the number of items recalled, rather than the strength of those items, a benefit for testing over restudying would be predicted when the final test is difficult, for example when there is a long interval between initial test/restudy and final test. This is because, when the threshold for retrieval is high, only those items highest in memory strength will exceed it and be recalled, and these will tend to be the items which were successfully retrieved at initial test and whose strength has therefore been boosted the most. When the final test is easy, for example at a short retention interval, and the threshold for retrieval is low, an advantage for restudied over tested items is predicted, since all of the restudied items have been strengthened but only a subset of the tested ones has. The model can therefore explain the test-delay interaction which has been found, whereby the testing effect sometimes emerges only after a delay when no feedback is given at initial test (e.g., Roediger & Karpicke, 2006b; Toppino & Cohen, 2009).

Another proposal is that initial testing is beneficial because it engages the same processes which are required at final test, a transfer-appropriate-processing effect. Evidence consistent with this viewpoint was found by A. K. Thomas and McDaniel (2007), but S. H.

K. Kang, McDermott, and Roediger (2007) observed a testing benefit even when initial and final tests were in different formats, suggesting that the transfer-appropriate processing account cannot be the whole story. Testing may also enhance memory by suppressing competing responses which come to mind during the initial test (R. C. Thomas & McDaniel, 2013; see also Anderson, Bjork, & Bjork, 1994). Alternatively, activation of many concepts during initial retrieval may elaborate the memory trace and create multiple retrieval routes to the target, maximizing its retrieval on a subsequent occasion (Carpenter & DeLosh, 2006; Glover, 1989; McDaniel & Masson, 1985). Some routes created during initial testing may lead to incorrect responses, so another proposal is that monitoring processes are engaged at final test to distinguish the target from these incorrect responses (R. C. Thomas & McDaniel, 2013).

The generation effect

Related to the testing effect is the generation effect (e.g., Slamecka & Graf, 1978) whereby items generated by participants in response to a cue (e.g. "opposite of hot: c___") are better remembered than the same items when they are simply read. 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 situation. 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.

Direct and indirect effects of testing

There are at least two ways in which testing can benefit memory. A direct effect of testing refers to memory enhancement that occurs as a result of taking the test itself, whether or not corrective feedback is given on the test. However, testing can also benefit memory indirectly. For example, Izawa (1970) provided evidence that a retrieval attempt (i.e., a test) may potentiate subsequent encoding of the correct response, and Bahrick and Hall (2005) proposed that even 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, 2013, for a useful discussion of this point).

Testing and interference

The testing effect and the generation effect are both highly replicated findings and suggest that there is something about the active process of recalling or generating which leads to memory enhancement for the recalled or generated items. Despite widespread evidence of the benefits of testing, however, there are situations in which it can be harmful to memory, for example when incorrect information is retrieved or generated on the test. McDermott (2006) found that testing increased false memory as well as memory for correct information, and Schooler, Foster, and Loftus (1988) found that forcing participants to choose between two alternatives, both of which were incorrect, at initial test, led to impaired memory for the studied material on a subsequent final multiple choice (MCQ) test, compared with taking no initial test.

Given that testing can enhance memory for incorrect as well as correct information, educators may be deterred from making optimal use of testing as a learning tool – for example, by setting difficult tests, which have the potential to enhance memory to a greater extent than easier ones – because of a concern that many errors will be made and that these errors will be reinforced by the act of testing, and interfere with students’ ability to remember the correct information. Certainly, there is evidence that errors are best avoided during learning. For example, Baddeley and Wilson (1994), in a typical "errorless learning" study, gave participants the first two letters of a word with many possible completions, e.g., "QU", and encouraged them to generate up to four guesses as to what the word might be, before giving them the target ("errorful" condition). They ensured that participants made at least one error for each item, by using a back-up target if the participant happened to produce the designated target on the first guess. In the "errorless" condition, the participant was given the stem, followed immediately by the correct answer. After each list of words had been presented once, the procedure was repeated twice more, so there was potential for many errors to be produced in the errorful condition. At a final free recall test, more targets were recalled in the errorless than the errorful condition.

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. Kornell, Hays and Bjork (2009), in a study which is described more fully below, found that participants remembered more weakly associated word pairs when they had generated incorrect guesses for the targets at study, before receiving corrective feedback, than when they had read the pairs intact. A worthwhile goal, then, is to identify the conditions in which errorful generation may be either helpful or harmful to subsequent retention. This thesis seeks to contribute towards achieving this goal. In this scenario, the test itself produces the erroneous material and the question is whether the act of generation, despite producing errors, can potentiate the encoding of subsequent feedback sufficiently to outweigh any negative effect of the errors.

The perpetuation of errors on a test is not the only way that testing could be harmful to memory. Although memory for the tested material is typically enhanced by testing there is evidence that it can be modified in other ways. Research in neurobiology, mainly involving non-human animals, has suggested that consolidated memories become "labilized" (made fragile) on reactivation and need to be reconsolidated (see Hardt, Einarsson, & Nader, 2010, for a review). Indeed, it may be through a reconsolidation process that memories become strengthened, leading to better subsequent recall (Finn & Roediger, 2011, Sara, 2000). However, during the reconsolidation period memories are particularly susceptible to interference which, far from strengthening them, may cause them to undergo modification or even unlearning.

So far few studies have investigated whether reconsolidation occurs in humans. Rubin (1976) found that electroconvulsive shock (ECS) could cure patients of their compulsive thoughts. ECS was only successful if applied when patients were focusing on their compulsions, and not when they were anaesthetised, consistent with the notion that reactivating these memories made them labile and susceptible to disruption by ECS. More recently, a handful of studies have investigated the issue of reconsolidation in human memory using new learning as the interfering agent (e.g., Walker, Brakefield, Hobson, & Stickgold, 2003; Forcato, Burgos, Argibay, Molina, Pedreira, & Maldonado, 2007; Forcato, Argibay, Pedreira, & Maldonado, 2009; Hupbach, Gomez, Hardt, & Nadel, 2007, Hupbach, Hardt, Gomez, & Nadel, 2008; Hupbach, Gomez, & Nadel, 2009). It is well-established, from studies of retroactive interference, that new learning can disrupt memory for recently acquired information (e.g., Barnes & Underwood, 1959). If reactivating a memory returns it to a labile state similar to when it was first acquired, will it again be sensitive to disruption from new learning? In this scenario, it is proposed that information learned *after* a test can disrupt memory for the tested material itself. This literature is reviewed more thoroughly in Chapter 8, so the remainder of this chapter will focus on the issue of error generation.

The effect of making errors during learning

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. For example, when asked for the capital of Canada, they may know that the correct answer is “Ottawa” but mistakenly respond with “Montreal”. In this case there is a pre-existing association between the question and the correct answer (*Canada - Ottawa*) at the time of initial retrieval, and corrective feedback may be used to reinforce this association, to maximize the chance of successful retrieval on future occasions. In this thesis this scenario is referred to as “unsuccessful retrieval” (following Kornell, Hays, & Bjork, 2009) and it has been the focus of several recent studies (Grimaldi & Karpicke, 2012; Hays, Kornell, & Bjork, 2013; Huelser & Metcalfe, 2012; Knight, 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. For example, they may be asked for the capital of Mali, but never have come across that piece of information before. 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 (e.g., “Jamestown” for the capital of Mali). In this case, to learn the correct answer, the individual has to learn a novel association between the unfamiliar cue material (Mali) and the corrective feedback provided (Bamako) – there is no pre-existing association to be reinforced. This scenario, which will be referred to as “errorful generation”, has received less attention and is the focus of Experiments 1-8 of this thesis. These experiments 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.

Is generating errors during learning helpful or harmful to memory?

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 be harmful to subsequent memory. In a typical errorless learning study, as noted above, a condition in which participants are encouraged to generate many erroneous responses to a test cue leads to worse subsequent memory performance than a condition in which they are presented with the correct answer intact. 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).

Unsuccessful retrieval

Despite these claims, there is also evidence that even tests which yield errors can benefit later retention, as long as corrective feedback is given. A study by Kane and Anderson (1978) found a benefit of generating errors over reading when participants were instructed to generate 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. 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 et al., 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.

The semantic relatedness hypothesis

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*). This covert retrieval or activation 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 as part of the search set, 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 activated at the same time. Since retrieval strengthens memory more than reading, and feedback enhances the benefits of testing (Butler & Roediger, 2008; Pashler et al., 2005), this led to an advantage for targets studied in the Generate condition. A related proposal was made by Slamecka and Fevreiski (1983) who suggested that generation could be beneficial even when participants failed to produce the target at initial study, if they had partially retrieved it by activating some of its semantic attributes but not its surface features. This study is described in more detail in Chapter 3.

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 also 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

The focus of Experiments 1-8 of this thesis is the related but 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 the 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.

Why might errorful generation benefit memory?

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.

Overview of thesis

Experiments 1-8 of this thesis examine 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 these experiments are very obscure English 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?

Chapter 1 establishes the basic procedure used for the errorful generation experiments described in Chapters 2 to 7. Participants learned definitions of rare English words, e.g., *roke – mist* (or translations of foreign language words in Experiments 2B and 8), each in one of three different ways. Previous studies examining the effect of error generation (e.g., Grimaldi & Karpicke, 2012; Hays, Kornell & Bjork, 2012; Huelser & Metcalfe, 2012; Kane &

Anderson, 1978; Kornell, Hays, & Bjork, 2009; Slamecka & Fevreiski, 1983) have typically compared a condition in which a cue and target are presented together for study (*Read condition*) with a condition in which the cue alone is presented and the participant's task is to guess the target before being shown the correct answer as feedback (*Generate condition*). Subsequently, participants take a final test of all the items. Our experiments also used Read and Generate conditions, and we added a novel third condition, in which participants were given a choice of definitions at study, from which they were to select the one they thought was correct (*Choice*).

Experiment 1, described in Chapter 2, demonstrates an advantage for generating errors over both reading and choosing during the learning of unusual English words, even though the correct answers are displayed for a much shorter time in the Generate condition. In Chapter 3 the effect is replicated both with unusual English words (Experiment 2A) and with foreign vocabulary (Experiment 2B) as stimuli. In addition, participants' metacognitive judgments of learning, and their relationship to test performance, are investigated and reveal a striking dissociation between participants' predictions and their actual test performance. Chapter 4 (Experiments 3 and 4) explores possible explanations for these findings, and replicates the effects of the previous experiments using a modified procedure in which participants could study correct answers for as long as they chose.

Chapter 5 describes an experiment (Experiment 5) which investigates the effect of different lures in the Choice condition, at study and at test. Chapter 6 (Experiments 6 and 7) asks whether the Generate advantage extends to a cued recall final test format, and examines more closely the effect, on final test performance, of errors generated at study. Chapter 7 looks at the differences between the procedure used in our experiments and that used in typical "errorless learning" studies in a quest to reconcile their divergent findings, and uses a modified procedure (Experiment 8) to investigate the effect of generating errors across multiple cycles of learning.

Chapter 8 (Experiments 9 and 10) examines the proposal that testing can be detrimental to memory because reactivation of a consolidated memory can make it more susceptible to interference, which can disrupt reconsolidation. We conclude that testing and interference exert opposite effects on memory but that testing, far from making memory more vulnerable to interference, has a greater benefit in the presence of interference than in its absence. Finally, Chapter 9 summarises and discusses the main findings and highlights avenues for future research.

CHAPTER TWO: A BENEFIT OF GENERATING ERRORS DURING THE LEARNING OF NOVEL INFORMATION

Our first experiment addresses the question of whether an advantage for generating over reading can be observed when the cues are novel vocabulary items. In this scenario there is no pre-existing association in participants' minds between cue and target because the words have never been previously encountered, and little or no opportunity to elaborate the cue in a meaningful way. If a benefit for errorful generation is observed in these circumstances, this would challenge the claim that such a benefit can only be observed when there is a pre-existing relationship between cue and target (Grimaldi & Karpicke, 2012; Hays, Kornell, & Bjork, 2012; Huelser & Metcalfe, 2012).

In Experiment 1 participants learned definitions of rare English words (e.g., *roke – mist*), using a computerised task in which words and their definitions were displayed on screen and participants typed their responses into the computer. There were three within subjects conditions: the Generate and Read conditions used in previous studies, and a novel "Choice" condition, in which participants were shown the cue and two possible definitions, one of which was the target, and they had to type in the one they thought was correct. Thus, participants learned definitions either by reading the word with its definition (*Read*), by guessing a definition followed by corrective feedback (*Generate*), or by choosing from two definitions followed by feedback (*Choice*). 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. With this design, exposure to the correct answer is longer for Read than for Generate items, since the target definition is on screen for the total trial time in the Read condition, but only the last few seconds of the trial time in the Generate condition. This disparity could disadvantage Generate items by comparison with Read items. However, we used this design because we were interested in determining which study method makes optimal use of study time. A benefit of generating errors over reading has been demonstrated even when total trial time was equated, with related paired associates (Hays, Kornell, & Bjork, 2012; Huelser & Metcalfe, 2012; Kornell, Hays, & Bjork, 2009; Slamecka & Fevreiski, 1983) and with both determined and undetermined sentence completions (Kane & Anderson, 1978). If the same benefit is observed for the learning of novel vocabulary, where there is no pre-existing relationship between cue and target in participants' minds, and no opportunity to elaborate the cue before exposure to the correct answer, this advantage will be all the more striking given that the correct answer is available for a shorter time in the Generate condition.

After studying 60 words and their definitions under these 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.

Because we were interested in the effect of making errors, we firmly encouraged participants to guess at study 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.

Experiment 1

The aim of Experiment 1 was to examine 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.

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

The design was within-subjects with one independent variable (Study Method) with three levels (Read, Generate, and Choice). The dependent variable was the number of items correctly selected at the final multiple choice test in each condition.

Materials

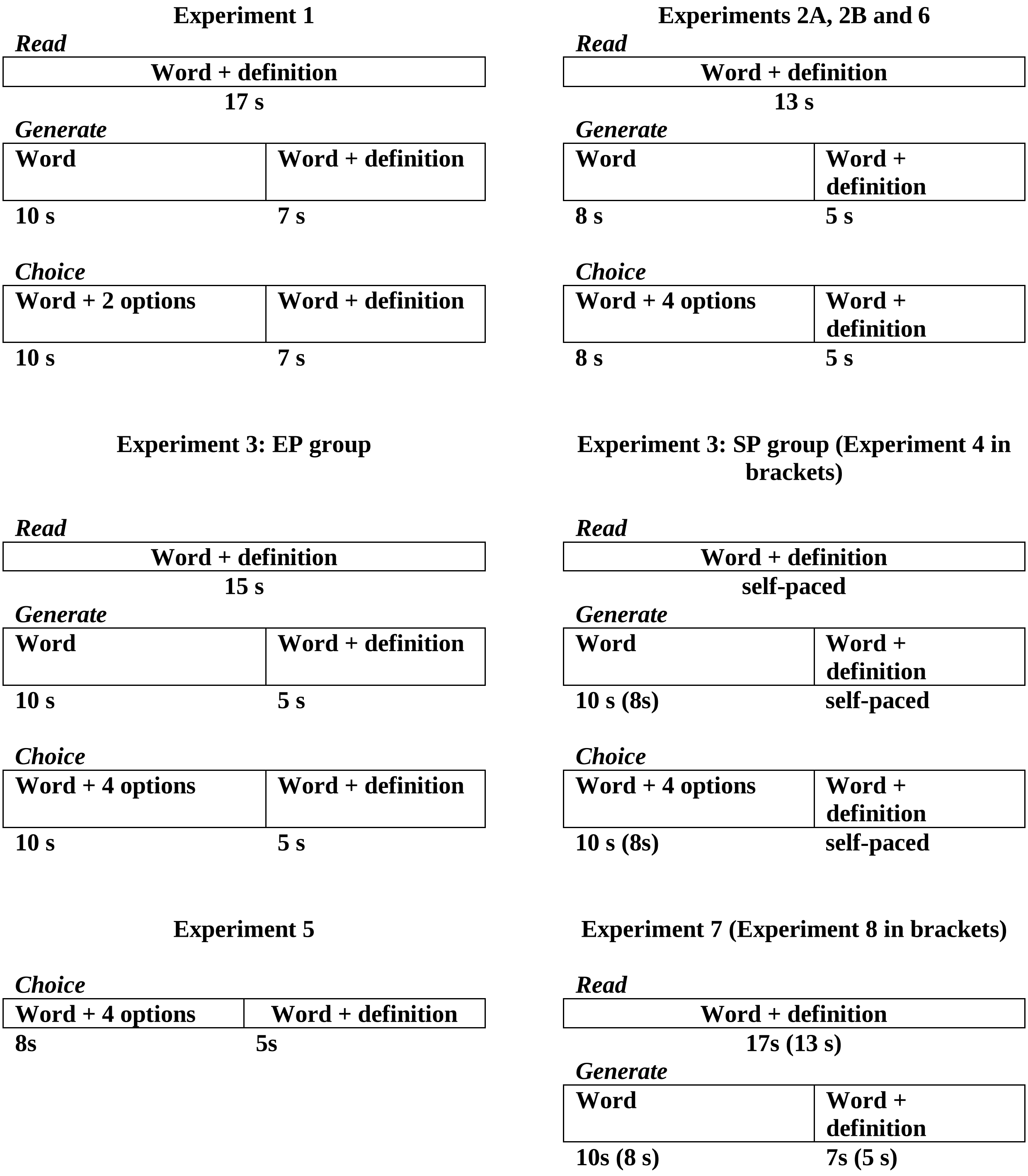
For the stimulus materials we created a pool of very unusual English words, each paired with a one-word definition e.g., *hispid - bristly, valinch - tube, frampold – quarrelsome,* from which we selected 60 pairs which were unfamiliar to participants in a pilot study. Appendix A shows the stimulus materials for all experiments. See Table A1 for the word pairs used in Experiment 1. 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. The order of the two choices on the screen was randomly determined.

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 at study in Experiments 1-8.

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. There were no time constraints in this phase of the experiment.



Experiment 1

Read

Word + definition

17 s

Generate

Word | Word + definition

10 s | 7 s

Choice

Word + 2 options | Word + definition

10 s | 7 s

Experiments 2A, 2B and 6

Read

Word + definition

13 s

Generate

Word | Word + definition

8 s | 5 s

Choice

Word + 4 options | Word + definition

8 s | 5 s

Experiment 3: EP group

Read

Word + definition

15 s

Generate

Word | Word + definition

10 s | 5 s

Choice

Word + 4 options | Word + definition

10 s | 5 s

Experiment 3: SP group (Experiment 4 in brackets)

Read

Word + definition

self-paced

Generate

Word | Word + definition

10 s (8s) | self-paced

Choice

Word + 4 options | Word + definition

10 s (8s) | self-paced

Experiment 5

Choice

Word + 4 options | Word + definition

8s | 5s

Experiment 7 (Experiment 8 in brackets)

Read

Word + definition

17s (13 s)

Generate

Word | Word + definition

10s (8 s) | 7s (5 s)

Figure 1. Study trial procedure for Experiments 1- 8. EP = Experimenter-paced, SP = Self-paced.

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, η_p^2 = .25, (Figure 2). 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 2). This analysis revealed a similar pattern, *F* (2,46) = 3.33, *p* = .044, η_p^2 = .13. Generate scores were higher than Read scores, *t*(23) = 2.40, *p* = .025, *d* = .35, 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.

[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.

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.

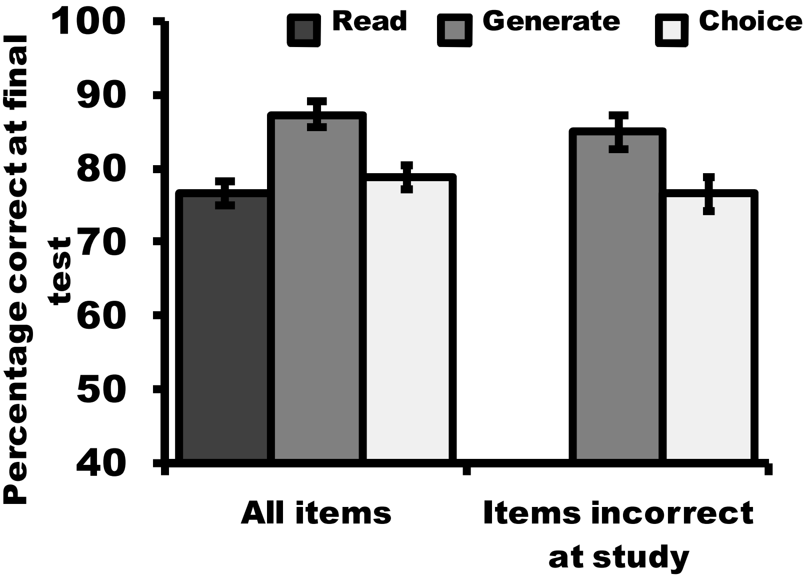


Figure 2: Mean percentage correct at final memory test in Experiment 1. Error bars indicate standard errors.

Discussion

Experiment 1 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.

Why is generation beneficial, even when many errors are generated, and even when there is no pre-existing relationship between cue and target to reinforce? One possibility is that, in the Generate condition, the effort involved in generating a response, together with the experience of finding that that response is (nearly) always incorrect, leads participants to perceive Generate items as more difficult to learn than Read or Choice items. If this is the case, they may apply more effort or attention to the encoding of corrective feedback in the Generate condition. Chapter 3 examines the role of metacognition in the errorful generation task.

CHAPTER THREE: METACOGNITIVE JUDGMENTS IN THE ERRORFUL GENERATION TASK.

The finding, in our first experiment, that generating led to better memory than reading even when cues were previously unfamiliar words is surprising in light of the proposal that an advantage for errorful generation can only be observed when there is a pre-existing association between cue and target. In our second experiment, therefore, the first aim was to replicate the errorful generation benefit observed in Experiment 1 for the learning of unusual English words (Experiment 2A), and to examine whether it extended to the learning of foreign language vocabulary (Experiment 2B).

Although most of the words used in Experiment 1 were unfamiliar to most participants pre-experimentally, and participants were at chance on the Choice test, a few words were correctly generated, indicating some prior knowledge. In Experiment 2A and subsequent experiments we replaced items for which the response had been correctly generated in Experiment 1. We also attempted to replicate the finding using rare foreign language vocabulary which would be highly unlikely to be familiar to participants. In Experiment 2B, therefore, the stimuli used were words from Euskara, the language of the Basque country in Northern Spain. We chose Euskara because it is a “language isolate”, a language with no known relations.

In addition to seeking to replicate our previous findings, a further aim of Experiment 2 was to investigate participants’ metacognitive awareness of their learning. 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 and 2B, we 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). Generating errors before encoding correct information may lead to corrective feedback being processed less fluently, which in turn may lead participants to give lower JOLs to Generate items. Processing of feedback may be less fluent when errors are generated 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 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 were interested in 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.

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 them. 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 as observed in Experiment 1.

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 percentage 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.

Experiments 2A and 2B

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 (see Table A2 in Appendix A). For Experiment 2B we selected 60 Euskara nouns with their English translations (e.g., *igel* - *frog*, *urmael* - *pond*, *untxi* - *rabbit*). The full set is shown in Table A3. 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 et al., 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 B. 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).

[2] Somewhat surprisingly, and unlike in Experiment 1, 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

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 3A).

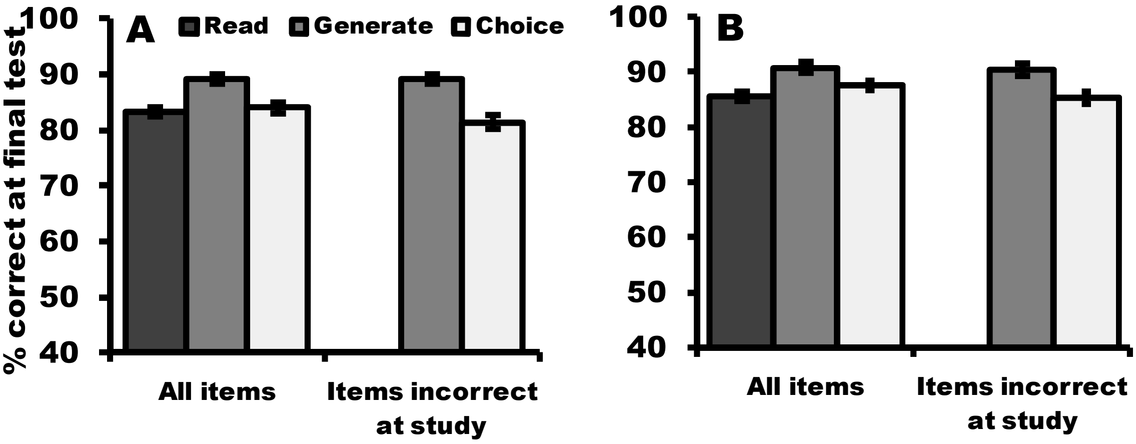


Figure 3: Mean percentage correct at final memory test in (A) Experiment 2A, (B) Experiment 2B. Error bars indicate standard errors.

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

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.

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] ($t(23) = 2.11$, $p = .046$).

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 3B). 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 4A), $\eta_p^2 = .42$. 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.

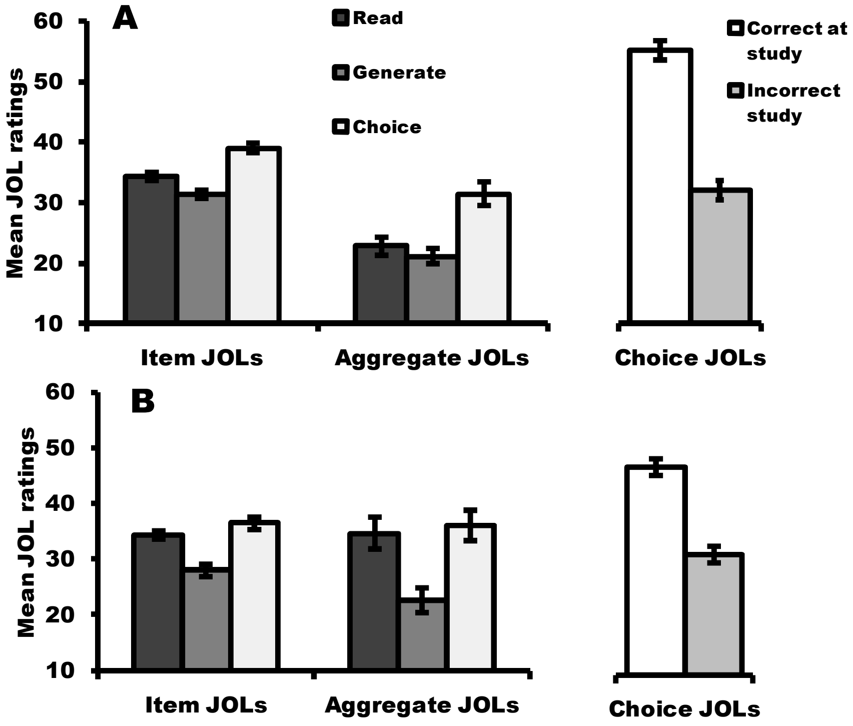


Figure 4: Mean item and aggregate judgments of learning (JOLs), and JOLs for items correct and incorrect at study, in (A) Experiment 2A, and (B) Experiment 2B. Error bars indicate standard errors.

Aggregate JOLs showed a largely similar pattern (Figure 4A). 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) = 7.77, p = .002, η_p^2 = .21). Choice JOLs were again higher than both Read (t(29) = 2.60, 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. 4A), 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 suggests 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 C for an alternative possibility.)

Appendix D reports the relationship between JOLs and test performance. Although JOLs showed some ability to predict final test scores (Figure 5A), 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 4B). 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 4B). 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 4B). 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 5B). See Appendix D for analysis of these data.

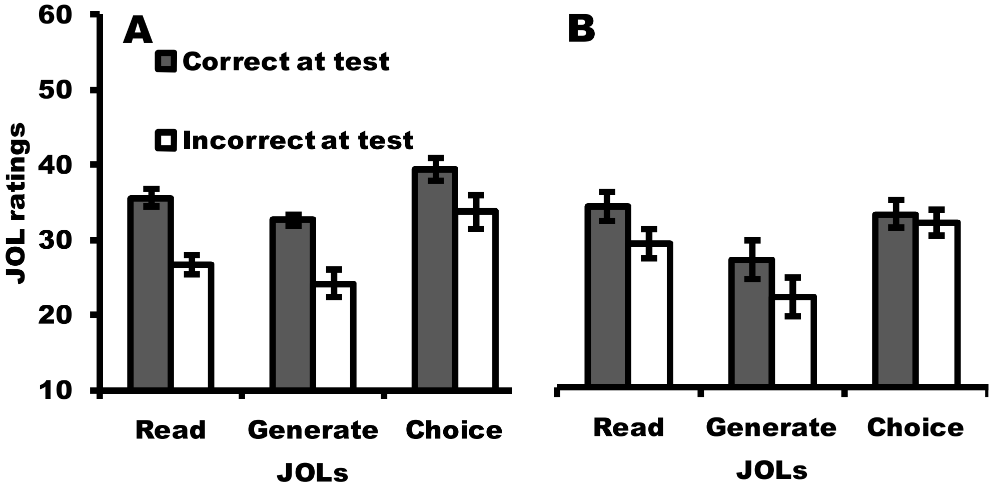


Figure 5: Mean judgments of learning (JOLs) for items correct and incorrect at test in (A) Experiment 2A, and (B) Experiment 2B. Error bars indicate standard errors.

Discussion

Experiment 2A replicated the benefit of generating over reading for unusual English words that we observed in Experiment 1, and Experiment 2B showed that this benefit 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. The findings of Experiment 2 are inconsistent with this proposal. 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 were described in Chapter 1 (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.

Is it possible that our participants in Experiment 2A had some knowledge of the definitions prior to the experiment? The cues were deliberately chosen to be very obscure words which were extremely unlikely to be known to participants pre-experimentally. However, perhaps our participants were avid crossword-solvers with a passion for Scottish dialect and archaic English. When participants generate an incorrect response, or no response, at study, we cannot conclude that they had no prior knowledge of that word. Slamecka and Fevreiski (1983) reported a generation benefit even when participants had failed to generate items at study. They proposed that, when generation failed, it was often the case that participants had in fact retrieved some semantic attributes of the correct answer but had been unable to retrieve its surface features. When these surface features were supplied, in the form of corrective feedback at study, this facilitated memory at the final test, both for recall (Experiment 1) and recognition (Experiment 2). This conclusion was supported by their third experiment which showed that, even when no feedback was given at study, performance on a forced-choice recognition test of the response items was above chance for items which had not been successfully generated during the study phase, suggesting that these items had in fact been partially retrieved.

Could our participants' failure to generate correct responses at study therefore reflect an inability to produce them in the time available, rather than a lack of knowledge, in line with Slamecka and Fevreiski's proposal? 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). We compared this with the average similarity of the targets to all of the other targets, as an example of a randomly unrelated set of items (*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. Together with the finding that the benefit of generating was replicated when participants learned previously unfamiliar foreign vocabulary, it seems unlikely that the Generate advantage is an artefact of prior familiarity with the cue words.

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.

CHAPTER FOUR: SELF-PACING OF STUDY TIME AND THE OPPORTUNITY TO SELECT GENERATED GUESSES AT TEST

The preceding experiments have demonstrated an advantage for generating over reading and choosing. Strikingly, Experiments 2A and 2B showed that participants failed to predict this benefit, judging Generate items as less likely to be remembered than either Read or Choice items when in fact they were better remembered. These findings suggest that there is a powerful benefit of errorful generation over reading and choosing, but that participants fail to predict this because they perceive the Generate condition as more difficult than the Read and Choice conditions and their JOLs are a reflection of this perception. However, there are alternative explanations which need to be eliminated. The two experiments in this chapter address three issues which may affect interpretation of these results: the reason why participants give lower JOLs to Generate items; the possibility that Generate items are better remembered because participants rehearse them during time allocated to the study of Read items; and the nature of the lures in the final multiple choice test.

Why do participants give lower JOLs to Generate items?

Both Experiments 2A and 2B suggested that participants' JOLs were strongly influenced by an event preceding study of the correct answer. In Experiment 2, participants gave lower JOLs to Generate items than to either Read or Choice items, consistent with our hypothesis that they experience the Generate condition as more difficult than the other study methods. However, participants had only 5 seconds in which to process corrective feedback in the Generate condition, but 13 seconds to process correct answers in the Read condition. One possible explanation for the higher JOLs given to Read than Generate items, therefore, 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. In Experiments 3 and 4, we allowed participants to study correct definitions for as long as they chose. If the same pattern of JOLs is observed when participants are free to study targets for as long as they choose, this will 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.

Displaced rehearsal of Generate items?

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. Slamecka and Katsaiti, (1987) using a traditional generation effect paradigm in which participants either generated synonyms from cues, e.g., "*lawyer*" from "*attorney – l_____*", or read the cue-synonym pair intact ("*attorney – lawyer"),* proposed that, when Read and Generate items are included in the same list, participants find Generate items harder and consequently engage in "displaced rehearsal" of Generate items by continuing to rehearse them while Read items are being presented. They found that the advantage for the generated targets at a final free recall test could be eliminated by requiring participants to rehearse the current item only, and concluded that the generation effect observed in previous studies was an artefact of using a mixed list design. Although this interpretation was challenged by subsequent researchers as an account of the classic generation effect (e.g., Hirshman & Bjork, 1988; McDaniel, Riegler, & Waddill, 1990), it remains a plausible alternative explanation for the benefit of errorful generation observed in our experiments.

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. In Experiment 4 processing of correct answers was self-paced for all participants.

Opportunity to select, at final test, a response generated at study

In Experiments 1, 2A and 2B, 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. A further aim of Experiment 3 was to examine whether the advantage for generating over both reading and choosing observed in our earlier 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 Experiments 1, 2A, and 2B, 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 Experiments 3 and 4 we equated familiarity of the lures between the three conditions. In Experiment 3 participants had the option to select, at test, their own generated guess as well as incorrect responses made in the Choice condition. If the only reason for the benefit of generating over choosing in the earlier experiments was because of the presence of interfering study lures at test in the Choice condition but not in the Generate condition, we would expect to see the Generate benefit eliminated in Experiment 3.

In Experiment 4 the test lures were all definitions for other items, so there was no possibility of selecting either an incorrectly generated or an incorrectly chosen response. In this case, eliminating the possibility of interference from Choice study lures may allow an advantage for choosing over reading to emerge. If incorrect choosing is just as beneficial as generating under these conditions, this would suggest that there is nothing special about generating: both generating and choosing can lead to a benefit over reading.

Experiment 3

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 maximize 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*). See Table A4 for the full set of items.

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.

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

		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*

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

An important 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) x 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, $F(2,76) = 1.57$, $p = .216$. Table 2 shows the means for each group and Figure 6 shows the means collapsed across groups. 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, $t(39) = 1.75$, $p = .089$, $d = .24$.

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

	Final test performance		Item JOLs		Aggregate JOLs	
	SP	EP	SP	EP	SP	EP
Read						
M	62.84	69.06	34.35	35.45	28.71	40.31
SD	18.07	19.75	15.03	14.30	15.90	17.27
Generate						
M	73.65	77.31	30.25	27.77	26.75	29.81
SD	17.70	12.84	13.62	12.77	17.61	14.21
Choice (All)						
M	71.40	69.89	44.29	36.51	40.83	34.25
SD	18.16	22.99	15.82	16.65	21.09	20.82
Choice (incorrect at study)						
M	69.84	67.34				
SD	21.85	25.35				

Next we repeated the analysis for just those items incorrect at study (Figure 6). The assumption of sphericity was not met, so the Huynh-Feldt correction was applied. There was a main effect of Study Method, $F(1.70, 64.44) = 7.35$, $p = .002$, $\eta_p^2 = .16$, no main effect of Group, $F(1,38) = .19$, $p = .662$, and no interaction, $F(1.70, 64.44) = 1.53$, $p = .227$. 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.

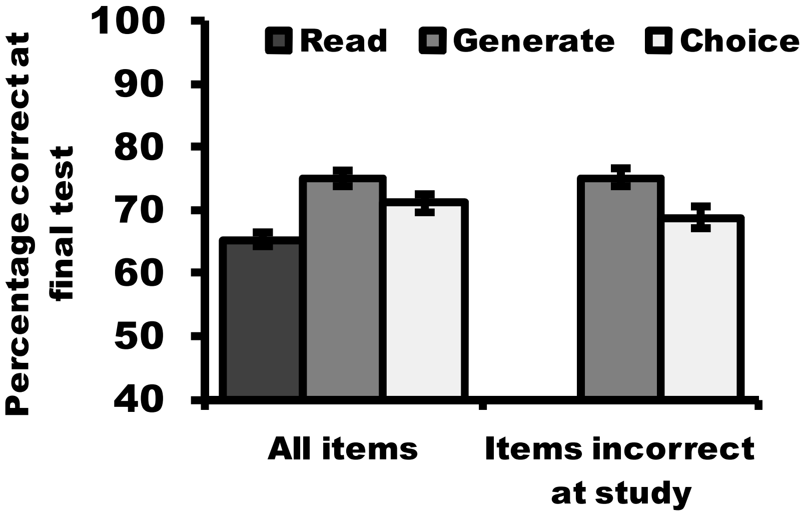


Figure 6: Final test performance in Experiment 3 (both groups combined). Error bars indicate standard errors.

Effect of interference from errors generated and chosen at study

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 E.

Judgments of learning

Did participants’ JOLs show any awareness of the benefits of generating over reading and choosing? A 3 (Study Method) x 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 7.

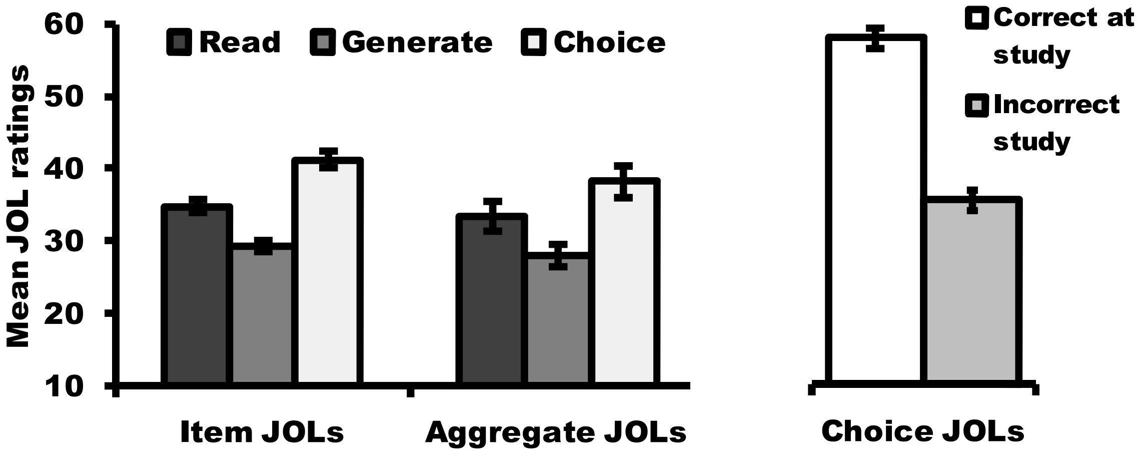


Figure 7: Mean item and aggregate judgments of learning (JOLs), and JOLs for items correct and incorrect at study, in Experiment 3. Error bars indicate standard errors.

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 = .29, 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, $t(15) = 4.25$, $p = .001$, $d = .57$, 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 7). 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, $p = .469$, $\eta_p^2 = .01$, but again there was an interaction between Study Method and Group, $F(1.62, 61.54) = 3.78$, $p = .037$, $\eta_p^2 = .09$. 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, $t(15) = 4.25$, $p = .001$, $d = .57$, and JOLs for Read items were higher than those for Choice items incorrect at study, $t(15) = 2.66$, $p = .018$, $d = .39$.

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 7). 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* = 77.8, *SD* = 26.2) was not significantly greater than the percentage of items incorrect at study which were converted to being correct at test (*M* = 73.1, *SD* = 23.0), *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 D. Figure 8 shows the means.

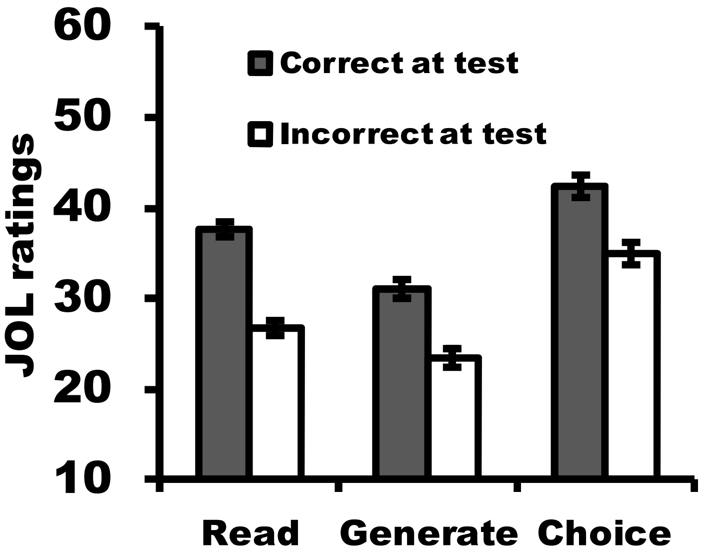


Figure 8: Mean judgments of learning (JOLs) for items correct and incorrect at test in Experiment 3.

For the aggregate JOLs there was a main effect of Study Method, *F*(2,76) = 4.10, *p* = .020, η_p^2 = .10 (Figure 7). 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, η_p^2 = .01, but there was an interaction between Study Method and Group, *F*(2,76) = 3.81, *p* = .027, η_p^2 = .09. The EP group gave higher aggregate JOLs to the Read condition (*M* = 40.3, *SD* = 17.3) than to the Choice (*M* = 34.3, *SD* = 20.8) or Generate (*M* = 29.8, *SD* = 14.2) conditions but the difference between the study methods fell short of significance, *F*(2,30) = 2.78, *p* = .078, η_p^2 = .16. For the SP group, the difference was significant, *F*(2, 46) = 5.80, *p* = .006, η_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? In the Read condition, “study time” of correct definitions is the total trial time, whereas for Generate and Choice items it is just that part of the trial in which corrective feedback is given, it does not include the first 10s of the trial which are spent generating or choosing responses. There was a significant difference between conditions in study time, *F*(2, 46) = 35.44, *p* < .001, η_p^2 = .61. Participants spent longer studying definitions for Read items (*M* = 8.02 s, *SD* = 4.35) than for both Generate (*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 F 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 final 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 in Experiment 2, 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.

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. Indeed, they were very good at rejecting their own incorrect generations, choosing these at test at a rate significantly below chance. 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.

Experiment 4

In Experiment 3, study of correct definitions was self-paced and participants had the opportunity to select, at test, responses they had themselves generated or chosen at study. In these conditions, the findings of our previous experiments were replicated: generating led to better memory than reading and incorrect choosing, and participants continued to give lower JOLs to Generate items than to Read or Choice items.

In Experiment 4 we again allowed participants to study correct answers for as long as they chose. However, the final multiple choice test equated familiarity of the lures between the three conditions in a different way. There were four final test options, the target and three lures. The three lures were all true definitions for other items from the experiment, one from each of the different study methods. Thus, all test lures were definitions which had previously been seen in the study phase, and there was no option to select either an incorrectly generated response or an incorrect choice made at study. In Experiment 3, the benefit of generating over incorrect choosing persisted in spite of the opportunity for interference from participants' own generated responses as well as from incorrect choices. Participants were better at recognising and rejecting their own generated responses than their own incorrect choices, which may have been responsible for this difference. In Experiment 4, with both those sources of interference eliminated, the benefit of generating over incorrect choosing might also be eliminated. Of course we still expected to see a benefit of generating over reading, as in previous experiments.

Method

Participants

There were 30 participants, 13 male, average age 29.2 (*SD* =10.78).

Materials

Stimuli were the same 60 word-definition pairs as in Experiment 3, with the same study lures for the Choice condition. Lures in the test phase were all true definitions from the study phase of the experiment, one from each of the three study methods (Table A5). Thus, each definition appeared three times in the final test, once as the correct target and twice as lure definitions for other items. Counterbalancing was as in the previous experiments.

Procedure

The procedure was identical to that for the self-paced group in Experiment 3 except that Generate and Choice cues were displayed for 8s rather than 10s. There was no experimenter-paced group. As in Experiment 3, participants clicked on a button labelled "Finished studying" when they had studied the definitions for as long as they wished. The final multiple choice test options consisted of the correct answer and three lures which were true definitions from other items in the experiment, one from each of the different study methods. Thus, there was potential for interference from lures familiar from the study phase, but not from participants' own errors and not from lures available for selection at study in the Choice condition.

Results

At study no Generate responses were correct. Correct responses to Choice items (M = 29.3, SD = 11.6) were above chance, $t(29) = 2.05$, $p = .049$[3].

Final test performance

Replicating the findings of previous experiments, when all items were considered, there was a main effect of Study Method, $F(2,58) = 4.19$, $p < .020$, $\eta_p^2 = .126$. Generating (M = 59.3, SD = 21.8) was superior to reading (M = 50.7, SD = 26.3), $t(29) = 2.90$, $p < .007$, d = .36. Choosing (M = 56.7, SD = 23.9) was also superior to reading ($t(29) = 2.09$, $p = .045$, d = .24), while the Generate and Choice conditions did not differ (t (29) = .80, $p = .430$, d = 0.12). Twenty participants remembered more Generate than Read items, while eight showed the opposite pattern. Seventeen participants remembered more Choice than Read items, with

[3] On removal of the data of one participant whose Choice score at study was more than two standard deviations above the mean and who correctly selected more than half of the items at study, correct responses to Choice items (M = 28.4, SD = 10.7) were no longer above chance, $t(28) = 1.74$, p = .094. We re-ran the main analyses without this participant's data. There was a main effect of Study Method, $F(2,56) = 4.56$, $p < .015$, $\eta_p^2 = .140$. Generating (M = 59.1, SD = 22.2) was superior to reading (M = 49.8, SD = 26.3), $t(28) = 3.09$, $p < .005$, $d = 0.38$. Choosing (M = 56.0, SD = 24.1) was also superior to reading, $t(28) = 2.09$, $p = .045$, d = 0.25, while the Generate and Choice conditions did not differ, t (29) = .91, $p = .371$, d = 0.13. These effects are similar to those in the analysis of all participants. However, when only items incorrect at study were included, the difference between study methods now became significant, $F(2,56) = 3.40$, $p = .040$, $\eta_p^2 = .108$. Since there were no incorrect generations at study, the advantage of generating over reading was as above. There was no difference between incorrect choosing (M = 56.1, SD = 25.7) and reading, $t(28) = 1.68$, $p = .104$, d = 0.24), nor between incorrect choosing and generating, $t(28) = .74$, $p = .465$, d = .12.

seven showing an advantage for reading. When the analysis was re-run excluding Choice items which were correct at study (but with all Read and Generate items included, since there were no Generate items which were correct at study), the difference between study methods fell just short of significance, $F(2,58) = 3.09$, $p = .053$, $\eta_p^2 = .010$ (but see footnote 3). The mean score for Choice items incorrect at study was 56.46 (*SD* = 25.33), falling between the Read and Generate scores and not significantly different from either.

Judgments of learning

Just as in previous experiments, an ANOVA conducted on the judgments of learning showed a main effect of Study Method, $F(2,58) = 14.92$, $p < .001$, $\eta_p^2 = .34$. Despite having as long as they chose to study correct feedback, participants gave lower JOLs to Generate items (*M* = 32.1, *SD* = 18.0) than they did to either Read items (*M* = 38.2, *SD* = 20.5), $t(29) = 3.6$, $p = .001$, or Choice items (*M* = 40.8, *SD* = 21.6), $t(29) = 5.1$, $p < .001$. There was no difference between Read and Choice JOLs, $t(29) = 1.71$, $p = .097$.

Also replicating the findings of previous experiments, JOLs for Choice items correct at study (*M* = 56.0, *SD* = 26.0) were significantly higher than JOLs for Choice items incorrect at study (*M* = 34.0, *SD* = 20.1), $t(29) = 8.13$, $p < .001$. In final test performance, however, there was no difference between the percentage of correct study choices which remained correct at test (*M* = 57.1, *SD* = 30.5) and the percentage of incorrect study choices which were converted to correct at test (*M* = 56.5, *SD* = 25.3), $t(29) = .12$, $p = .905$. Choosing correctly at study did not benefit later memory any more than choosing incorrectly did, though participants predicted that it would.

JOLs were higher for items correct at test (Read: *M* = 40.4, *SD* = 21.8; Generate: *M* = 34.5, *SD* =19.8; Choice: *M* = 40.7, *SD* = 21.3) than for items incorrect at test (Read: *M* = 33.5, *SD* = 18.2; Generate: *M* = 27.0, *SD* =16.4; Choice: *M* = 37.1, *SD* = 20.3). As for the previous experiments, analysis of these data can be found in Appendix D.

Aggregate JOLs also showed a main effect of Study Method, $F(2,58) = 4.81$, $p = .012$, $\eta_p^2 = .14$. Choice JOLs ($M = 34.9$, $SD = 20.1$), were higher than both Generate JOLs ($M = 26.9$, $SD = 20.5$), $t(29) = 2.8$, $p = .009$, and Read JOLs ($M = 28.7$, $SD = 22.5$), $t(29) = 2.17$, $p = .038$, but there was no difference between Read and Generate JOLs, $t(29) = .78$, $p = .445$.

Study time

Study time data showed the same pattern as in Experiment 3. An ANOVA revealed a main effect of Study Method, $F(2,58) = 16.64$, $p < .001$, $\eta_p^2 = .365$. Participants spent longer studying Read items ($M = 7.34$s, $SD = 3.58$) than either Generate ($M = 6.48$s, $SD = 3.41$), $t(29) = 2.82$, $p = .008$, or Choice items ($M = 5.33$s, $SD = 2.61$), $t(29) = 4.93$, $p < .001$, and longer studying Generate than Choice items, $t(29) = 3.50$, $p = .002$.

Discussion

Experiment 4 replicated the findings of Experiment 3. Even though participants could study correct definitions for as long as they chose, generating led to better test performance than reading while JOLs showed the opposite pattern. Choosing was also superior to reading when all items were considered, though this difference fell short of significance when only items incorrect at study were included in the analysis.

Effect of final test options on test performance in Experiments 1-4

In Experiments 3 and 4, generating produced better test performance than reading, replicating previous findings. In both these experiments, and unlike in previous ones, choosing was also superior to reading when all items were considered. However, this advantage disappeared when only items incorrect at study were included in the analysis.

The multiple choice format of the final test in Experiments 1-4 allowed us to examine the effect of different lures at test. Generating was superior to reading in all experiments, no matter what options were used at final test. How did Choice performance differ across these experiments? When all Choice items were included in the analyses, whether correct or incorrect at study, choosing was disadvantaged relative to generating in Experiments 1, 2A, and 2B, in which Choice study lures all appeared at test for Choice items, while lures for Read and Generate items were all new. However, this disadvantage was eliminated when, in Experiments 3 and 4, familiarity of test lures was equated, placing the three conditions on a more equal footing. In these experiments there was no difference between generating and choosing, and choosing was superior to reading when all items were considered.

When items correct at study were excluded from the analyses, incorrect choosing never led to better test performance than reading in any of the five experiments, and it led to significantly worse performance than generating in all experiments except Experiment 1, where the difference was marginally significant, and Experiment 4, where there was a non-significant numerical advantage for generating. Thus, equating familiarity of the lures (Experiments 3 and 4) at test boosted Choice performance sufficiently to wipe out the advantage of generating over choosing when all items were considered, but not enough to eliminate the advantage when just those items incorrect at study were included.

Taken together, these findings suggest that the inclusion of all of the Choice study lures among the final test options (Experiments 1 and 2) interfered with ability to select the correct answer in the Choice condition. When this source of interference was reduced (Experiment 3) or removed (Experiment 4), choosing was not always disadvantaged relative to generating, at least when all items were considered. In Experiments 3 and 4 choosing was even superior to reading when all items were considered. However, no matter what lures were used at test, *incorrect* choosing never conferred the advantage over reading that generating did. Generation, even when it produces errors, has a powerful enhancing effect on memory for corrective feedback by comparison with passive reading. Choosing, when it produces errors, does not have this same enhancing effect.

Although incorrect choosing generally 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.

The benefit of errorful generation in Experiments 1-4

In all five experiments described so far, 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. The errorful generation benefit was observed whether or not the participant's initial guess appeared as one of the final test options, and even when participants regulated their own study time. Generating errors followed by feedback was also more beneficial than incorrect choice in Experiments 2A, 2B, and 3, with the difference falling short of significance in Experiments 1 and 4.

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 five experiments.

Table 3. Mean values for final test accuracy and JOLs across Experiments 1-4 (95% confidence intervals in brackets)

	Read	Generate	Choice	
			Correct at study	Incorrect at study
Test accuracy	71.14 (69.8 – 72.5)	79.29 (77.9 – 80.7)	78.77 (76.7 – 80.9)	73.96 (72.0 – 76.0)
JOLs	35.38 (34.3 – 36.5)	30.22 (29.1 – 31.4)	54.68 (52.6 – 56.8)	33.40 (32.2 – 34.6)

CHAPTER FIVE: THE EFFECT OF DIFFERENT CHOICE OPTIONS AT STUDY AND TEST, ON FINAL TEST PERFORMANCE AND ON JUDGMENTS OF LEARNING

Grimaldi and Karpicke (2012) found a benefit of errorful generation for related but not unrelated word pairs and proposed that this was because encoding is facilitated when the correct answer is present in the "search set". They suggested that, when participants search for a target which is related to the cue, they retrieve many possible associations, including the correct one. When weakly associated pairs are being learned, as in Grimaldi and Karpicke's experiments, the participant may give, as their response, a stronger associate to the cue but the weaker, correct one, has also been activated and is strengthened by this activation, leading to an advantage for generating over reading. For example, participants may produce "*wave*" as a response to the cue "*tide*" but other candidates (such as *beach*, *surf, ocean*) also receive some activation. If one of these is the target, this activation facilitates its encoding when it is presented as feedback. By contrast, when unrelated pairs are used, there is little likelihood that the correct answer will be present in the search set as the participant's guesses are likely to be related to the cue but not to the target.

Experiments 1-4 provided evidence that a pre-existing cue-target relationship is not necessary for the errorful generation benefit to be observed. When participants learned novel English words and foreign language words, their responses were unrelated to the targets and it is implausible that the target would have been activated along with the correct response. The facilitation observed for the learning of Generate items cannot therefore be explained by Grimaldi and Karpicke's account. In the Choice condition, however, the options available at study may be considered as constituting the search set. An interesting question, therefore, is whether the presence of the target among the options facilitates its encoding when it appears as corrective feedback, and whether encoding of the target is enhanced to a greater degree if it is the option chosen by the participant. Experiment 5 examined the idea that encoding is facilitated when the correct answer is present in the "search set" in the Choice condition. In this experiment all items were studied under Choice conditions, but the options available at study were varied so that they either included, or did not include, the target. We also examined participants' metacognitive judgments of learning with regard to this question. Experiments 2-4 showed that participants' JOLs were heavily influenced by their experience immediately prior to studying corrective feedback. Experiment 5 examined how participants' JOLs would be affected by the presence or absence of the target at study.

Experiment 5

The first aim of Experiment 5 was to examine whether the presence of the correct target among the options in the Choice condition affects its encoding when it is presented as feedback. The presence of the target among Choice options could enhance its encoding when presented as feedback, consistent with Grimaldi and Karpicke's proposal, particularly if it were selected at study. Alternatively, rejecting the correct target in favour of one of the lures could lead to impaired memory for the target at final test. In Experiment 5 all items were studied under Choice conditions. One third of the items were studied using the standard Choice condition (STD), so that the target and three lures were presented. One third were studied under a "Target Absent" (TA) condition. In this condition the correct target was replaced by an option labelled "other" which was presented with three lures. The remaining items were filler items, where the study options comprised the target, two lures and an "other" option. The filler items were included to avoid participants learning that the option labelled "other" was the correct choice whenever it was available.

If presence of the correct target among the study options enhances its encoding when it is selected, the proportion of items correctly selected at study which persist as correct at test in the STD condition will be greater than the proportion of items for which correct "other" responses are made at study in the TA condition and which are also correct at test. If presence of the correct target is harmful when it is not chosen at study, then the proportion of items incorrectly chosen at study in the STD condition which are converted to correct at test will be lower than the proportion of items incorrectly chosen in the TA condition which are converted to correct at test.

We also collected judgments of learning. In Experiment 3 we found that participants gave much higher JOLs in the Choice condition when they had guessed correctly at study than when they had not. In the current experiment we examined whether responding correctly at study, either with a target in the STD condition or an “other” response in the TA condition, would lead to higher JOLs than responding incorrectly, or if only selection of the target itself would lead to higher JOLs. We also examined whether absence of the target among the options in the TA condition would lead to lower JOLs, due to reduced familiarity of the target when presented as feedback.

Lastly, we manipulated final test format as a between-subjects factor. For both groups, the final test was in multiple choice format but for one group the lures were all lures from the study phase whereas for the other group they were new items. Experiments 1-4 provided suggestive evidence that inclusion of study lures as final test options interfered to some extent with participants’ ability to select the correct answer at test, in that choosing was no better than reading in Experiments 1 and 2, when test options for Choice items included all of the study lures, but it was more beneficial than reading in Experiment 4, when no study lures were included. Experiment 5 enabled us to compare, in the same experiment, the effect of a final test where all lures were the same as at study with one where no options had been seen at study, either as lures or as incorrect definitions. If availability at test of the same lures as were present at study interferes with selection of the correct answer, we should see an advantage in test performance for the No Lure group.

Method

Design

A mixed 2 x 2 design was used with one within-subjects factor, Study Method, which had two levels (Target Absent and Standard). The between-subjects factor was Group (Lure and No Lure). For the Lure group the options in the final multiple choice test consisted of the target and the three lures presented with that item at study. For the No Lure group the options were the correct target and three new items. The dependent measure was the number of definitions correctly selected on the final multiple choice test.

Participants

Participants were 30 members of the general population, 13 males and 17 females, average age 26 (SD = 8.0), recruited via the UCL subject pool. They were each paid £4. They were randomly allocated to two groups, with 18 in the Lure group and 12 in the No Lure group.

Materials

The stimulus set consisted of 60 unusual English words, taken from the same pool of items as was used for Experiment 1. Each word had a one-word definition and three lures which were approximately matched with the correct definition for word frequency according to the Kucera and Francis (1967) norms, and for number of syllables. Items were counterbalanced such that each item appeared equally often in each condition across participants. See Table A6 for the full set of materials.

Procedure

The study procedure was identical to that in Experiment 2 except that the two study methods were both variants of the Choice condition. In each case, the cue word was presented with four possible options. Participants were told that, for some of the items, one of the options would be “other”. In this case, the correct answer might or might not be present among the options. If they believed the correct answer was not among the options, they should select the option labelled “other”. In the STD condition, the four options consisted of the target and three lures. In the TA condition, the options were three lures and the “other” option. In this case “other” was the correct response. For the filler items, the options were the target, two lures and an “other” option. In this case “other” was an incorrect response. Item JOLs were elicited after study of each correct definition, but participants were not asked to make aggregate JOLs at the end of the study phase.

Results

Initial study phase performance

Selection of the correct answer at study in the STD condition was above the chance rate of 25%: M = 31.3, SD = 12.0, t (29) = 2.9, p = .007. Correct selection of “other” in the TA condition was also above chance, M = 32.7, SD = 15.6, t (29) = 2.5, p = .019.

Final test performance

Of primary interest was whether the presence of the target at study would affect final test performance in the STD and TA conditions. In each of these conditions, three lures are present at study. The remaining option is either the target (STD condition) or “other” (TA condition). Thus, the two conditions differ only in whether or not the target is present at study. A mixed 2x2 ANOVA, with Study Method (STD or TA) as the within subjects factor and Group (Lure or No Lure) as the between subjects factor, revealed no difference in final test score between the study methods when all items were considered, F(1,28) = .33, p = .572, nor between the groups, F(1,28) = 2.33, p = .138, and no interaction, F(1,28) = .33, p = .572. Figure 5A shows the means collapsed across groups. In the Lure group, the mean score was 75.0 (SD = 22.2) for the STD condition and 72.5 (SD = 21.6) for the TA condition. In the No Lure group, the mean score was 85.0 (SD = 18.1) for the STD condition and 85.0 (SD =

18.8) for the TA condition. Although scores were numerically higher for the No Lure group, as might be expected, the difference was not significant.

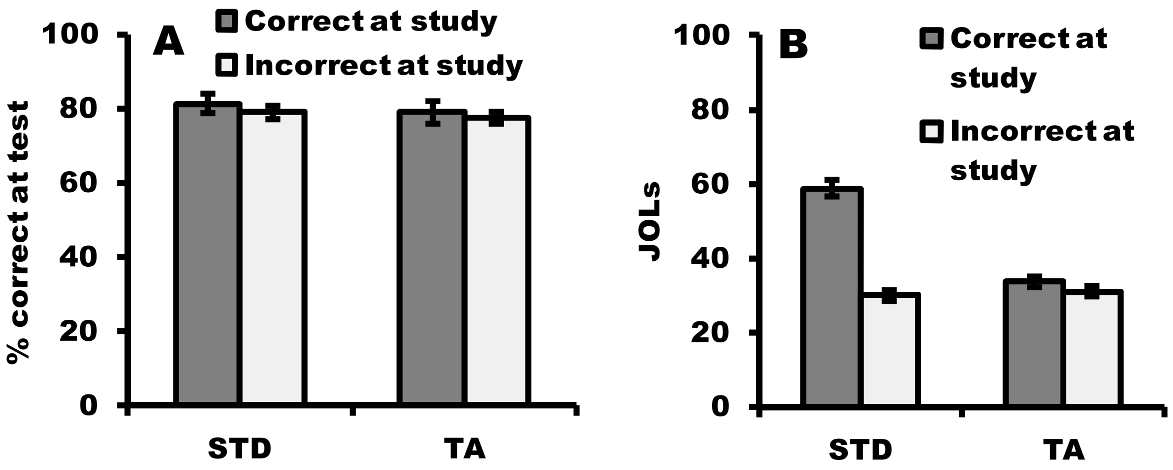


Figure 9: Experiment 5: (A) Final test performance by study method and accuracy at study. (B) JOLs by study method and accuracy at study.

Selection of the target at study

Did selecting the target at study in the STD condition lead to better final test performance than selecting "other" in the TA condition? In neither of these cases is an incorrect option picked, but in the STD condition the target is present throughout the trial, whereas in the TA condition it only appears in the last 5 seconds as feedback. STD items correctly selected at study might be better remembered than TA targets because of this difference. We conducted an analysis for just those items which were correct at study (i.e., the correct option picked in the STD condition, or "Other" picked in the TA condition). A mixed 2 (Study Method) x 2 (Group) ANOVA revealed no difference between study methods, *F*(1,24) = .42, *p* = . 524. Thus, the presence and selection of the target in the STD condition did not enhance final test performance compared with correctly selecting "other" in the TA condition. There was a significant difference between the groups, *F*(1,24) = 4.6, *p* = .

042, but no interaction between Study Method and Group, *F*(1,24) = .12, *p* = .730. Means were higher for the No Lure group (*M* = 89.7, *SD* = 16.1 for the STD condition, *M* = 88.1, *SD* = 18.9 for the TA condition) than for the Lure group (*M* = 77.6, *SD* = 23.6 for the STD condition, *M* = 72.3, *SD* = 23.0 for the TA condition). It was generally easier for the No Lure group, who had no possible interference from study lures, to pick the right answer than it was for the Lure group, but it made no difference whether the target or "other" had been correctly selected.

Effect of incorrect rejection of target at study?

Did the presence of the target impair final test performance when it was incorrectly rejected at study? We compared the percentage of items incorrect at study and converted to correct at test in the STD and TA conditions. A 2 (Study Method) x 2 (Group) ANOVA showed no effect of Study Method, *F*(1,28) = .34, *p* = .565, nor of Group, *F*(1,28) = 1.96, *p* = .173, and no interaction, *F*(1,28) = .08, *p* = .703. For the No Lure group, the means were 84.6 (*SD* = 19.0) for the TA condition and 85.3 (*SD* = 22.2) for the STD condition. For the Lure group, the means were 72.8 (*SD* = 22.3) and 74.8 (*SD* = 24.2) respectively. Selecting an incorrect lure when the target was present among the options at study did not impair performance relative to when the target was not present.

Effect of accuracy at study on test accuracy

Was it more likely that a correct response at study would also be correct at test than that an incorrect study response would be converted to correct at test? A 2 (Accuracy at study: correct/incorrect) x 2 (Study Method: STD/TA) x 2 (Group: Lure/No Lure) ANOVA showed there was no effect of Accuracy at study, *F*(1,24) = .36, *p* = .557. It was no more likely that a correct study response would persist as correct at final test (*M* = 80.4, *SD* = 22.3) than that an incorrect study response would be converted to correct at test (*M* = 78.3, *SD* =

22.4). There was no main effect of Study Method, *F*(1,24) = .56, *p* = .461, no interaction between Study Method and Accuracy, *F*(1,24) = .14, *p* = .708, and no difference between the groups, *F*(1,24) = 2.5, *p* = .124.

Judgments of learning

There was no difference in final test performance between the STD and TA conditions, whether all items were considered or just those correct or incorrect at study. Was this reflected in the JOL ratings? We analysed the data for all participants together, since the groups differed only in the options available at test, a difference which does not affect the JOLs, since these are made at study. A paired samples *t* test revealed that significantly higher JOLs were given to STD items than to TA items, *t*(29) = 4.8, *p* < .001 (Figure 5B). Thus, the absence of the target at study in the TA condition did not lead to worse memory performance but it did lead to lower JOLs.

Did accuracy at study affect JOLs? A 2 (Study Method) x 2 (Accuracy: correct or "other" vs. incorrect) ANOVA revealed a main effect of Study Method, *F* (1,25) = 43.7, *p* < .001, a main effect of Accuracy, *F*(1,25) = 51.9, *p* < .001, and an interaction between them, *F* (1,25) = 38.1, *p* < .001. JOLs were significantly higher for correct (*M* = 58.7, *SD* = 21.2) than incorrect (*M* = 29.9, *SD* = 14.0) responses in the STD condition, *t*(29) = 8.6, *p* < .001, but were no higher for TA "other" responses (*M* = 33.6, *SD* = 17.4) than they were for TA incorrect responses (*M* = 30.4, *SD* = 13.3), *t*(25) = 1.34, *p* = .192. Thus, the mere fact of getting the answer right, and correctly rejecting the incorrect lures, was not sufficient to lead to higher JOLs: Only selection of the correct target led to higher JOLs for correct than incorrect responses. JOLs were also higher for correct STD responses than for TA "other" responses, *t*(25) = 7.2, *p* < .001, but there was no difference between JOLs for incorrect STD and incorrect TA responses, *t*(29) = .59, *p* = . 563. Although JOLs for TA "other" responses were no different from TA lure responses, they were significantly higher than STD lure responses, *t*(25) = 2.47, *p* = .021. In other words, correctly rejecting the incorrect lures in the TA condition (by selecting “other”) led to higher confidence that the target would be remembered than incorrectly selecting a lure in the STD condition, again suggesting that JOLs are strongly influenced by an event immediately preceding study of corrective feedback.

Discussion

Grimaldi and Karpicke (2012) proposed that activation of the target at study, in the course of generating a guess, facilitates its encoding when it is presented as feedback, leading to a Generate advantage for related word pairs (since the target is likely to be activated during guessing) but not for unrelated word pairs (since it is very unlikely to be activated). The aim of Experiment 5 was to examine the effect, on both final test performance and judgments of learning, of the presence or absence of the target among the options in the Choice condition.

Strikingly, there was no difference in final test performance between the STD condition, where the target was included among the options at study, and the TA condition, where it was not. If encoding of corrective feedback is facilitated by the presence of the target in the search set, then we would have expected to see better performance for STD items than for TA items, in this scenario where cues are unfamiliar to participants. Of course, if participants knew some of the definitions pre-experimentally, then it could be argued that, when they select the TA “other” option, they are also activating, through retrieval, the correct definition. In this case we would not expect to see a detriment to the TA condition. However, the finding that the percentage of items correctly selected at study which remained correct at test was no greater than the percentage of items incorrect at study which were converted to correct at test, makes it unlikely that when correct responses were given at study this was because of prior knowledge. Thus, the presence of the target in the STD condition did not facilitate its encoding when it was provided as feedback, by comparison with the TA condition where it was not present.

In all analyses, the presence of the correct target among the study options had no effect on final test performance. It did, however, affect judgments of learning. In Experiment 5, JOLs for items correctly selected in the STD condition were higher than JOLs for items incorrectly selected. These data add further support to the proposal that JOLs are heavily influenced by the fortuitous selection of a correct answer at study, replicating the pattern we observed in Experiments 2 to 4. The fact that JOLs for TA “other” responses did not differ from JOLs for TA lure responses extends our findings to show that it is the selection of the target itself which is important, not simply the experience of being correct. Participants are correct both when they select the target in the STD condition, and also when they select “other” in the TA condition, but only selection of the target led to higher JOLs than for incorrect answers, adding further support to our proposal that participants experience an “illusion of knowing” when they select an option which turns out to be correct. When “other” is selected in the TA condition, the correct answer, when it appears, is unfamiliar, whereas in the STD condition, the feedback is congruent with the selected correct response and is processed more fluently as a result, leading to higher JOLs for these items.

Furthermore, while JOLs did not differ for TA “other” and TA lure responses, JOLs for TA “other” responses were higher than for STD lure responses. Even though participants did not know the answer in either case, confidence was lower when an incorrect item was picked and a correct one was present (in the STD condition) than when “other” was correctly identified as the right choice (in the TA condition).

CHAPTER SIX: ASSESSING CUED RECALL AND THE EFFECT OF INTERFERENCE FROM RESPONSES AT STUDY

In Experiments 1 and 2 we found that generating an incorrect definition for an unfamiliar English word and being provided with corrective feedback led to significantly better performance on a later multiple choice test than spending the same length of time studying the word with its definition. In Experiments 3 and 4 we observed the same effect when study of corrective feedback was self-paced. Experiments 6 and 7 were designed to explore whether an advantage for errorful generation would also be observed when memory was tested with a final test in cued recall format.

A recall test cued by the word alone would, by its very nature, enable participants to reproduce the same responses they gave at study for Generate or Choice items. In Experiment 3, where participants' study responses were also available for selection in the final multiple choice test, participants were very good at rejecting their own generated responses and better than they were at rejecting their incorrect choices. A multiple choice test differs from a recall test in the amount of cue support available to memory. In the multiple choice test of Experiment 3 both the correct target and the incorrect response either generated or chosen by the participant were available for selection. Under these conditions, it may have been relatively easy for a participant to distinguish between their own response and the experimenter-provided response. With a cued recall test, neither response is presented to the participant, who has to retrieve the relevant information from memory. If generating strengthens the item which is generated, in this case the participant's own incorrect response, this response may dominate and interfere with the ability to recall the correct target, consistent with the predictions of the "errorless learning" literature discussed in Chapter 1. In this case the advantage for generating over reading observed in Experiments 1-4 may be eliminated or even reversed when memory is tested by a cued recall test.

Because we wanted to separate the benefits of generation from the potential detriments of interference, in Experiment 6 we manipulated the opportunity for participants to respond, at test, with the same response they had given at study. If the advantage for generating over reading remains when participants have the opportunity to respond with their own errors, this would suggest that the benefits of generating are sufficient to outweigh any negative effect of interference from generated guesses. If the benefit of generating over incorrect choosing also remains in these circumstances, this would suggest that there is something powerful about the act of generating one's own response, even when erroneous, which potentiates encoding of corrective feedback, and that choosing does not confer a comparable benefit.

Experiment 6

Experiment 6 had two principal aims. First, we asked whether the benefit of generating over reading and choosing can be generalized to a cued recall final test in addition to a multiple choice test. Second, we asked whether final test performance would be affected by whether or not it was possible to select, at test, the same response which had been generated at study. To examine the effect of allowing or eliminating the possibility for interference from responses generated at study to manifest itself at final test, Experiment 6 used two slightly different versions of the cued recall final test. The first enabled intrusions from the study phase to occur at final test in both the Generate and the Choice conditions, the second allowed them in neither condition. We crossed these two final test conditions with the three study conditions in a mixed 2 x 3 design.

In one test version, *Lure-Present*, participants were presented, at final test, with a choice of two possible initial letters for the definition. One of these was the initial letter of the correct definition and the other was either the initial letter of the word generated by the participant at study in the Generate condition (for Generate items) or of the lure in the Choice condition (for Read and Choice conditions). Note that in this experiment, the Choice condition had just two options at study, as in Experiment 1, the target and one lure. In the second version of the final test, *Lure-Absent*, one letter was the initial letter of the correct definition and the other was a letter other than the initial letter of the definition generated by the participant or the lure from the Choice condition. Thus, in the Lure-Present version it was possible for participants to select the same incorrect answer they had chosen at study, while in the Lure-Absent version it was almost impossible. (It was possible in the few cases where the participant happened to generate a definition with the same initial letter as the correct definition.) Comparison of these two groups will enable us to examine the effect of interference, during retrieval at final test, of a guess generated by the participant at study.

Method

Design

A mixed design was used with one within-subjects variable, Study Method, which had three levels (Read, Generate and Choice), and one between-subjects variable, Group, with two levels (Lure-Present and Lure-Absent). The dependent variable was the number of definitions correctly recalled at final test.

Participants

Participants were 130 students enrolled in an introductory Psychology course. There were 102 females and 28 males, average age 18.8 (*SD* = 1.2). Eighty-one participants reported that English was their first language and 49 reported that it was not.

Materials

The stimulus set consisted of 60 unusual English words, each with a one-word definition, e.g., kedge – anchor, taken from the same pool of items as used in the earlier experiments. For the Choice condition, a lure was created for each of the definitions. Each lure had the same number of syllables as the correct definition and approximately the same word frequency according to the Kucera and Francis (1967) norms (Table A7). The set 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 contained the same number of nouns, verbs and adjectives. Each subset was also matched for the average word frequency of the definitions. Within each subset, no two definitions had the same initial letter. For each initial letter which appeared in one subset, the same letter also occurred in the other two subsets. For example, each of the three subsets had a definition beginning with "a", one beginning with "b" etc. The same was also true of the lures. No lure had the same initial letter as its corresponding correct definition.

Procedure

The experiment involved a single session which took around 30 minutes to complete. As in Experiment 1, there were just two options at study in the Choice condition, and no JOLs were captured. The procedure differed from that of Experiment 1 in just two respects. First, in the study phase the total trial time was 13 secs (as in Experiment 2). In the Generate and Choice conditions participants had 8 seconds to produce a response and 5 seconds to read the correct answer. Second, the final phase of the experiment was a cued recall test. All 60 English words were presented, one at a time in random order. Along with the cue, participants were presented with a choice of initial letter for the definition, one of which was the correct initial letter. For the Lure-Present group the other letter was the initial letter of the definition generated by the participant (for words in the Generate condition) or the initial letter of the lure from the Choice condition (for items in the Read or Choice conditions). For example, imagine that a participant in the Lure-Present group generated, at study, the (incorrect) definition *tasty* for the cue *gallionic* (whose true definition is *careless*). At final test, the cue *gallionic* would appear with two options: *c*____ and *t*_____. The participant's task was to enter a definition beginning with either *c* or *t*. Similarly, for the item *esurient*, the two Choice options at study were *lonely* (incorrect) and *hungry* (correct). At test, the cue *esurient* appeared with the options *l*_____ and *h*_____. These options allowed the participant to enter an incorrect definition they had generated or seen at study for the given item. For the Lure-Absent group, the alternative letter was different from the initial letter of either the generated item or the study lure item. In this case, the participant might generate the definition *tyre* for the cue word *mechlin* (whose true definition is *lace*) at study. At test, they would see the cue word *mechlin* with two options, *l*____ and *c*____. In this case, neither option allowed them to enter the same definition they had generated at study. 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 recall and enter the correct definition. Only responses beginning with one of the initial letters presented were accepted: where other responses were entered, an error message appeared and the participant was prompted to amend their answer. No feedback was given.

Results

Initial study phase performance

In the Generate condition the percentage of correctly generated definitions was very low (*M* = 0.4, *SD* =1.3). The percentage of correct answers given on the initial test in the Choice condition (*M* = 50.7, *SD* = 10.8) did not differ from the chance level of 50%, *t*(129) = .69, *p* = .492. These data confirm that the definitions were largely unknown to participants prior to the experiment. There was no difference between the groups in the number of correct selections in the Choice condition.

Final test performance

Figure 10 shows the mean number of items recalled in each of the three study methods at the final test. First we analysed performance for all items, irrespective of accuracy of initial response. The data were positively skewed and a Shapiro-Wilks test revealed that all variables were significantly non-normally distributed, probably due to the high number of zero scores in the data (around 7% of scores were zero), so non-parametric tests were used[4]. As a non-parametric means of determining whether the two versions of the task interacted with study type, we performed a rank transformation of the data and a mixed 3 x 2 (Study Method x Group) ANOVA on the transformed scores (Conover & Iman, 1981). The interaction was not significant, $F(2,256) = .694$, $p = .500$, and a Mann Whitney test showed no difference between the groups ($Z = 1.33$, $p = .185$) so we combined the data from the two groups for further analysis. A Friedman test on final test data from all 130 participants revealed a significant difference between the three study methods (χ^2= 11.731, $p = .003$). Wilcoxon signed ranks tests revealed a significant benefit of the Generate over the Read condition, $Z = 2.16$, $p = .031$. There was no significant difference between the Generate and Choice conditions, $Z = .97$, $p = .332$ or between the Read and Choice conditions, Z= 1.47, p = .142. Of 130 participants, 75 scored higher in the Generate than the Read condition, 42 scored higher in the Read than the Generate condition and there were 13 ties. A Sign test showed that the probability of this result occurring by chance was very low, $Z = 2.96$, p = .003.

Final test performance for items answered incorrectly at initial test

Next we considered final test performance for just those items which were 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. These data were also non-normally distributed. A Friedman test revealed a significant difference between study

[4] In our other experiments, scores were sometimes normally distributed and sometimes not but they all showed significant effects with both ANOVA and non-parametric tests. In Experiment 6 the ANOVA and non-parametric tests diverge. Since the data were not normally distributed and were skewed, we used a non-parametric test in this instance.

methods, $\chi^2 = 10.51$, $p = .005$. Wilcoxon signed ranks tests showed a significant benefit of the Generate over the Read conditions, $Z = 2.00$, $p = .045$, and of the Generate over the Choice conditions, $Z = 2.80$, $p = .005$, but no difference between the Read and Choice conditions, $Z = .71$, $p = .478$. These data are illustrated in Figure 10.

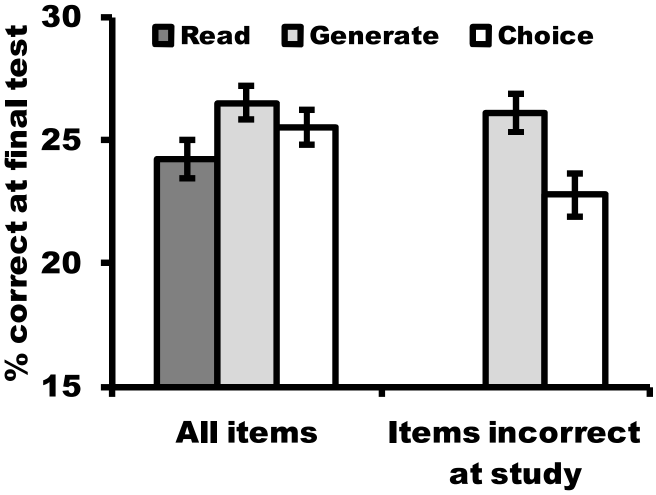


Figure 10: Mean percentage correct at the final memory test in Experiment 6. Error bars indicate standard errors.

Choice condition

In the Choice condition, where a response at final test was correct, it was significantly more likely to have been answered correctly (*M* = 14.9, *SD* = 12.5) in the initial test than incorrectly (*M* = 10.7, *SD* = 9.8), $Z = 4.41$, $p < .001$.

Intrusions at final test of errors made during study

For the Lure-Present group, we examined to what extent incorrect responses given at study were reproduced at final test as a proportion of all incorrect responses at study in each condition. Participants were more likely to reproduce incorrect definitions they had generated at initial study in the Generate condition (*M* = 9.7, *SD* = 9.4) than they were to reproduce incorrect definitions they had selected in the Choice condition (*M* = 4.7, *SD* = 7.4), and this difference was highly significant (*Z* = 3.76, *p* < .001).

Discussion

The aim of Experiment 6 was to examine whether the errorful generation benefit observed in our previous experiments would also occur when the final test was in cued recall format, and to examine more closely the effect of interference at final test. When we considered the final scores of all 130 participants, regardless of whether the item had been answered correctly or incorrectly at study, we observed the same benefit of generating over reading that we had observed in Experiments 1 to 4. When we considered only those items which had been incorrectly answered at initial study, generating an incorrect answer and receiving corrective feedback also led to significantly better performance than simply reading the cue with its definition, and to better performance than selecting an incorrect answer from a choice of two. Thus we have now observed this benefit of generating over reading both with a multiple choice final test and with a cued recall final test.

In Experiment 6 two versions of the final test were created in order to examine the effect of interference from responses which had been generated or chosen at study. There was no difference in final recall between the Lure-Present and Lure-Absent groups, suggesting that the opportunity to select the same response did not impair performance relative to when there was no such opportunity. However our analysis of the persistence of errors in the Lure-Present group showed that participants were significantly more likely to reproduce an error made in the Generate condition than one made in the Choice condition, the opposite of the pattern observed in Experiment 3, where the final test was in multiple choice format. Despite these error intrusions, there was still a significant advantage for the Generate over the Read condition in the Lure-Present group but no advantage for choosing over reading. This striking result suggests that there is something about the act of generation, even when it results in an error, which enables corrective feedback to be processed effectively enough to outweigh the negative effects of interference. Errors made at study in the Choice condition were less likely to persist to final test, but this did not lead to an advantage for choosing over reading. Generating, then, enabled both errors and correct responses to be better remembered than in the Choice or Read conditions.

Perhaps the act of generation, because it involves activating many cues, leads to a more distinctive memory representation being formed than in either of the other two conditions. As long as participants are able to distinguish their own, erroneous, response from the correct answer provided as feedback, this will lead to better recall in the Generate condition than in the other conditions. Although a final test in multiple choice format may facilitate this task by presenting the correct target alongside the incorrect responses, the outcome of Experiment 6 suggests that Generate items are strengthened sufficiently to lead to an advantage for generating even when participants have to retrieve the answers themselves in a cued recall format final test. Choosing also strengthens items, but to a lesser extent: Choosing led to better final test performance than reading in Experiment 3, when the final multiple choice test included both correct target and incorrect responses among the options, but was no better than reading in the cued recall test of Experiment 6.

Experiment 7

Experiment 6 demonstrated that the Generate benefit observed in our earlier experiments extends to a cued recall final test. However, scores were very low and performance was at floor for many participants. The aim of Experiment 7 was to attempt to replicate the finding of an errorful generation effect using a cued recall final test while achieving higher scores. We therefore used 40 items instead of 60, and only two study methods, Read and Generate, with 20 items in each. In addition, we used the same, slightly longer, timings that we used in Experiment 1, to give participants more time to process the definitions. In Experiment 7, all participants took the same version of the final test, which was cued by the word alone. This will enable us to confirm that the effect can be replicated with a more traditional format of cued recall test.

Method

Participants

Participants were 24 members of the general population. 10 males and 14 females, average age 27.8 (SD = 8.4), recruited via the UCL subject pool. They were each paid £4.

Materials

The stimulus set consisted of 40 unusual English words, taken from the same pool of items as was used in the earlier experiments (Table A8). Half the items were presented in the Read condition, half in the Generate condition. Each item appeared equally often in each condition across participants.

Procedure

The study procedure was identical to that in Experiment 6 with three exceptions. First, there was no Choice condition at study and only 40 word-definition pairs were used. Second, items in the Read condition were displayed for 17 seconds while, in the Generate condition, participants had 10 seconds to generate a response and 7 seconds to view feedback (as in Experiment 1). Third, at test, all 40 cues were presented alone – unlike in Experiment 6, the initial letters of targets and lures were not presented – and participants were instructed to type in the correct definition. As in the previous experiments, each cue remained on screen until the participant entered a response.

Results

Initial study phase performance

In the Generate condition, across all participants, only 1 item was answered correctly at study, though this item was not correctly answered at test, suggesting that at study it may have been a fortuitous guess rather than an item which was already known to the participant.

Final test performance

Scores were higher in the Generate (*M* = 30.8, *SD* = 18.7) than the Read (*M* = 26.9, *SD* = 26.1) condition, though this difference failed to reach significance, *t*(23) = 1.34, *p* = .195. Fifteen participants scored higher on Generate than Read items, with eight showing the opposite pattern. Inspection of the data revealed that two participants scored substantially higher than the others in the Read condition, scoring 90% and 95% respectively, where the next highest score was 55%. These scores were 2.4 and 2.6 standard deviations above the mean respectively. When data from these two participants were excluded from the analysis, a paired samples *t* test showed a significant advantage for generating (*M* = 27.7, *SD* = 16.1) over reading (*M* = 20.9, *SD* = 17.3), *t*(21) = 2.83, *p* = .010.

There was some tendency for errors made at study to be repeated at final test. Participants reproduced, at final test, errors they had generated at study 6.9% of the time (*SD* = 9.3). Seventeen of the 24 participants reproduced at least one of their own errors.

Discussion

Although the advantage of generating over reading did not reach significance in Experiment 7 when data from all participants were included in the analysis, there was no evidence of any harmful effect of generating errors, despite the fact that participants had less time to process correct answers in the Generate condition (just 7s) than in the Read condition (17s). When data from two participants who scored substantially higher than the other participants in the Read condition were removed from the analysis, a strong generation benefit was again observed. These two were also the highest scorers overall. Brewer and Unsworth (2012) found that testing provided more benefit to students with poor memory abilities, as measured by performance on a series of episodic memory tests, than to those with high memory abilities. It is possible that high scoring participants use highly effective strategies to process Read items, including, perhaps, rehearsal and self-testing, during the time that Read items are on the screen, leading to a reduction or elimination of the Generate benefit. It would be interesting for future research to examine the factors affecting individual differences in the errorful generation benefit.

Discussion of Experiments 6 and 7

Experiments 6 and 7 replicated our previous finding that generating a definition for an unfamiliar English word, even when the generated response is incorrect, has a powerful effect on subsequent retention of the correct definition. We have now observed an advantage of generating errors over reading both when memory is tested with a multiple choice test (Experiments 1-4) and also with a cued recall test (Experiments 6 and 7, though the difference fell short of significance in Experiment 7 when all participants were included). We have also shown that the advantage for the Generate condition persists when participants have the opportunity to reproduce their own errors at test, whether the final test is in cued recall format or multiple choice format. This is interesting in light of the fact that, when the test was in multiple choice format (Experiment 3) participants were very good at rejecting responses they had themselves generated at study, selecting these less often than they selected Choice lures they had selected at study. When the test was in cued recall format, however (Experiment 6), they were significantly more likely to give, at test, the same incorrect response they had given in the Generate condition than the same incorrect response they had given in the Choice condition. In spite of this interference from responses made at study, the advantage for generating over reading remained, and generating was also more beneficial than choosing.

CHAPTER SEVEN: IS INTERFERENCE INCREASED BY GENERATING MULTIPLE ERRORS?

The preceding experiments have demonstrated an advantage for errorful generation over reading whether the final test is in multiple choice or cued recall format. This benefit of generating errors is striking given that a large body of literature on “errorless learning” has proposed that errors generated during learning can have a detrimental effect on later memory performance (e.g., Baddeley & Wilson 1994). Errorless learning studies typically compare a condition in which participants generate many errors in response to a test cue with a condition in which they are presented with the correct answer, with the latter proving more beneficial to later memory. An intriguing question concerns why these studies so often show an advantage for a passive errorless condition relative to a condition involving the generation of errors, while our experiments have consistently shown the reverse pattern. There are several differences between our task and the ones typically used in errorless learning studies. One difference which may be important is in the amount of potential for interference from errors that is created in the typical errorless learning task by comparison with ours. In errorless learning studies, not only are participants encouraged to make many erroneous guesses on each presentation of the item, but they also go through several cycles of learning, increasing the potential for errors. In contrast, in our task only one error is generated. It is possible that multiple incorrect responses cause greater interference, resulting in a detriment to the errorful condition in the errorless learning studies.

In Experiment 8 we aimed to bring our procedure closer to the procedure used in a typical errorless learning study by having some participants go through several cycles of learning. In this experiment items were 40 Euskara – English word pairs, and they were each studied in either the Read or the Generate condition. The final test was a cued recall test. Participants underwent one, two or three cycles of learning. If the errorless learning advantage observed in some studies occurs because multiple cycles of learning create more interference from errors in the errorful condition, we would expect to see a reversal of the Generate benefit between the one-cycle group and the three-cycle group. That is, we would predict a Generate over Read advantage in the one-cycle group, replicating our previous findings, but a Read over Generate advantage in the three-cycle group, due to increased interference from multiple errors.

Experiment 8

Method

Design

The design was a mixed 3 x 2 design, with Group (number of cycles: 1, 2, or 3) as a between subjects factor, and Study Method (Read or Generate) as a within subjects factor.

Participants

Participants were 118 students enrolled in an introductory Psychology course. There were 97 females and 21 males, average age 18.7 (SD= 0.9). Seventy-five participants reported that English was their first language and 43 reported that it was not. They were randomly allocated to three groups, which underwent one (N = 40), two (N = 38) or three (N = 40) cycles of learning respectively.

Materials

The stimulus set consisted of 40 Euskara words with their English translations, taken from the same pool of items as was used in Experiment 2B (Table A9). Half the items were presented in the Read condition, half in the Generate condition. Each item appeared equally often in each condition across participants.

Procedure

The study procedure was identical to that in Experiment 7 with two exceptions. First, items in the Read condition were displayed for 13 seconds, while in the Generate condition participants had 8 seconds to generate a response and 5 seconds to view feedback. Second, the number of cycles of learning varied according to the group to which the participant had been allocated. For the one-cycle group, the study phase was similar to that in Experiment 7, as the set of 40 items was presented once only. In the two-cycle and three-cycle groups, following initial presentation of all 40 items, they were presented again, in a different random order each time, for either one or two additional cycles. In each cycle, items were presented in the study method to which they were initially assigned. That is, Read items were presented in Read format in each cycle in which they appeared, and Generate items appeared in Generate format. As in Experiment 7, the final test was in cued recall format and there were no time constraints.

Results

Initial study phase performance

No items were correctly generated on the first cycle of study, confirming that the words were unknown to participants pre-experimentally.

Final test performance

Minor variations in spelling at final test were allowed. A 3 (Group) x 2 (Study Method) ANOVA on the final test scores showed no main effect of Study Method, *F*(1, 115) = 1.71, *p* = .194, a significant effect of Group (number of cycles), *F*(2,115) = 59.44, *p* < .001, and no interaction, *F*(2, 115) = 1.82, *p* = .167. The effect of Group was as expected: final test scores were significantly higher for the two-cycle group than the one-cycle group, *t*(76) = 5.72, *p* < .001, and higher for the three-cycle group than for the two-cycle group, *t*(76) = 5.05, *p* < .001. The means are shown in Figure 11.

The mean percentage of correct responses at final test, across all groups, was 42.71 (*SD* = 29.4) for the Read condition and 40.85 (*SD* = 27.8) for the Generate condition. The means for the three groups are shown in Figure 11.

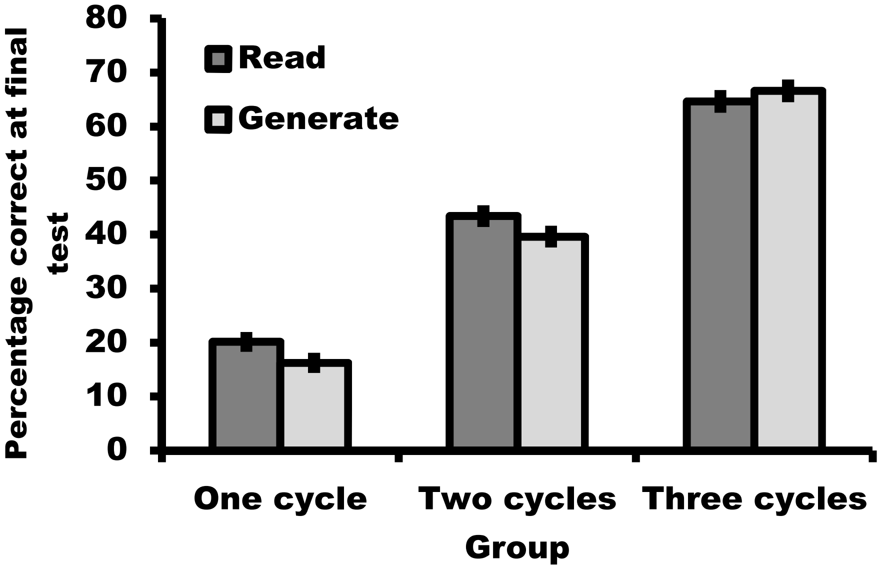


Figure 11: Mean percentage correct at the final memory test in Experiment 8. Error bars indicate standard errors.

Discussion

In the light of our previous experiments (especially Experiments 6 and 7 which also used a cued recall test format) it is surprising that there was no advantage for generating over reading in the one-cycle group. However, there was no disadvantage to generating, despite the fact that in the Read condition the correct answer was available for the full 13s of the trial time, whereas it was displayed for just 5s in the Generate condition. Indeed, in the three-cycle group, there was a numerical advantage for Generate items at final test, even though at least one error (and in some cases three) had been produced. Correct answers for these items had been presented for a total of 15s across the three cycles, while correct translations for Read items had been displayed for a total of 39s. Thus generating incorrect responses was not harmful to learning compared with simply reading the translation. There was no evidence that increasing the number of errors generated for each item led to any detriment to Generate items compared with Read items.

However, in errorless learning studies participants typically make several guesses per trial, in addition to undergoing several cycles of learning, so there is potential for many more erroneous responses to be generated. It is possible that our manipulation, by allowing only one error per trial, did not create sufficient interference to cause any detriment to memory. It will be interesting for future research to examine whether generating several errors on each trial, in addition to undergoing several cycles of learning, will increase the amount of interference to the extent that it nullifies, or reverses, the benefit of generation. In errorless learning studies it is usual for responses to be given orally in the errorful condition. A possible future modification of our task could therefore have participants respond orally to cues in the study phase and be encouraged to generate up to four errors per trial, as in the typical errorless learning study. Responding orally reduces the amount of time needed for each trial, compared with typing in four responses, and will mirror the typical errorless learning procedure more closely.

Errorless learning studies also often constrain responses by giving participants the initial letter or letters of the target. For example, Haslam, Hodder, and Yates (2011) had participants generate names in response to pictures of faces and provided them with the initial letter (e.g., “S”) of the name to be guessed. Participants were encouraged to make up to four guesses (e.g., *Susan, Sarah, Suzanne, Sonya*) before being given the correct answer. This method may increase interference by increasing similarity between the incorrect guesses and the target. Because responding is constrained to a limited number of possibilities it may also increase the subjective plausibility of the incorrect responses generated: in other words, participants may have more confidence that one of their responses will be correct. In the task used in the experiments reported here, where the set of possible responses is unconstrained, participants are unlikely to believe that they will happen to give a correct response to a cue they have never seen before and they may therefore be less committed to their own responses. A possible avenue for future research, therefore, is to constrain participants' responses in the Generate condition. This could be done by providing the initial letter, increasing phonological similarity between responses, and thereby potentially increasing interference. Alternatively, a category to which the item belongs could be provided (e.g., fruits, animals etc.). This would increase the probability of generating a correct response, and therefore may lead participants to commit more firmly to their own responses, which may also increase the potential for interference.

CHAPTER EIGHT: TESTING CAN PROTECT MEMORIES AGAINST INTERFERENCE

Experiments 1-8 examined the proposal that, while testing is often beneficial to memory, taking a test that generates many errors can be harmful to subsequent memory, because the errors we make ourselves may interfere with memory for corrective feedback presented after the generation attempt. In contrast, those experiments showed that the act of generation potentiates the encoding of subsequent feedback sufficiently to outweigh any negative effect of the errors produced by the test itself. Testing can therefore be beneficial even when it produces many errors, and even during novel learning when there is no pre-existing association between cues and targets, by potentiating the processing of feedback.

Experiments 9 and 10 examine a different scenario in which the benefits of testing may be affected by interference. In this instance it is proposed that testing can be harmful to memory when correct information is retrieved but interfering new learning is introduced immediately after retrieval. According to this proposal, the very act of reactivating a consolidated memory can destabilize it, so that it needs to be reconsolidated in order to persist in its original form. During the reconsolidation period memories are particularly susceptible to interference, which may cause them to undergo modification or even unlearning. In Experiments 9 and 10 we investigated whether retrieving memories could make them more susceptible to interference from new learning following retrieval or whether, by contrast, testing can protect memories against interference. I begin with a brief review of the literature in this area.

Does retrieval make memories more susceptible to interference?

As noted in Chapter 1, the beneficial effect of retrieval practice on later recall of the retrieved items is well established (e.g., Allen, Mahler, & Estes, 1969; Karpicke & Roediger, 2008) and is observed even when no corrective feedback is given (Allen et al., 1969;

Carpenter & DeLosh, 2006; Kuo & Hirshman, 1996), and when testing is compared with a restudy condition (e.g., Carpenter, Pashler, Wixted, & Vul, 2008; Cull, 2000), suggesting that something specific to the retrieval process itself leads to enhanced memory for the tested items. Bjork and Bjork (1992) proposed that testing increases the retrieval strength, or accessibility, of successfully retrieved items and the more difficult the retrieval, the greater the boost to retrieval strength. When retrieval is easy, a test confers little benefit. The recent distribution-based framework proposed by Bjork and colleagues (Halamish and Bjork, 2011; Kornell, Bjork, &, Garcia, 2011) for understanding when testing is and is not beneficial, proposes that successful final recall is a function of the difficulty of the final test and retrieval strength at the time of this final test. A more difficult final test means that retrieval strength has to be higher in order for an item to reach the threshold for successful retrieval. Thus retrieval practice, which boosts retrieval strength, is most beneficial when the subsequent final test is difficult, because the additional strength provided by practice allows items to be retrieved which otherwise would not be.

Despite widespread evidence of the benefits of testing, it can sometimes have detrimental effects on memory, for example by increasing false memory as well as correct memory (McDermott, 2006) or by increasing susceptibility to misinformation (Chan, Thomas, & Bulevich, 2009). In these studies learning and testing took place within a single session, before consolidation would have had time to occur. Memory impairment or updating in declarative memory has also been observed, however, when consolidated memories were reactivated, a day or two after learning, by a brief reminder such as the presentation of a single cue from the learned list, or a brief allusion to the original study episode, immediately before the learning of new information (e.g., Forcato, Argibay, Pedreira, & Maldonado, 2009; Forcato, Burgos, Argibay, Molina, Pedreira, & Maldonado, 2007; Forcato, Rodriguez, Pedreira, & Maldonado, 2010; Hupbach, Gomez, Hardt, & Nadel, 2007; Hupbach, Gomez, &

Nadel, 2009). In these demonstrations of disruption to reconsolidation the reminder did not include a complete test of the original memory so the question of whether the benefits of testing would be sufficient to outweigh impairment or modification caused by interfering learning remains open. However, a study of motor skill learning by Walker, Brakefield, Hobson, and Stickgold (2003) did include a complete test of the original material as the reminder treatment. Walker et al. had participants learn two different finger-tapping sequences a day apart. At a final test on the third day, memory for the first sequence was poorer when participants had briefly rehearsed this sequence before learning the second sequence on Day 2 than when they had not. The authors proposed that reactivating the memory had made it labile and susceptible to disruption by the learning of the new sequence. This raises the interesting question of whether taking a test could, in similar fashion, lead to impairment of declarative memory.

Does retrieval lead to an increased susceptibility to interference or can it immunize memory against interference? There is some evidence that testing may protect against proactive interference when a memory is first acquired. Szpunar, McDermott, and Roediger (2008) found that participants who were tested after studying each of five lists showed better recall of List 5 than participants who were only tested on List 5, and they made fewer intrusions from earlier lists. The effect was specific to retrieval: only testing, not restudying, after each list protected against the build up of proactive interference.

Halamish and Bjork (2011, Experiment 3) investigated whether retroactive interference is moderated by testing, also at the time of initial memory acquisition. Participants studied cue-target pairs followed by two cycles of either restudy or test. They then read cue-target pairs, some of which had the same cues as studied pairs but different targets (*interfering pairs*). This was followed by a final cued recall test of the original studied pairs. Testing led to better recall than restudying only when it was followed by interfering pairs. Testing did not completely immunize against interference, however: recall was significantly lower on the final test than on the initial test when interfering pairs were studied in between.

The study by Halamish and Bjork (2011) suggests that testing may provide some protection against retroactive interference during memory acquisition but does not immunize the memory completely. Here, again, learning and testing took place within a single session, before consolidation would have taken place, and participants were required only to read the interfering pairs, not to learn them. In the current study we examine what happens when a memory has already been consolidated and is retrieved for the purposes of a test taken a day after learning. The reconsolidation hypothesis proposes that retrieving a consolidated memory may labilize it, making it vulnerable to unlearning if new material is learned immediately afterwards. In Experiments 9 and 10 participants learned items over a three-day period, to allow time for memory consolidation and reconsolidation to occur. After learning English-Swahili word pairs (List 1) on Day 1, some participants took a "reminder" test of these words, thereby reactivating the memory, before learning interfering English-Finnish words (List 2) on Day 2. Others learned the English-Finnish pairs without the reminder test. All were given a final test on Day 3. The reconsolidation studies discussed earlier showed impairment or modification to memories when they were reactivated using just a subtle reminder. The goal of Experiments 9 and 10 was to investigate the effect of memory retrieval in the more usual circumstances in which retrieval practice takes place, such as when taking a vocabulary test. Under these circumstances does retrieval make memory vulnerable to interference or does the benefit of retrieval practice outweigh, or even immunize against, any negative effect of interference? By having participants learn, rather than simply read, the interfering List 2 we aimed to create the conditions necessary for disruption to the already-consolidated List 1 memory.

Experiment 9 investigates whether reactivating memory of List 1 by taking a test leads to a detriment to that memory when new material is learned during the reconsolidation period, leading to poorer subsequent recall of List 1 in the tested group. By contrast, we find superior recall in this group. Experiment 10 examines whether the magnitude of the testing benefit observed in Experiment 9 is reduced by the learning of List 2 or whether testing immunizes List 1 memory to the extent that it is invulnerable to the detrimental effects of interference. By using four groups which between them represent all possible combinations of the reminder test and the List 2 interference task, we show that testing and interference exert opposite effects on memory but that testing, far from making memory more vulnerable to interference, protects memory against interference.

Experiment 9

Method

Participants

Thirty-two participants, 16 female, average age 33.31 (*SD* =18.01), took part in return for a small payment and were equally divided between two groups. A further five were excluded because they failed to complete all three sessions and one because she already knew some Swahili words.

Materials

A list of 20 English-Swahili word pairs (List 1) was constructed, for study by both groups of participants on Day 1. A second list was constructed, comprising 20 English-Finnish word pairs (List 2), for study by both groups on Day 2. Half of the word pairs in List 2 shared the same English cues as List 1 (Table A10). Thus the English word *bed* appeared only with its Swahili translation *kitanda,* whereas the word *cat* appeared on both lists with its translations *paka* (Swahili) and *kissa* (Finnish).

Procedure

Both groups studied List 1 on Day 1 and List 2 on Day 2. One group (R+I+) received a test of List 1 immediately before learning List 2, while the other group (R-I+) did not (R=Reminder, I=Interference). All participants received a cued recall test of both lists on Day 3.

At study (of List 1 on Day 1, and of List 2 on Day 2), word pairs were presented one at a time for 6 secs, in randomized order. After every four word pairs, recall of the four translations was tested, cued by the English word, until each word had been correctly translated twice. Corrective feedback (in black for a correct response, red for incorrect) was provided on every trial. After study of all 20 items, six blocks of practice followed. In each block, each word was tested, with feedback, until correctly answered, at which point it was dropped from practice for the remainder of that block. In the first two blocks, five words were presented until correctly answered, and then a new set of five was tested and so on until all 20 had been answered correctly. In the third block, words were tested in two sets of 10. In the remaining three blocks, all 20 words were tested in randomized order and participants were shown how many they had translated correctly on first presentation in that block. On completion of the task, each word had been correctly translated exactly eight times, while number of presentations varied. On Day 2, the R+I+ group received a reminder test of List 1 immediately before studying List 2. For the reminder test, all twenty cue words were presented, one at a time, and the participant was instructed to type in the correct Swahili translation for each. No feedback was given. The R-I+ group did not receive the reminder test.

On Day 3, participants took four tests without feedback. First they were tested on List 1 Forwards (English to Swahili). Each cue word was presented and participants typed in the translation. This process was repeated for List 1 Backwards (Swahili to English), then for List 2, first forwards and then backwards.

Results and Discussion

Was memory for List 1 affected by the taking of a reminder test on Day 2? Table 4 shows the mean scores for words with shared and unique cues for each group in all four tests.

Table 4. Mean scores out of 10 (SDs in brackets) obtained on all four final tests in Experiment 9, by cue type.

Day 3 Test	Group	Shared cues	Unique cues
List 1 Forwards	R+I+	6.4 (2.7)	7.9 (1.9)
	R-I+	5.0 (2.4)	5.8 (2.5)
List 2 Forwards	R+I+	6.4 (2.8)	7.4 (2.4)
	R-I+	5.8 (2.3)	6.9 (2.0)
List 1 Backwards	R+I+	7.6 (2.7)	8.4 (2.0)
	R-I+	7.0 (1.8)	7.9 (2.3)
List 2 Backwards	R+I+	7.8 (3.2)	8.4 (1.3)
	R-I+	8.8(1.5)	8.5 (1.6)

Figure 12 shows the total scores collapsed across cue type. For the List 1 Forwards test, a 2 (Group) x 2 (Cue Type) mixed ANOVA did indeed show a main effect of Group, F(1,30) = 5.04, *p* = .032, but recall was higher for the R+I+ group. Far from being deleterious to the List 1 memory, the reminder test enhanced it, that is, there was a testing effect.

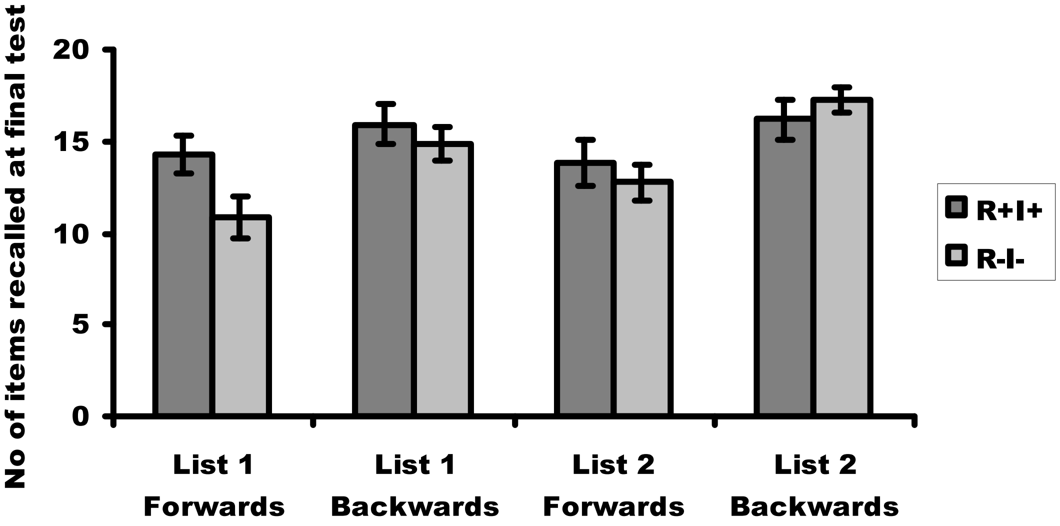


Figure 12 Mean scores (out of 20) on all four final tests in Experiment 9, collapsed across cue type.

There was also a main effect of Cue Type, *F*(1,30) = 14.98, *p* = .001. Unique cue words were better recalled than shared cue words, but there was no interaction between Group and Cue Type, *F*(1,30) = 1.49, *p* = .231, indicating that the group effect was similar for shared and unique cues. Unique cue words were also better recalled than shared cue words in the List 2 Forwards test, *F*(1,30) = 18.01, *p* < .001. However, there was no difference between the groups, *F*(1,30) = .45, *p* = .510, and no interaction, *F*(1,30) <1, *p* = .702.

The advantage for the unique cued words over the shared cue words was also present in the List 1 Backwards test, *F*(1,30) = 7.58, *p* = .010 but again there was no difference between the groups, *F*(1,30) < 1, *p* = .469, and no interaction, *F*(1,30) < 1, *p* = .919. For the List 2 Backwards test, no effects were significant: Cue Type, *F* (1,30) < 1, *p* = .646; Group, *F*(1,30) < 1, *p* = .42; with no interaction, *F*(1,30) = 1.45, *p* = .237. Thus the effect of the reminder test was specific to forwards and not backwards testing of List 1. Indeed, a 2 (Group: R+I+ vs. R-I+) x 2 (Test: List 1 Forwards vs. List 1 Backwards) ANOVA showed a significant interaction between Group and Test, *F*(1,30) = 9.34, *p* = .005.

Did taking a reminder test make memory for List 1 more susceptible to updating with words from the other list? It was marginally more common for List 2 translations to be mistakenly given in the List 1 test (*M* = .59, *SD* = .91) than vice versa (*M* = .25, *SD* = .67), *F*(1, 30) =3.95, *p* = .056, but there was no difference between the groups, *F*(1,30) < 1, *p* = .89, and no interaction, *F*(1, 30) = 2.65, *p* = .11. Taking the reminder test made no difference to the extent to which items were erroneously translated using a word from the other language.

Experiment 10

Experiment 9 provided no evidence that reactivating the List 1 memory by taking a reminder test made it more susceptible to interference from List 2 learning. On the contrary, instead of a reconsolidation effect we found a testing effect: the reminder test led to enhanced memory for List 1. However, it is possible that we did not observe disruption to List 1 memory because List 2 was not an effective interfering agent. In Experiment 10, we therefore included a group which learned List 2 immediately after List 1 (I+), and a control group which just learned List 1 (I-), to determine whether List 2 was capable of interfering with consolidation of List 1 when that memory was first acquired. The I+ and I- groups took a final test of List 1 one day after List 1 learning, in order to equate the interval between acquisition and test with the interval between retrieval (reminder test) and final test in the groups which received a reminder. Secondly, it is possible that there was some disruption to the List 1 memory in the R+I+ group in Experiment 9, but it was offset by the beneficial effect of the extra practice this group received by having the reminder. In other words, the test did not completely immunize List 1 memory against interference and participants in the R+I+ group might have performed even better had they not learned List 2 after the List 1 reminder. In Experiment 10, in addition to the R+I+ and R-I+ groups of Experiment 9, we included two additional groups. The R+I- group took a reminder test but did not learn List 2, and the R-I- group did neither. Thus we had four groups which between them represented every combination of reminder testing and interference, plus two groups to test for an interfering effect of List 2 on learning (rather than recall) of List 1. Table 5 shows the schedule of learning and testing for all six groups.

Table 5: Schedule of learning and testing for the six groups in Experiment 10. L1 = Swahili learning task. L2 = Finnish learning task. ? = test.

Group	Day 1	Day 2	Test (Day 3)
I-		L1	L1?
I+		L1 - L2	L1? L2?
R+I+	L1	L1? - L2	L1? L2?
R-I+	L1	L2	L1? L2?
R+I-	L1	L1?	L1?
R-I-	L1		L1?

Method

Participants

Participants took part in return for a small payment. After exclusion of five who failed to attend all sessions and seven who failed to complete the task within the prescribed number of trials, 100 participants remained, 62 female, average age 23.88 (*SD* 7.19). They were randomly assigned to one of the six groups such that there were 14 participants in each of the I+ and I- groups and 18 participants in each of the remaining four groups (R+I+, R+I-, R-I+, R-I-).

Materials

Materials were similar to those used in Experiment 9, except that all 20 English cues were shared by the two lists in order to maximize the potential for interference (Table A11).

Procedure

For the I+ and I- groups, there were two sessions on consecutive days. The I+ group learned List 1 immediately followed by List 2 on their first day (designated Day 2 in Table 5 in order to align the final tests), while the I- group learned List 1 only. Both groups were tested the following day. The remaining four groups learned List 1 on Day 1 and were tested on Day 3. On Day 2, the R+I+ group took a reminder test of List 1 followed by learning of List 2; the R-I+ group learned List 2 only; the R+I- group took the reminder test only, and the R-I- group did neither.

For all groups, study of Lists 1 and 2 proceeded in the same way as in Experiment 9 except that there were five blocks of practice instead of six and, in the last two blocks, the block was terminated after 45 trials even if participants had not answered every item correctly, in order to limit the length of the task for participants who found it difficult. Data from participants who reached the 45-trial limit without having answered all items correctly were excluded from the analysis. On completion of the task, each word had been correctly translated exactly seven times, while number of presentations varied. On Day 3, first List 1 was tested, forwards and backwards, then List 2, forwards and backwards.

Results

There was no significant difference between the six groups in the scores achieved in the final block of List 1 learning, $F(5,94) = 1.40$, $p = .23$, nor in the number of trials taken to complete the task, $F(5,78) < 1$, $p = .51$. Table 6 shows the means.

Table 6: Mean number of items correctly recalled first time in final block of List 1 (out of 20) and number of trials required to complete the task in Experiment 10 (SD in brackets).

Group	Mean no. of items correctly recalled first time in final block of List 1	Mean number of trials taken to learn List 1
I-	17.4 (2.8)	201.7 (44.3)
I+	18.4 (1.7)	195.7 (40.0)
R-I+	17.7 (2.6)	206.6 (45.6)
R+I+	18.7 (1.6)	194.9 (38.5)
R+I-	18.7 (1.8)	215.4 (64.0)
R-I-	18.8 (1.3)	186.7 (33.7)

Confirming that List 2 was capable of interfering with List 1 on initial acquisition, the I- group scored significantly higher than the I+ group on both the List 1 Forwards test $t(28) = 2.45$, $p = .021$, and the List 1 Backwards test $t(28) = 2.14$, $p = .042$. Learning List 2 immediately after List 1 impaired recall of List 1 a day later (Figure 13A). Thus List 2 is an effective interfering agent.

To assess the effects of taking the reminder test and of learning List 2, and their interaction, on performance in the List 1 Forwards test we carried out a 2 x 2 between-subjects factorial ANOVA, with Reminder (Reminder test, R+/No reminder test, R-) and Interference (List 2 learning, I+/No List 2 learning, I-) as the factors, for the remaining four groups. This revealed main effects of Reminder, $F(1,68) = 6.37$, $p = .014$, and of Interference, $F(1,68) = 29.06$, $p < .001$, and a significant interaction, $F(1,68) = 9.59$, $p = .003$. Final test recall performance for these four groups is shown in Figure 13B. Simple effects analyses showed that the reminder test had a highly significant effect on List 1 Forwards recall in the presence of interference, F (1,68) = 15.79, $p < .001$, revealing a testing effect, but not in its absence, *F* (1, 68) < 1, *p* = .686. Interference, in the form of List 2 learning, had a significant effect on List 1 Forwards recall when there was no reminder test, *F* (1, 68) = 36.02, *p* < .001, reducing List 1 recall by nearly half. However, taking a reminder test before the List 2 learning improved performance substantially, so that there was no longer a significant effect of interference on List 1 Forwards recall, *F* (1, 68) = 2.63, *p* = .109. These results suggest that taking a reminder test immunized the List 1 memory against interference caused by the learning of List 2. Strikingly, this powerful benefit to subsequent recall of taking a reminder test was only evident when the test was followed by interfering learning. Recall of List 1 was just as good in the R-I-group, which did nothing on Day 2, as in the R+I- group, which took the reminder test alone, suggesting that, when there was no interfering task, memory for the original List 1 was as good 48 hours after learning as it was after 24 hours and there was no additional benefit of taking a test.

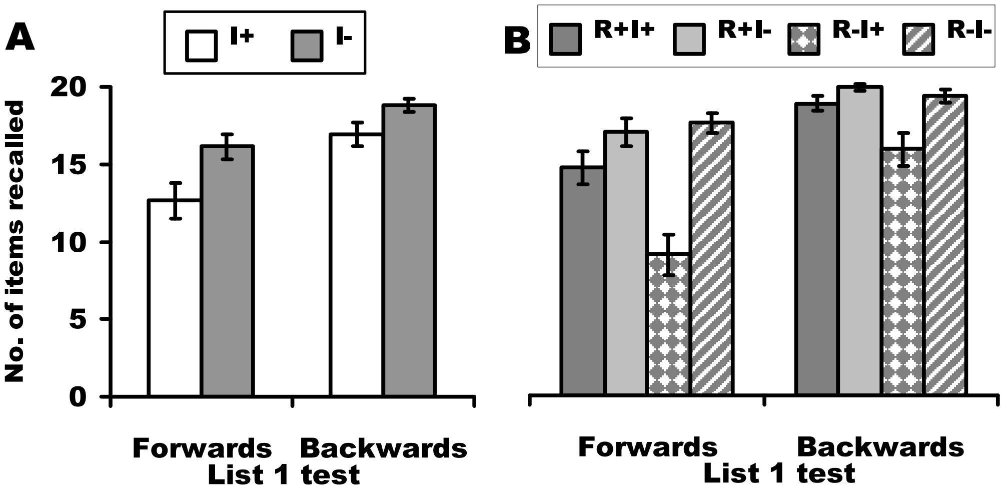


Figure 13. No. of List 1 items correctly recalled at final test (out of 20) in Experiment 10 for (A) the I + and I- groups and (B) the R-I+, R-I-, R+I+ and R+I- groups.

There was no significant difference in performance on the List 2 Forwards test between the R-I+ (*M* = 12.3, *SD* = 4.0) and R+I+ (*M* = 12.2, *SD* = 5.6) groups, (*t*(34) < 1, *p* =

.919), and no difference on the List 2 Backwards test (R-I+: *M* = 18.1, *SD* = 1.6, R+I+: *M* = 18.6, *SD* = 1.7), *t* (34) < 1, *p* = .314. Taking the reminder test enhanced memory for List 1 without affecting memory for List 2.

Although the equivalent performance on the List 1 Forwards test in the R+I- and R+I+ groups suggests that the reminder test was sufficient to neutralize the effects of interference altogether, another way to analyse the data is to compare performance on the reminder test and the final test for these two groups. A mixed 2 (Test: reminder vs. final) x 2 (Group: R+I- vs. R+I+) ANOVA revealed a main effect of Test, *F*(1,34) = 11.84, *p* = .002, and an interaction between Test and Group, *F*(1,34) = 18.78, *p* < .001, such that there was a significant increase in recall for the R+I- group between the Day 2 reminder test (*M* = 16.1, *SD* = 3.8) and the Day 3 List 1 Forwards test (*M* = 16.4, *SD* = 3.6), *t*(17) = 2.12, *p* = .049, but a significant decrease in recall for the R+I+ group (*M* = 17.6, *SD* = 3.0 and *M* = 14.2, *SD*= 4.4, respectively), *t*(17) = 3.97, *p* = .001. These results suggest that the reminder test did not completely immunize List 1 memory against interference, consistent with the findings of Halamish and Bjork (2011, Experiment 3), though the significant advantage for the R+I+ group over the R-I+ group, and the equivalent final test performance of the R+I+ and R+I- groups, suggest it provided substantial protection.

Discussion

In two experiments we observed a testing effect when participants took a reminder test before learning interfering material. In Experiment 9, taking a reminder test of List 1 before learning List 2, far from disrupting List 1 memory, led to higher final recall scores in the List 1 Forwards test, and in Experiment 10 this benefit also transferred to the List 1 Backwards test. This was not because List 2 was not an effective interfering agent: In Experiment 10, poorer recall in the I+ than the I- group confirmed that List 2 was capable of interfering with memory for List 1 when that memory was first acquired. The other four groups in that experiment, by covering all possible combinations of taking the reminder test and learning List 2, enabled us to examine whether the magnitude of the testing benefit was reduced by the learning of List 2 immediately after the reminder.

Could there have been some disruption to the List 1 memory in the R+I+ group that was offset by the beneficial effect of the extra practice this group received by having the reminder? In fact, performance in the R+I+ group, which took the reminder test and learned List 2, was not significantly poorer than that in the R+I- group, which did not learn List 2, suggesting that testing immunized List 1 memory against interference. Comparison of the reminder test and the final test for the R+I+ and R+I- groups, however, suggests that, although the reminder test provided substantial protection against interference, there remained some detrimental effect of List 2 learning.

In the R-I+ group, learning List 2 reduced List 1 final test performance by nearly half compared with the R-I- group. Is it possible that the presentation, in the List 2 learning task, of cues which were familiar from the List 1 task, acted as a subtle reminder of List 1, reactivating and labilizing List 1 memory and disrupting its subsequent reconsolidation? While some propose that every retrieval triggers reconsolidation (e.g., Sara, 2000), others claim that reconsolidation only occurs when the structure of the reminder meets certain conditions (Forcato et al., 2009). Perhaps a complete test, such as our List 1 reminder test, fails to labilize the memory but a more subtle reminder, the presentation of the shared cues in the List 2 task, does lead to labilization. However, these shared cues were presented in exactly the same way in the List 2 task performed by the R+I+ group, following the reminder test, but this group’s final test performance was no poorer than that of the R+I- group which did not learn List 2. Furthermore, in Experiment 9, List 2 learning disrupted List 1 memory for words with unique cues to the same extent as it did for shared cue words, both when List

2 learning was preceded by a reminder and when it was not, suggesting that the effect on List 1 memory was not due to labilization caused by the presentation of shared cues during List 2 learning. An alternative explanation, discussed below, is that interference from List 2 increased final test difficulty to the extent that many List 1 items became inaccessible. Whatever was the mechanism by which List 2 caused a detriment to List 1 memory in the R-I+ group in the current study, the reminder test in the R+I+ group had a powerful enhancing effect which outweighed the detrimental effect of List 2 learning, protecting List 1 memory against interference.

In Experiment 10, the reminder test provided more benefit in a situation where the List 1 memory was vulnerable to interference than when there was no interference. In fact, there was no difference in performance between the R+I- and R-I- groups, that is, no testing effect when there was no interfering list to learn. Halamish and Bjork (2011, Experiment 3) also found that taking a test could protect memory from interference even when the test provided no benefit when it was not followed by interference. In their study learning, testing, and interference all took place in the same session. Our findings suggest that testing can confer a similar benefit a day after learning, when memory has had time to consolidate.

Why did we find no testing effect for the R+I- group when compared with the R-I- group and why did the reminder test protect List 1 memory from interference even when it provided no enhancement to memory in the absence of interference? The conditions in which retrieval practice is beneficial are determined by many factors including the timing and spacing of tests and the nature of the material studied between tests. Testing is most effective at a point when items are still sufficiently accessible to be recalled on the test but are in danger of becoming less accessible without a test. When items are still easily recallable, a test confers little or no benefit (see, for example, Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008). In our study, List 1 learning involved many test cycles, which may have ensured there was little forgetting during the following 48 hours in the R-I- group even in the absence of retrieval practice, and therefore there was no advantage of taking an intervening test for the R+I- group. Consistent with Halamish and Bjork's (2011) framework, the reminder test may increase the memory strength of the items successfully recalled on the test but if these items are already overlearned, this may not translate into a higher level of final test performance for the tested (R+I-) versus the untested group (R-I-). Halamish and Bjork (2011) proposed that testing will confer a greater benefit in the presence of any factor which makes the final test more difficult, such as final test format, test delay or, in our study, the learning of the interfering List 2. Interference from List 2 makes recall of List 1 at final test more difficult. Put differently, items need to be higher in memory strength in order to overcome the interference and exceed the threshold for retrieval. With no reminder test, several items fail to reach this threshold, leading to the dramatic cut in recall we see in the R-I+ group compared with the R-I- group, for whom the final test is easier because there is no interference from List 2. By contrast, the reminder test increases the strength of the retrieved items in both the R+I+ and R+I- groups. If items are already overlearned, as they are likely to be following our thorough Day 1 training regime, this increase in memory strength does not translate into *more* items recalled on Day 3 in the R+I- group than in the R-I- group, it simply maintains at the same level the number of items recalled. In the R+I+ group, however, the boost to retrieval strength induced by the reminder test ensures that, even though the final test is made more difficult by interference from List 2, items are strong enough to exceed the threshold for retrieval on Day 3. Thus, a reminder test which is unnecessary when there is no following interference, becomes crucially important when there is.

An example of a similar interaction between testing and interference comes from a study by Storm, Bjork, and Storm (2010) who found an advantage for an expanding over a uniform schedule of retrieval practice only when the practiced material was made vulnerable to forgetting by the interpolation of interfering material. They suggested this was because expanding retrieval practice involved an earlier test than uniform practice. When the interpolated material was non-interfering there was little forgetting before the first test with either schedule, but the earlier test became crucially important in the presence of interference as it allowed more items to remain accessible. Similarly, our reminder test only became important when the interfering List 2 was introduced, making List 1 vulnerable to forgetting. In this situation, the reminder test enabled List 1 to remain accessible despite the negative effects of interference.

We therefore found no evidence that taking a reminder test a day after learning led to any detriment to List 1 memory when tested on the third day. Rather, testing had an immunizing effect against interference. If reconsolidation occurs every time a memory is retrieved (e.g., Sara, 2000; Hardt et al., 2010) we might have expected to see some detriment to List 1 memory when retrieval was followed by interfering learning, but we did not. It is possible, however, that List 1 memory was not labilized by our reminder test and there was consequently no need for reconsolidation, consistent with an alternative proposal that retrieval does not always lead to labilization of a memory (Forcato et al., 2009). From a practical point of view, our findings suggest that, in a situation in which one is learning new information with the potential to interfere with previously-learned material, a brief test of the latter may help to keep that memory accessible with little or no detriment to the new learning.

CHAPTER NINE: GENERAL DISCUSSION

The aim of this thesis was to examine whether the well-known benefits of generation and retrieval are sufficient to outweigh the negative effects of interference, in two scenarios in which it has been proposed that testing can be harmful to memory because it increases susceptibility to interference. The first scenario involves tests in which many errors are generated, which have the potential to interfere with memory for correct information, and the second is the situation in which retrieval of a consolidated memory is thought to render it labile and vulnerable to interference from the learning of new information.

Experiments 1-8 examined whether generating errors was helpful or harmful to memory when the cues were novel items and there was no pre-existing relationship between the cue and the target in participants' minds. We consistently observed a benefit of generating errors over reading correct responses, except in Experiment 8 where generating and reading led to equivalent performance, despite correct answers being available for a much shorter time in the Generate condition. The benefit of errorful generation was observed whether the final test was in multiple choice format (Experiments 1-4) or cued recall format (Experiments 6 and 7), whether study of correct answers was experimenter-paced (Experiments 1-3 and 6-7), or self-paced (Experiments 3 and 4), and including conditions where it was possible for the participant to respond, at final test, with an error generated at study (Experiments 3, 6, and 7).

Experiments 9 and 10 examined whether reactivating a previously consolidated memory, in the form of a test taken a day after learning, would make that memory more susceptible to interference from new learning following the reactivation, in line with the reconsolidation hypothesis. In contrast, we found that testing immunized memory against interference.

Testing, and interference from errors: The benefit of generating errors during learning

Experiments 1-8 provide evidence that taking a test can be beneficial to memory even when the test produces many errors, and even when those errors are unrelated to the correct answers presented as feedback. Generating responses, even when they were nearly all errors, enabled corrective feedback to be processed more effectively than when cue and target were presented intact for passive study. Although there was evidence that participants remembered their own generated responses, the benefit of generation on the encoding of corrective feedback was sufficient to outweigh any negative effect of interference from those erroneous responses.

Theoretical accounts of the benefits of errorful generation

Strengthening of a pre-existing association

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 (Hays et al., 2013; Huelser & Metcalfe, 2012) have also argued that an existing cue-target relationship is necessary for this benefit to be observed. These explanations all assume that the benefit of errorful generation occurs because the correct answer, or items with a pre-existing association to it, are activated and strengthened during the generation attempt and that this strengthening enables corrective feedback to be encoded more effectively. 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, and even when these errors are unrelated to the target, by making the encoding of subsequent feedback more effective than when it is not preceded by a generation attempt. Of course, this is not to say that semantic relatedness between cue and target is not helpful to memory. When cue and target are related, activation of the target or other concepts associated to it may well enhance encoding of corrective feedback in the way suggested by Grimaldi and Karpicke (2012), consistent with their findings of a generation benefit for semantically related, but not unrelated, pairs (see also Huelser and Metcalfe, 2012). While semantic relatedness may be helpful to memory when incorrect responses are generated, our findings suggest that it is not a necessary condition for the errorful generation benefit to be observed.

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 the experiments presented here, their procedure 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, which are strongly related to the cue but unrelated to the target, are likely to have come to mind again at final test, on presentation of the cue, and may therefore have interfered with retrieval of the designated "correct" but unrelated answer. For example, a participant may generate "*sand*" in response to the cue "*beach*" and then be told that the target is "*window*". Presentation of "*beach*" at test may elicit "*sand*" because of its strong pre-existing association, and its recent activation, and this may be sufficient to interfere with retrieval of the unrelated "*window*" even if encoding of "*window*" has also been enhanced by generation. In the experiments reported here 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. 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. By using novel vocabulary items, which had no pre-existing associations in participants' minds, we were able 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 have been able to show that the errorful generation benefit does not depend solely on a strengthening of pre-existing associations that enhance memory as a result of also being associated to the target, though this mechanism undoubtedly plays a part in some situations. Instead, the act of generation per se, regardless of what is generated, is able to make the encoding of subsequent feedback more effective. This is important from a theoretical perspective but it is also important from a practical point of view, since our task is closer to the kind of learning situations people are likely to encounter in educational settings.

Future research could investigate the role of semantic relatedness by including, within the same experiment, word pairs studied under three different cue conditions: related (e.g., *tree - leaf),* unrelated (*pillow-leaf*) and novel (where the cue is a foreign language or non word, e.g., *plitsu-leaf*). If pre-existing semantic relatedness between cue and target is helpful to the errorful generation benefit, then we would expect to see a greater benefit for related than for novel pairs. If semantic activation of information which is related to the cue but not to the target interferes with memory to the extent that it outweighs any other benefit of generating a guess, then we would expect to see a similar benefit for novel and for related pairs, but no benefit for unrelated pairs. If semantic activation of information related to the cue is neither helpful nor harmful to memory, having no effect on the benefit of errorful generation, then we would expect to see a similar errorful generation benefit for all types of pair - related, unrelated and novel, though this last possibility seems unlikely, given the findings of Grimaldi and Karpicke (2012) and Huelser and Metcalfe (2012).

Enhanced attention to corrective feedback - the role of curiosity and surprise

Generating errors 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.

One possibility 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. Similarly, generating may evoke more curiosity than reading, where the correct answer is on display from the start. If heightened curiosity leads to better recall, it is possible that one reason why Generate items are better remembered than Read items is that generating guesses increases curiosity.

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).

It will be interesting for future research to examine the role of curiosity, or motivation, in the errorful generation task. For Generate items, participants could be asked to rate their curiosity to learn the correct answer, giving their rating either before or after generating their guess, but before seeing corrective feedback. For Read items, they would see the cue, give their rating, and then be shown the correct answer. If generating a guess increases curiosity to learn the correct answer, curiosity ratings given after generating a guess will be higher than those given before a guess, and higher than for Read items, whereas curiosity ratings will be similar for Read and Generate items when the rating is given immediately following the cue, since no guess has yet been generated. On the other hand, generating would be predicted to lead to superior test performance than reading, regardless of when the curiosity rating is made, consistent with our previous findings. When ratings are made immediately before processing correct answers (i.e., after a guess for a Generate item or after the cue for a Read item) a positive correlation between test performance and curiosity ratings would be predicted.

A further condition in which there is a delay between presentation of the cue and the request for a curiosity rating in the Read condition will enable measurement of the extent to which curiosity is increased simply by having to wait longer for the answer to be displayed, regardless of whether or not an error is produced. If curiosity is increased by waiting, curiosity ratings will be higher in the Read condition after a delay than when given immediately. If curiosity leads to memory enhancement, then higher curiosity ratings in the Read condition will correspond to higher test performance. If generating an error increases curiosity more than waiting does then curiosity ratings will be higher following a generated response than following a delay in the Read condition, and if curiosity leads to enhanced memory, then test performance should also be higher for these Generate items. If, on the other hand, curiosity ratings do not differ in these two conditions but test performance does, this would suggest that the benefit of generating errors cannot be attributed solely to curiosity.

Generation of additional retrieval cues

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 in the “related” conditions used by 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.

Certainly, there was evidence that participants often remembered the guesses they had generated at study, suggesting that they were attending to the cue and trying to find some association with it, rather than generating random guesses regardless of the cue. In Experiments 3 and 6, the final test format enabled participants to enter, at test, a response they had generated or chosen at study. In Experiment 3, the lures in the multiple choice final test consisted of true but incorrect definitions from the study phase, borrowed from other items in the experiment, and incorrect responses that participants had generated or chosen at study. In that experiment, when participants chose an incorrect answer in the final MCQ test, they were very good at rejecting their own generated responses, selecting them at a rate lower than chance, and at a significantly lower rate than they selected their incorrect choices. Incorrect choices were selected no more often than the chance rate. In contrast, Experiment 6 produced a different pattern of results. In that experiment, the final test was in cued recall format and the Lure group was presented with two choices, consisting of the initial letter of the generated or chosen response and the initial letter of the target. Participants reproduced, at final test, the same incorrect response they had given at study nearly 10% of the time and at a significantly *higher* rate than they reproduced incorrect study choices.

These data suggest that incorrect generated responses were better remembered than incorrect choices. In a multiple choice test, the items are all presented on screen, so participants do not have to retrieve the correct definition from memory. Instead, the task becomes one of identifying the source of the available options in order to distinguish the target from the lures. In Experiment 3, this meant distinguishing the target from other familiar definitions shown in the study phase, as well as incorrect choices seen and selected at study, and incorrect guesses generated by the participant. Most often the participant was able to identify the definition which was matched with the cue at study, but when an incorrect response was ultimately chosen at test, participants were able to distinguish their own generated guesses from the other lures, enabling them to reject their own study responses at a rate higher than chance. In contrast, they selected their incorrect choices at chance rates. This suggests that incorrectly generated guesses were better remembered, and recognised as self-generated, than incorrect choices.

The cued recall test of Experiment 6 provides less cue support. Here the participant has to retrieve items from memory, with only a choice of two alternative initial letters to constrain selection. The finding that, when participants repeated an error made at test, it was more often a generated guess than an incorrect choice suggests that generation strengthened the generated error more than incorrect choosing strengthened the incorrect choice, making it more accessible at the final test. Despite these intrusions from study errors in the Generate condition, generating still produced superior memory performance to reading and choosing overall. This could be because, as described above, generating enhances the encoding of corrective feedback by means of increased attention to the feedback, or because participants are able to use their generated guess as an extra retrieval cue at final test.

To investigate further the possibility that generated guesses, even when unrelated to the target, may benefit memory by providing extra retrieval cues which assist with retrieval or recognition at final test, a series of experiments could be designed to examine the extent to which a response generated by the participant at study may assist retrieval of the correct response at test by acting as a mediator. There are several ways this question could be addressed. Pyc and Rawson (2010) note that, for a mediator to benefit memory, two factors are important: the mediator must be recallable in response to the cue, and, having been recalled, it must be able to elicit the target from memory (see also Dunlosky, Hertzog, & Powell-Moman, 2005, for other factors affecting the usefulness of mediators). An experiment which eliminates the need for the mediator to be recalled from memory should therefore produce improved test performance as long as the mediator is effective in eliciting the target. Our task could be modified such that, at final test, half of the Generate items are presented with the response the participant generated at study for that item. If the generated response can assist with retrieval of the target by acting as a mediator, correct recall of targets should be better for the items for which the generated response is provided than for those for which it is not. Of course, if participants were already recalling their own responses at a high rate, providing them might not increase test scores significantly.

Another approach could be to present participants, at final test, with a list of cues, a list of items they had generated at study, and a list of correct targets and ask them to match them. If, in cases where participants matched targets correctly to their cues, they correctly matched them to their associated generated responses at a rate no higher than chance, this would enable us to rule out the explanation that the errorful generation benefit is due to the generated response acting as a mediator. If, on the other hand, correct cue-target pairs were also correctly matched to generated responses at a rate higher than chance, this would be consistent with the mediator account, but not conclusive evidence since good memory for the association between the generated response and the cue or target may be correlated with good memory for the cue-target association without causing it.

Therefore, if evidence consistent with a mediator account is found, a further experiment could examine whether the benefit of errorful generation can be wholly accounted for by the fact that it involves the production of a generated response which can be used as a mediator. Participants would experience a modified version of the Read condition in which, after studying both cue and target, they would be asked to give a response which they might have generated for that cue had they not been shown the target (or a response suitable for use as a MCQ test lure). If generating something acts as a mediator, then this modified Read condition should lead to better performance than the standard Read condition. If the mediator is the *only* mechanism by which generating produces better performance than reading, then the modified Read condition should lead to similar performance to the Generate condition. If, however, the generation benefit is at least partly attributable to some other factor, e.g., curiosity or surprise, then memory for Generate items will still be superior to that for items studied in the modified Read condition.

Metacognition in the errorful generation task

This is the first set of errorful generation experiments 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. The 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. Moreover, 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 than to Read or Choice items, even when, in Experiments 3 and 4, 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 used in their experiment (two Read conditions of different presentation durations, and a Generate condition) 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 the present 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-test or post-study 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 Experiment 2B and Experiment 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, across Experiments 2-4, 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.

Experiment 5, which used only Choice conditions at study, extended our findings by showing that simply being correct was not sufficient to elicit higher JOLs than for incorrect choices - only selection of the target itself, and not correct selection of "other" (in the TA condition), led to higher JOLs than selection of lures. This adds further support to the proposal that (fortuitously) correct selection of the target induces a feeling of familiarity when that target appears as feedback, and this leads to higher JOLs for these items. Participants may even experience an "illusion of knowing", as discussed in Chapter 3. It is interesting that there was no difference in final test performance between items correctly guessed at study in the STD condition and those correctly given an "other" response in the TA condition, in spite of the extra exposure correct targets in the STD condition have at study.

This is the first errorful generation study to include a Choice condition, and the first to collect JOLs. 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 happen 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 (Experiments

2B and 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 & 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.

Format of the final test

Grimaldi and Karpicke (2012), Huelser and Metcalfe (2012) and Hays, Kornell, and Bjork (2012) all used final tests in cued recall format. In our experiments, we consistently observed a benefit of generating over reading when a multiple choice test was used (Experiments 1-4), even when the test enabled selection of erroneous responses generated at study (Experiment 3). With a cued recall test, there was a clear benefit of generating in Experiment 6, and a numerical advantage in Experiment 7, which reached significance after exclusion of two outliers. In Experiment 8, however, there was no difference between generating and reading. It may be that correct answers are strengthened to a greater extent for Generate than for Read items, leading to the consistent benefit observed on the multiple choice tests in Experiments 1-4, but that this strengthening is not always sufficient to lead to a benefit when the task is to retrieve the item from memory, not simply to recognise it.

Furthermore, in the three experiments which used a recall test (Experiments 6-8), correct answers for Generate items were presented for a much shorter time at study than correct answers for Read items were. Any boost to memory strength due to generating may therefore have been offset by an advantage for Read items in terms of the additional time available for study of correct answers. The larger sample used in Experiment 6 may have enabled the Generate advantage to emerge in spite of this. By using the multiple choice format final test in Experiments 1-4, we were able to observe a clear benefit of errorful generation which may not always be so easy to detect with a cued recall test.

Timing of the final test, and timing of feedback

Experiments 1- 8 all used a short retention interval, with just one minute's filled delay between the study phase and the test phase. An important question for future research is whether the errorful generation benefit will persist over a longer retention interval of, for example, a day or a week. While Kane and Anderson (1978) found a benefit of errorful generation at a retention interval of one week, they also found evidence of an increase in interference from errors generated at study. One explanation for the benefit of errorless learning in memory-impaired populations is that memory-impaired people have difficulty identifying the source of their memories. An incorrect item they have generated themselves is strengthened more than an item provided by the experimenter, and comes to mind more readily at final test. Participants with poor source memory are unable to identify it as self-generated, leading to perseverative errors at test. With healthy young participants, source memory may decline over long retention intervals, e.g., of a week, so it is possible that the errorful generation advantage may be reduced or eliminated in these circumstances.

Another important issue for future research is whether the errorful generation benefit depends on corrective feedback being provided immediately following the generation of a response at study. Vaughn and Rawson (2012) found a benefit of errorful generation for weakly associated word pairs when feedback was given immediately after generation (Experiment 2) but not when it was given after a delay in which other items were presented for generation (Experiment 1). On the other hand, studies in which participants are questioned before studying a text inevitably involve a delay between generating responses to the questions and the subsequent studying of the text, yet these studies have found a benefit of incorrect guessing even under these circumstances (e.g., Pressley, Tanenbaum, McDaniel, & Wood, 1990; Richland, Kornell & Kao, 2009). Of course, the nature of the materials is very different in these cases. With text materials, the benefit may arise because questions presented before the text is studied draw participants' attention to testable content in the text. It will be important for future research to try to identify the optimal timing of feedback for the learning of different kinds of material.

Errorless learning

Experiment 8 was designed to bring our task closer to the typical task used in "errorless learning" studies, in an effort to explore why those studies so often show a benefit of passive studying over generating errors while our experiments consistently showed the opposite effect. As discussed in Chapter 7, there are many differences between our task and the typical errorless learning tasks. A difference which may be important as far as these differing outcomes are concerned is that the typical errorless learning task encourages the generation of many errors per item, both by eliciting many errors on each trial and also by having participants go through several cycles of learning for each item. The task used in Experiments

1-4 and 6-7 elicits only one error per item, and each item is studied just once. Experiment 8 was designed to explore whether encouraging generation of multiple errors, by having participants experience several study cycles, would yield a benefit of reading over generating rather than the opposite pattern observed in our other experiments. Unfortunately, in this experiment the Generate benefit did not emerge even in the one-cycle group. However, there was no evidence that increasing the number of cycles, and therefore the number of errors generated for each Generate item, led to an increasing benefit in favour of the Read condition: On the contrary, for the three-cycle group there was a numerical advantage for generating. One important avenue for future research will be to increase the number of errors participants make on each trial, as well as having them go through several cycles of learning, to increase the opportunity for erroneous responses to interfere with recognition or recall of the correct answer at test.

Other possibilities, discussed in more detail in Chapter 7, include constraining responses by providing the initial letter or letters of the target. Generating multiple responses per trial, with the same initial letter, would increase phonological similarity between errors and the target, potentially creating interference with learning the correct response. The subjective plausibility of responses may also be an important factor. In our task participants may not make a strong association between the cue and their own generated response because they are highly unlikely to be correct. Constraining responses, either by providing one or more letters, or by providing a semantic cue, such as a category, would increase the probability that the participant will guess the target, and may lead participants to make a stronger association between their own error and the cue.

Finally, while some errorless learning studies have included healthy young people among their participants, often their main focus is on people with memory impairments or healthy older adults. It would be interesting to examine the performance of, for example, healthy older adults on the errorful generation task used in our experiments, to determine if errorful generation is as beneficial for that population as it has been shown to be with our, typically younger, participants.

Errorful generation: Benefits and costs

Experiments 1-8 found no detriment of testing when it produced many errors. On the contrary, generating typically produced superior memory performance to errorless reading. Even tests which involved choosing never led to a detriment from making errors, by comparison with reading, and choosing led to superior performance when final test options were all equally familiar and when all items were considered, whether correct or incorrect at study (Experiments 3 and 4). On the other hand, testing led to metacognitive errors, in that participants were strongly influenced by whether or not they happened to guess correctly at study, leading to erroneous predictions about their future performance. Given that metacognitive judgments can affect learners' study strategies, the misperception that generating is less beneficial to memory than passive studying, may lead to underuse of a potentially powerful learning tool. Methods to attenuate this misperception are worth exploring in future research.

Testing and interference from new learning

Experiments 9 and 10 examined whether reactivating a memory in the form of a test taken a day after learning would make it more susceptible to interference. In Experiment 9 testing, far from creating a detriment to List 1 memory, enhanced it, showing a testing effect. In Experiment 10, interference, in the form of List 2 learning, had a substantial detrimental effect on memory for List 1, reducing recall of List 1 by nearly half compared with when there was no List 2 learning. However, a single test of List 1, without feedback, on Day 2 eliminated that detriment and protected List 1 memory against interference from the learning of List 2. There was no evidence that reactivating List 1 memory, then learning List 2, disrupted reconsolidation of List 1.

Few studies have investigated reconsolidation in human memory, and findings have been mixed among those which have. Some researchers (e.g., Forcato et al., 2009) have proposed that labilization of consolidated memories only occurs when a memory is reactivated without being fully retrieved, in other words when there is a mismatch between what is expected and what occurs. However, others have reported reconsolidation effects following a full test of the memory (Chan & LaPaglia, 2013; Finn & Roediger, 2011; Walker et al., 2003), though in Finn and Roediger’s study retrieval occurred so soon after learning that it is debatable whether consolidation, let alone reconsolidation, had had time to occur. Our findings suggest that a full test, even without feedback, protects the memory from the harmful effects of interference from new learning.

Testing and interference: Application to educational settings

Testing is increasingly being advocated as a means of enhancing learning in educational settings (e.g., Pashler, Bain, Bottge, Graesser, McDaniel, & Metcalfe, 2007). To maximize its usefulness it is important to identify in what conditions testing is most beneficial, and when it might be harmful to memory. The discovery that generation can be beneficial to memory even when it produces many errors and even when there is no pre-existing association between the cue and the target (Experiments 1-8), 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.

A few studies have examined testing effects in real classroom settings, though there are considerable difficulties in carrying out this kind of applied research (see, for example, Agarwal, Bain, & Chamberlain, 2012). One study (McDaniel, Agarwal, Huelser, McDermott, & Roediger, 2011) investigated the benefits of testing in an American middle school setting, and included a condition in which students were given a multiple choice test before the teacher's lesson on the tested topic (pre-lesson test), but after reading the relevant textbook chapter. This condition therefore has some, though not exact, similarity with the Choice condition used in the errorful generation experiments reported here, though it differs in terms of the materials used and in the fact that pupils had the opportunity to read about the topic in the textbook before taking the test. McDaniel et al. found that there was neither a benefit nor a detriment of pre-lesson testing, compared with giving no test, on end-of-unit summative test performance up to three weeks later, nor on summative tests given 3 months (end of semester) or 8 months (end of year) after initial learning, although this last result was marginally significant in favour of the pre-lesson test.

While a handful of classroom studies have incorporated, in their design, pre-lecture tests in short answer format rather than MCQ format (e.g., Narloch, Garbin, & Turnage, 2006; Nevid & Mahon, 2009), these studies have tended to focus on indirect benefits of pre-lecture testing, such as motivating students to do the assigned reading, activating relevant schemas prior to the lecture, and focussing attention on key concepts subsequently covered in the lecture. For example, Narloch, Garbin, and Turnage (2006) examined the efficacy of giving students pre-lecture tests based on assigned reading, and found a benefit for tested over nontested groups in terms of later performance on both multiple choice and essay exam questions. Students' self-reports of their preparation time for the pre-lecture tests suggest that at least part of this benefit was due to more time spent on the assigned reading in the tested groups, and the design of the study makes it impossible to isolate the effect of generating responses per se. While indirect benefits of tests such as students' increased engagement with pre-lecture assigned readings are undoubtedly important, it would be interesting to investigate whether the generation benefit observed in the present experiments, where material has not been previously studied, would transfer to applied settings. No study, to my knowledge, has yet addressed this.

One line of research, with some similarity to the present one, has examined the benefit of "productive failure" in the learning of novel mathematics concepts (see Kapur & Bielaczyk, 2011, for review). In one study (Kapur, 2012), school children learned the concept of variance either by "direct instruction" (DI, i.e., being taught the canonical solution by the teacher using worked examples, and then being given data analysis problems to solve) or, in the "productive failure" (PF) condition, worked in groups to invent a method for solving those same data analysis problems, prior to being taught the canonical solution. The productive failure group offered many solutions based on their prior knowledge, but not the canonical one which they were subsequently taught. On a subsequent test, however, they outperformed the direct instruction group in both conceptual understanding of the solution, and ability to apply it to new problems. This was in spite of the fact that the study design meant that the DI group had had the opportunity to practise more data analysis problems overall. While there are many differences between Kapur's study and the experiments reported here, his findings suggest a real benefit of incorrect generation in the classroom, which will be important to follow up using a similar task to the one used in our experiments.

Similarly, the finding (in Experiments 9 and 10) that testing can protect memories against interference from new learning which otherwise could disrupt even well-learned memories, is of real practical value. When learning new material with some similarity to previously learned information, learners would do well to undergo a brief rehearsal of the original material before embarking on learning the new information, and need not fear that this will lead to increased confusion between the two.

Concluding remarks

This thesis examined whether testing, which is known to enhance memory in many situations, would be helpful or harmful to memory in cases where the test itself has the potential to increase interference, a factor which is often detrimental to memory. In fact, the evidence from the experiments reported here suggests that the benefit of testing is remarkably robust in the face of interference, at least in the two scenarios considered in this thesis.

Experiments 1-8 showed a consistent benefit of generation, even when it produced many errors, and even when the opportunity for participants to persist with their errors was maximized. That this was so even though items to be learned were unfamiliar, challenges the proposal that generating errors during learning can only benefit memory when there is a pre-existing association between cue and target which can be reinforced by corrective feedback. While the reinforcement of pre-existing associations between cue and target may play an important role in the classic testing effect, and in some errorful generation situations, the benefit we observed during the learning of novel items must be attributable to a different mechanism, and determining the exact nature of that mechanism is an important goal for future research. One disadvantage to testing, though, was in its effect on participants' metacognitive judgments, which were unduly influenced by the outcome of fortuitous guesses.

Likewise, the reconsolidation hypothesis predicts modification or unlearning when a memory is reactivated immediately prior to interfering new learning. In contrast, Experiments 9-10 not only found no detriment to memory from testing under these conditions, but found a benefit such that testing immunized the memory against interference from the new learning. Indeed, testing conferred a greater benefit in the presence of interference than in its absence.

REFERENCES

Agarwal, P. K., Bain, P. M., & Chamberlain, R. W. (2012). The value of applied research: retrieval practice improves classroom learning and recommendations from a teacher, a principal, and a scientist. *Educational Psychology Review, 24,* 437–448.

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.

Anderson, M. C., Bjork, R.A., & Bjork, E. L. (1994). Remembering can cause forgetting: retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1063–1087.

Arnold, K. M., & McDermott, K. B. (2013). Test-potentiated learning: Distinguishing between direct and indirect effects of tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39,* 940–945.

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

Barnes, B. M., & Underwood, R. J. (1959). The fate of 1st list associations in transfer theory. *Journal of Experimental Psychology, 58,* 97–105.

Begg, I., Vinski, E., Frankovich, L., & Holgate, B. (1991). Generating makes words memorable, but so does effective reading. *Memory and Cognition, 19,* 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, 7,* 750–759.

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

Brewer, G. A., and Unsworth, N. (2012). Individual differences in the effects of retrieval from long-term memory. *Journal of Memory and Language, 66,* 407–415.

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

Butler, A. C., Karpicke, J.D, & Roediger, H. L. III. (2008). Correcting a metacognitive error: feedback increases retention of low-confidence correct responses. *Journal of Experimental Psychology, Learning, Memory, and Cognition, 34,* 918 –928.

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,* 604–616.

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

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

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,* 268–276.

Carpenter, S. K., Pashler, H., Wixted, J. T., & Vul, E. (2008). The effects of tests on learning and forgetting**.** *Memory & Cognition, 36,* 438–448.

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,* 107–111.

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning. A temporal ridgeline of optimal retention. *Psychological Science, 19,* 1095–1102.

Chan J. C. K., & LaPaglia, J. A. (2013). Impairing existing declarative memory in humans by disrupting reconsolidation. *Proc. Natl. Acad. Sci. USA, 110,* 9309-9313.

Chan, J. C. K., Thomas, A. K., & Bulevich, J. B. (2009). Recalling a witnessed event increases eyewitness suggestibility: the reversed testing effect. *Psychological Science, 20,* 66–73.

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

Conover, W. J., & Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician, 35,* 124–129.

Cull, W. L. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology, 14,* 215–235.

Dunlosky, J., Hertzog, C., & Powell-Moman, A. (2005). The contribution of mediator-based deficiencies to age differences in associative learning. *Developmental Psychology, 41,* 389 –400.

Dunlosky, J., Serra, M. J., Matvey, G., & Rawson, K. A. (2005). Second order judgments about judgments of learning. *Journal of General Psychology, 132,* 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,* 470–482.

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

Finn, B., & Roediger, H.L.III. (2011). Enhancing retention through reconsolidation: negative emotional arousal following retrieval enhances later recall. *Psychological Science, 22,* 781–786.

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

Forcato, C., Argibay, P. F., Pedreira, M. E., & Maldonado, H. (2009). Human reconsolidation does not always occur when a memory is retrieved: the relevance of the reminder structure. *Neurobiology of Learning and Memory, 91,* 50–57.

Forcato, C., Burgos, V. L., Argibay, P. F., Molina, V.A., Pedreira, M. E., & Maldonado, H. (2007). Reconsolidation of declarative memory in humans. *Learning and Memory, 14,* 295–303.

Forcato, C., Rodriguez, M.L.C., Pedreira, M. E., & Maldonado, H. (2010). Reconsolidation in humans opens up declarative memory to the entrance of new information. *Neurobiology of Learning and Memory, 93,* 77–84.

Forlano, G., & Hoffman, N. M. H. (1937). Guessing and telling methods in learning words of a foreign language. *Journal of Educational Psychology, 28*, 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*, 392 –399.

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

Halamish, V., & Bjork, R.A. (2011). When does testing enhance retention? A distribution-based interpretation of retrieval as a memory modifier. *Journal of Experimental Psychology: Learning, Memory and Cognition, 37,* 801–812.

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.

Hardt, O., Einarsson, E.I., & Nader, K. (2010). A bridge over troubled water: reconsolidation as a link between cognitive and neuroscientific memory traditions. *Annual Review of Psychology, 61,* 141–167.

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,* 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*, 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*, 22–34.

Hirshman, E., & Bjork, R. A. (1988). The generation effect: support for a 2-factor theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 484–494.

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,* 514–527.

Hupbach, A., Gomez, R., Hardt, O. & Nadel, L. (2007). Reconsolidation of episodic memories: a subtle reminder triggers integration of new information. *Learning and Memory, 14,* 47–53.

Hupbach, A., Gomez, R., & Nadel, L. (2009). Episodic memory reconsolidation: updating or source confusion? *Memory, 17,* 502–510.

Hupbach, A., Hardt, O., Gomez, R., & Nadel, L. (2008). The dynamics of memory: context-dependent updating. *Learning and Memor,* 574–579.

Izawa, C. (1970). Optimal potentiating effects and forgetting-prevention effects of tests in paired-associate learning. *Journal of Experimental Psychology, 83,* 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,* 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*, 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*, 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*, 48–59.

Kapur, M. (2012). Productive failure in learning the concept of variance. *Instructional Science, 40,* 651–672.

Kapur, M., & Bielaczyk, K. (2011). Designing for productive failure. *Journal of the Learning Sciences, 21,* 45–83.

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,* 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*, 416–428.

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: a distribution based bifurcation model. *Journal of Memory and Language, 65,* 85–97.

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

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*, 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*, 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*, 494–513.

McDaniel, M. A., Riegler, G. L. & Waddill, P. J. (1990). Generation effects in free recall: further support for a 3-factor theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 789–798.

McDaniel, M. A., & Masson, M. E. J. (1985). Altering memory representations through retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11,* 371–385.

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*, 18–26.

McDermott, K.B. (2006). The paradoxical effects of testing: repeated retrieval attempts enhance the likelihood of later accurate and false recall. *Memory & Cognition, 34,* 261–267.

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*, 249–259.

Narloch, R., Garbin, C. P., & Turnage, K. D. (2006). Benefits of prelecture quizzes. *Teaching of Psychology, 33,* 109–112.

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*, 269–273.

Nevid, J. S., & Mahon, K. (2009). Mastery quizzing as a signaling device to cue attention to lecture material. *Teaching and Learning of Psychology, 36,* 29–32.

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). Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education.

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,* 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*, 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*, 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*, 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*, 1155–1159.

Rubin, R. D. (1976). Clinical use of retrograde amnesia produced by electroconvulsive shock: a conditioning hypothesis. *Canadian Psychiatric Association Journal, 21,* 87–90.

Sara, S. (2000). Strengthening the shaky trace through retrieval. *Nature Reviews Neuroscience, 1,* 212–213.

Schooler, J. W., Foster, R. A., & Loftus, E. F. (1988). Some deleterious consequences of the act of recollection. *Memory and Cognition, 16,* 243–251.

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

Slamecka, N. J., & Graf, P. (1978). The generation effect: delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory, 4,* 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,* 589–607.

Storm, B. C., Bjork, R. A., & Storm, J. C. (2010). Optimizing retrieval as a learning event: When and why expanding retrieval practice enhances long-term retention. *Memory & Cognition, 38,* 244–253.

Szpunar, K.K., McDermott, K.B., & Roediger, H.L.III. (2008). Testing during study insulates against the build up of proactive interference. *Journal of Experimental Psychology: Learning, Memory and Cognition, 34,* 1392–1399.

Thomas, A. K., & McDaniel, M.A. (2007). The negative cascade of incongruent generative study-test processing in memory and metacognition. *Memory and Cognition, 35,* 668–678.

Thomas, R. C., & McDaniel, M. A. (2013). Testing and feedback effects on front-end control over later retrieval. *Journal of Experimental Psychology: Learning, Memory and Cognition, 39,* 437–450.

Toppino, T. C., & Cohen, M. S. (2009). The testing effect and the retention interval: questions and answers. *Experimental Psychology 56,* 252–257.

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.

Walker, M. P., Brakefield, T., Hobson, J. A., & Stickgold, R. (2003). Dissociable stages of human memory consolidation and reconsolidation. *Nature, 425,* 616–620.

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*, 552–557.

APPENDICES

Appendix A – Stimulus materials for each experiment.

Table A1: Stimulus materials for Experiment 1

Cue	Target	Lure 1	Lure 2	Lure 3
compendious	concise	elegant	overhanging	scholarly
descry	discover	betray	demolish	outline
encomium	tribute	theatre	meeting	committee
esculent	edible	dainty	precise	moist
frond	leaf	edge	group	stream
gossamer	flimsy	greedy	false	huge
immolate	sacrifice	soften	pacify	criticize
imprecation	curse	stain	error	lie
inculcate	instil	insert	enter	fill
interstice	space	pillar	layer	archway
lassitude	tiredness	flexibility	pessimism	gratitude
maunder	ramble	moan	grieve	pester
minatory	threatening	exploratory	insignificant	ridiculous
plangent	noisy	heavy	funny	pleasant
roke	mist	pole	wheel	stick
spoliate	plunder	blemish	disrupt	indulge
sprauncy	smart	jolly	lean	spicy
stentorian	loud	stern	firm	old
subduce	withdraw	bewitch	conclude	underlie
subsecive	spare	later	repeated	jealous
contumacious	rebellious	meddling	spacious	expensive
effulgent	shining	generous	bulging	exuberant
inimical	hostile	unique	notable	impaired
limpid	clear	weak	sticky	supple
objurgate	rebuke	argue	oppress	oppose
opprobrium	disgrace	pleasure	approval	excitement
orotund	pompous	stout	respected	boring
peculate	embezzle	complain	gamble	misbehave
picaroon	cheat	urchin	clown	tourist
recreant	surrendering	hypocritical	unreliable	dishonest
roil	billow	roast	tumble	simmer
slake	quench	attack	slice	climb
spoffish	fussy	proud	cute	young
stanchion	support	dwelling	companion	mentor
stolid	unemotional	unalterable	inevitable	unavoidable
burgeon	flourish	hesitate	reconcile	hurry
subluxation	dislocation	sufficiency	entirety	imbalance
subvention	grant	custom	trick	trial
succursal	branch	journey	revision	overview
superate	overcome	exceed	overestimate	maximise
everse	overthrow	prevent	deny	alternate

blandish	flatter	fortify	promise	condemn
immure	imprison	accustom	secure	abandon
inchoate	undeveloped	independent	unknown	inexplicable
valinch	tube	rope	clamp	screw
intractable	unmanageable	untouchable	inaccessible	inadequate
mendicant	beggar	misfit	tailor	criminal
perpend	consider	foretell	await	delay
pertinacious	unyielding	relevant	discerning	energetic
abjure	reject	ensure	believe	praise
rebarbative	repellent	sarcastic	influential	aggressive
recondite	hidden	needed	guilty	unsure
renitent	resistant	sorrowful	continuous	intuitive
sententious	moralising	controversial	passionate	irritable
sodality	fellowship	misfortune	loneliness	harmony
subdolous	crafty	gloomy	naughty	silly
surquedry	arrogance	abundance	politeness	deceitfulness
threnody	lament	excess	frenzy	fear
trammel	impede	trouble	destroy	mumble
tyro	beginner	dictator	soldier	juggler

Table A2: Stimulus materials for Experiment 2A

Cue	**Target**	**Lure 1**	**Lure 2**	**Lure 3**
menald	spotty	brainy	chatty	drowsy
frampold	quarrelsome	dishonest	insolent	ungrateful
floccose	woolly	crooked	flagrant	husky
hispid	bristly	mundane	soggy	reddish
orotund	pompous	burly	cheery	lavish
leggiadrous	elegant	protective	impersonal	sensible
stentorian	loud	stern	trim	bold
subsecive	extra	welcome	tiny	brilliant
selcouth	strange	fresh	bright	round
gaberlunzie	beggar	culprit	dweller	healer
opprobrium	disgrace	ordeal	pitfall	forfeit
carcanet	necklace	packet	bonnet	slipper
roke	mist	creek	grove	bush
succursal	branch	gift	scheme	drill
achene	fruit	tape	myth	knee
umiak	boat	hill	frame	song
famulus	assistant	scholarship	consumer	illusion
objurgate	rebuke	disperse	begrudge	fiddle
descry	detect	accuse	reject	tremble
subduce	withdraw	predict	concede	devise
rebarbative	repellent	bubbly	victorious	obstinate
spoffish	fussy	bogus	unkind	risky
recreant	cowardly	flamboyant	defiant	mischievous
plangent	noisy	clumsy	dreary	sleepy
esurient	hungry	nervous	fortunate	stupid
infandous	horrible	constructive	reluctant	intensive
sprauncy	smart	lean	brave	dim
coriaceous	tough	wise	calm	glad
clancular	secret	evil	pointed	quiet
surquedry	arrogance	fortitude	merriment	gallantry
desman	mole	vase	peg	toad
mechlin	lace	shrine	herb	chord
kedge	anchor	razor	needle	doorway
quidnunc	gossip	nonsense	moonlight	romance
valinch	tube	bird	horn	maid
gadoid	fish	tongue	bench	yard
stanchion	post	bank	youth	ship
spoliate	plunder	grapple	blemish	dwindle
conspue	despise	inflict	exclude	postpone
trammel	restrict	complain	exert	frighten

morigerous	obedient	meticulous	amicable	inquisitive
levisomnous	observant	nocturnal	expectant	versatile
compendious	concise	devout	weighty	sulky
insulse	boring	aloof	cunning	moody
gallionic	careless	furious	dismal	stormy
minatory	threatening	conclusive	forthcoming	eccentric
recondite	obscure	absurd	clever	weary
glabrous	smooth	rare	odd	mad
tantivy	quick	warm	dry	flat
hauberk	garment	banquet	fortress	puppet
falchion	sword	barge	hood	plank
frisket	mask	glove	cork	pearl
zamindar	landlord	dentist	salesman	clergy
quant	pole	brick	lamp	tray
kaross	jacket	mountain	owner	campus
sodality	fellowship	sympathy	intention	dignity
bistoury	knife	fair	judge	rock
droze	drip	coax	starve	groan
maunder	ramble	amaze	beware	prosper
mundify	wash	hurt	plot	fail

Table A3: Stimulus materials for Experiment 2B

Cue	Target	Lure 1	Lure 2	Lure 3
ahuntz	goat	towel	spoon	jug
aker	ram	hedge	stair	broom
apaiz	priest	sketch	quest	flag
apal	shelf	tar	layer	nerve
basamortu	desert	planet	marble	crossing
bizar	beard	storm	gear	card
danbor	drum	sleeve	rake	cliff
ezti	honey	saddle	parade	cocktail
gerezi	cherry	wallet	peanut	feather
hezur	bone	lunch	prince	grave
hosto	leaf	tank	silk	crest
igel	frog	shrub	wand	gnome
kajoi	drawer	pig	glue	sack
lepoko	necklace	turret	kettle	doorstep
lurrikara	earthquake	turkey	nightmare	hammer
oihan	jungle	suitcase	quarrel	border
opari	gift	scheme	rice	fate
orratz	needle	scholar	reward	anchor
poltsa	bag	pride	wire	skill
sagu	mouse	wool	doll	coin
aldapa	hillside	monkey	salad	berry
arrautza	egg	tin	lens	cab
belarri	ear	stem	jet	barn
dordoka	turtle	wardrobe	pillow	mansion
erbi	hare	fern	crust	plum
erizain	nurse	lion	fleet	dome
erle	bee	tomb	pump	juice
gurdi	cart	mat	hook	flask
gutunazal	envelope	summary	luxury	heritage
hileta	funeral	slavery	harmony	accident
kokotz	chin	shirt	mirror	grain
kometa	kite	flute	crease	hinge
leize	cave	wheat	pearl	mask
hegal	wing	tray	pen	brick
ohitura	custom	verdict	thunder	sunset
soka	rope	thread	lawn	clue
sutondo	fireplace	puppet	garment	banquet
tresna	tool	code	joy	proof
urmael	pond	star	fog	vein
zorro	packet	saucepan	ruler	orchard
ardi	sheep	plug	bowl	flower
arkatz	pencil	sugar	ocean	mistake
bidaia	journey	witness	profit	folklore

borda	shed	log	ghost	calf
eskularru	glove	hose	duck	cork
euritako	umbrella	veranda	hurricane	caravan
gazta	cheese	moss	rack	cage
hodei	cloud	sand	prize	dawn
ijito	gypsy	vulture	treasure	rainbow
iloba	niece	veil	stool	trunk
iturri	fountain	candle	lemon	hunter
katilu	mug	pram	claw	rib
mahats	grape	swan	cloak	thorn
margotu	paint	noise	gate	fool
marraskilo	snail	weed	jewel	skate
oinetako	shoe	purse	bush	hawk
sarbide	driveway	arrow	neighbour	insect
saski	basket	voyage	portrait	meadow
untxi	rabbit	walnut	violin	infant
oilo	hen	plate	mess	soap

Table A4: Stimulus materials for Experiment 3

Cue	Target	Study lure 1	Study lure 2	Study lure 3	Test lure 1	Test lure 2	“New” lure
compendious	concise	devout	weighty	sulky	glassy	obedient	hospitable
frampold	quarrelsome	dishonest	insolent	ungrateful	observant	cowardly	grubby
spoffish	fussy	bogus	unkind	risky	woolly	noisy	sloppy
hispid	bristly	mundane	soggy	reddish	boring	threatening	fragrant
orotund	pompous	burly	cheery	lavish	careless	lasting	discreet
acherontic	dismal	feeble	corrupt	selfish	horrible	smart	impartial
leggiadrous	elegant	protective	impersonal	sensible	obscure	hungry	sturdy
stentorian	loud	stern	trim	bold	extra	tough	distinctive
glabrous	smooth	rare	odd	mad	secret	quick	varied
selcouth	strange	fresh	bright	round	spotty	repellent	impossible
rapparee	bandit	chariot	gateway	steamboat	seaweed	necklace	snowstorm
hauberk	garment	banquet	fortress	puppet	gem	mole	ledge
falchion	sword	barge	hood	plank	jug	herb	prisoner
mechlin	lace	shrine	fold	spear	button	gossip	castle
frisket	mask	glove	cork	pearl	mist	anchor	trifle
zamindar	landlord	dentist	salesman	clergy	bird	pole	mineral
kaross	jacket	mountain	owner	campus	fruit	tube	magic
withe	branch	gift	scheme	drill	fool	fish	plot
hanap	cup	load	sheet	meat	knife	post	wave
karimption	crowd	band	fort	stream	beggar	camel	chest
rebarbative	repellent	bubbly	victorious	obstinate	loud	secret	jovial
morigerous	obedient	meticulous	amiable	inquisitive	smooth	spotty	arrogant
recreant	cowardly	flamboyant	defiant	mischievous	strange	glassy	erratic
plangent	noisy	clumsy	dreary	sleepy	concise	observant	tedious
minatory	threatening	conclusive	forthcoming	eccentric	quarrelsome	woolly	childish
olamic	lasting	blessed	monstrous	sweeping	fussy	boring	suspicious
sprauncy	smart	lean	brave	dim	bristly	careless	lucky
esurient	hungry	nervous	fortunate	stupid	pompous	horrible	unexpected
coriaceous	tough	wise	calm	glad	dismal	obscure	dirty
tantivy	quick	warm	dry	flat	elegant	extra	clean
mehari	camel	almond	fanfare	helmet	branch	knife	goldfish
carcanet	necklace	packet	bonnet	slipper	cup	beggar	orchard
desman	mole	vase	peg	toad	crowd	seaweed	worm
munjeet	herb	booth	grudge	prey	bandit	gem	chimney
quidnunc	gossip	nonsense	moonlight	romance	garment	jug	ribbon
kedge	anchor	razor	needle	doorway	sword	button	onion
quant	pole	brick	lamp	tray	lace	mist	bell
valinch	tube	lane	horn	maid	mask	bird	slave
gadoid	fish	tongue	bench	yard	landlord	fruit	chicken
stanchion	post	bank	youth	ship	jacket	fool	cover
menald	spotty	brainy	chatty	drowsy	quick	dismal	craggy
hyaline	glassy	thorny	spongy	lumpy	repellent	elegant	sickly
levisomnous	observant	nocturnal	expectant	versatile	obedient	loud	grandiose

floccose	woolly	crooked	flagrant	husky	cowardly	smooth	dainty
insulse	boring	aloof	cunning	moody	noisy	strange	greedy
gallionic	careless	furious	kindly	stormy	threatening	concise	merry
infandous	horrible	constructive	reluctant	intensive	lasting	quarrelsome	persistent
recondite	obscure	absurd	clever	weary	smart	fussy	imaginary
subsecive	extra	welcome	tiny	brilliant	hungry	bristly	formal
clancular	secret	evil	pointed	quiet	tough	pompous	official
gaberlunzie	beggar	culprit	dweller	healer	post	landlord	spider
varec	seaweed	ordeal	pitfall	forfeit	camel	jacket	ointment
peridot	gem	dove	marsh	trout	necklace	branch	vine
urceus	jug	wolf	mast	cord	mole	cup	feather
netsuke	button	puzzle	token	harness	herb	crowd	costume
roke	mist	creek	grove	bush	gossip	bandit	pine
barbet	bird	fence	trail	root	anchor	garment	gallery
achene	fruit	tape	myth	knee	pole	sword	soldier
balatron	fool	guide	bomb	gate	tube	lace	palace
bistoury	knife	fair	judge	rock	fish	mask	leader

Table A5: Stimulus materials for Experiment 4

Cue	Target	Study lure 1	Study lure 2	Study lure 3	Test lure 1	Test lure 2	Test lure 3
compendious	concise	devout	weighty	sulky	glassy	obedient	elegant
frampold	quarrelsome	dishonest	insolent	ungrateful	observant	cowardly	fussy
spoffish	fussy	bogus	unkind	risky	woolly	noisy	bristly
hispid	bristly	mundane	soggy	reddish	boring	threatening	pompous
orotund	pompous	burly	cheery	lavish	careless	lasting	quarrelsome
acherontic	dismal	feeble	corrupt	selfish	horrible	smart	concise
leggiadrous	elegant	protective	impersonal	sensible	obscure	hungry	dismal
stentorian	loud	stern	trim	bold	extra	tough	smooth
glabrous	smooth	rare	odd	mad	secret	quick	strange
selcouth	strange	fresh	bright	round	spotty	repellent	loud
rapparee	bandit	chariot	gateway	steamboat	seaweed	necklace	garment
hauberk	garment	banquet	fortress	puppet	gem	mole	sword
falchion	sword	barge	hood	plank	jug	herb	bandit
mechlin	lace	shrine	fold	spear	button	gossip	landlord
frisket	mask	glove	cork	pearl	mist	anchor	lace
zamindar	landlord	dentist	salesman	clergy	bird	pole	mask
kaross	jacket	mountain	owner	campus	fruit	tube	cup
withe	branch	gift	scheme	drill	fool	fish	jacket
hanap	cup	load	sheet	meat	knife	post	crowd
karimption	crowd	band	fort	stream	beggar	camel	branch
rebarbative	repellent	bubbly	victorious	obstinate	smooth	secret	noisy
morigerous	obedient	meticulous	amiable	inquisitive	loud	spotty	cowardly
recreant	cowardly	flamboyant	defiant	mischievous	strange	glassy	repellent
plangent	noisy	clumsy	dreary	sleepy	concise	observant	threatening
minatory	threatening	conclusive	forthcoming	eccentric	quarrelsome	woolly	lasting
olamic	lasting	blessed	monstrous	sweeping	fussy	boring	smart
sprauncy	smart	lean	brave	dim	bristly	careless	quick
esurient	hungry	nervous	fortunate	stupid	pompous	horrible	obedient
coriaceous	tough	wise	calm	glad	dismal	obscure	hungry
tantivy	quick	warm	dry	flat	elegant	extra	tough
mehari	camel	almond	fanfare	helmet	branch	knife	necklace
carcanet	necklace	packet	bonnet	slipper	cup	beggar	herb
desman	mole	vase	peg	toad	crowd	seaweed	camel
munjeet	herb	booth	grudge	prey	bandit	gem	mole
quidnunc	gossip	nonsense	moonlight	romance	garment	jug	anchor
kedge	anchor	razor	needle	doorway	sword	button	pole
quant	pole	brick	lamp	tray	lace	mist	tube
valinch	tube	lane	horn	maid	mask	bird	post
gadoid	fish	tongue	bench	yard	landlord	fruit	gossip
stanchion	post	bank	youth	ship	jacket	fool	fish
menald	spotty	brainy	chatty	drowsy	quick	dismal	woolly
hyaline	glassy	thorny	spongy	lumpy	repellent	elegant	spotty
levisomnous	observant	nocturnal	expectant	versatile	obedient	loud	horrible

floccose	woolly	crooked	flagrant	husky	cowardly	smooth	glassy
insulse	boring	aloof	cunning	moody	noisy	strange	observant
gallionic	careless	furious	kindly	stormy	threatening	concise	boring
infandous	horrible	constructive	reluctant	intensive	lasting	quarrelsome	careless
recondite	obscure	absurd	clever	weary	smart	fussy	extra
subsecive	extra	welcome	tiny	brilliant	hungry	bristly	secret
clancular	secret	evil	pointed	quiet	tough	pompous	obscure
gaberlunzie	beggar	culprit	dweller	healer	post	landlord	secret
varec	seaweed	ordeal	pitfall	forfeit	camel	jacket	gem
peridot	gem	dove	marsh	trout	necklace	branch	beggar
urceus	jug	wolf	mast	cord	mole	cup	mist
netsuke	button	puzzle	token	harness	herb	crowd	jug
roke	mist	creek	grove	bush	gossip	bandit	button
barbet	bird	fence	trail	root	anchor	garment	fruit
achene	fruit	tape	myth	knee	pole	sword	fool
balatron	fool	guide	bomb	gate	tube	lace	knife
bistoury	knife	fair	judge	rock	fish	mask	bird

Table A6: Stimulus materials for Experiment 5

Cue	Target	Study lure 1	Study lure 2	Study lure 3	Test lure 1	Test lure 2	Test lure 3
compendious	concise	devout	weighty	sulky	elegant	adept	hospitable
frampold	quarrelsome	dishonest	insolent	ungrateful	fussy	prickly	grubby
spoffish	fussy	bogus	unkind	risky	bristly	tiresome	sloppy
hispid	bristly	mundane	soggy	reddish	pompous	husky	fragrant
orotund	pompous	burly	cheery	lavish	quarrelsome	insecure	discreet
acherontic	dismal	feeble	corrupt	selfish	concise	unstable	impartial
leggiadrous	elegant	protective	impersonal	sensible	dismal	muscular	sturdy
stentorian	loud	stern	trim	bold	smooth	notable	distinctive
glabrous	smooth	rare	odd	mad	strange	golden	varied
selcouth	strange	fresh	bright	round	loud	patient	impossible
rapparee	bandit	chariot	gateway	steamboat	garment	admirer	snowstorm
hauberk	garment	banquet	fortress	puppet	sword	maze	ledge
falchion	sword	barge	hood	plank	bandit	ambush	prisoner
mechlin	lace	shrine	fold	chord	landlord	mint	castle
frisket	mask	glove	cork	pearl	lace	hose	trifle
zamindar	landlord	dentist	salesman	clergy	mask	cavity	mineral
kaross	jacket	mountain	owner	campus	cup	camera	magic
withe	branch	gift	scheme	drill	jacket	shadow	plot
hanap	cup	load	sheet	meat	crowd	snake	wave
karimption	crowd	band	fort	stream	branch	grass	chest
rebarbative	repellent	bubbly	victorious	obstinate	noisy	immovable	jovial
morigerous	obedient	meticulous	amiable	inquisitive	cowardly	chivalrous	arrogant
recreant	cowardly	flamboyant	defiant	mischievous	repellent	impetuous	erratic
plangent	noisy	clumsy	dreary	sleepy	threatening	outspoken	tedious
minatory	threatening	conclusive	forthcoming	eccentric	lasting	trivial	childish
olamic	lasting	blessed	monstrous	sweeping	smart	unfair	suspicious
sprauncy	smart	lean	brave	dim	quick	heroic	lucky
esurient	hungry	nervous	fortunate	stupid	obedient	noble	unexpected
coriaceous	tough	wise	calm	glad	hungry	dirty	wise
tantivy	quick	warm	dry	flat	tough	frank	clean
mehari	camel	almond	fanfare	helmet	necklace	ladle	goldfish
carcanet	necklace	packet	bonnet	slipper	herb	cloak	orchard
desman	mole	vase	peg	toad	camel	falcon	worm
munjeet	herb	booth	grudge	lamb	mole	bucket	chimney
quidnunc	gossip	nonsense	moonlight	romance	anchor	lantern	ribbon
kedge	anchor	razor	needle	doorway	pole	fabric	onion
quant	pole	brick	lamp	tray	tube	pen	bell
valinch	tube	lane	horn	maid	post	blanket	slave
gadoid	fish	tongue	bench	yard	gossip	joint	chicken
stanchion	post	bank	youth	ship	fish	enemy	cover
menald	spotty	brainy	chatty	drowsy	woolly	silky	craggy
hyaline	glassy	thorny	spongy	lumpy	spotty	flimsy	sickly
levisomnous	observant	nocturnal	expectant	versatile	horrible	hesitant	grandiose

floccose	woolly	crooked	flagrant	husky	glassy	bland	dainty
insulse	boring	aloof	cunning	moody	observant	dizzy	greedy
gallionic	careless	furious	kindly	stormy	boring	insistent	merry
infandous	horrible	constructive	reluctant	intensive	careless	sentimental	persistent
recondite	obscure	absurd	clever	weary	extra	formidable	imaginary
subsecive	extra	welcome	tiny	brilliant	secret	proud	formal
clancular	secret	evil	pointed	quiet	obscure	traditional	official
gaberlunzie	beggar	culprit	dweller	healer	seaweed	patchwork	spider
varec	seaweed	ordeal	pitfall	forfeit	gem	hurdle	ointment
peridot	gem	dove	marsh	trout	beggar	gadget	vine
urceus	jug	wolf	mast	cord	mist	plague	feather
netsuke	button	puzzle	token	harness	jug	coin	costume
roke	mist	creek	grove	bush	button	tower	pine
barbet	bird	fence	trail	root	fruit	dancer	gallery
achene	fruit	tape	myth	knee	fool	paint	soldier
balatron	fool	guide	bomb	gate	knife	concert	palace
bistoury	knife	fair	judge	rock	bird	bottle	leader

Table A7: Stimulus materials for Experiment 6

Cue	Target	Lure
achene	fruit	tape
algid	chilly	shallow
carcanet	necklace	packet
coriaceous	tough	glad
descry	detect	exploit
esurient	hungry	lonely
famulus	assistant	illusion
floccose	woolly	jerky
frampold	quarrelsome	dishonest
gaberlunzie	beggar	hovel
kaross	jacket	mountain
objurgate	rebuke	amaze
orotund	pompous	burly
quidnunc	gossip	nonsense
roke	mist	creek
selcouth	strange	fresh
subsecive	extra	welcome
trammel	impede	outwit
unguinous	oily	kindly
zamindar	landlord	ribbon
aegagrus	goat	maze
clancular	secret	obvious
conspue	despise	inflict
desman	mole	toad
facinorous	infamous	judicious
gadoid	fish	knee
infandous	horrible	luminous
inosculate	join	hang
insulse	boring	righteous
leggiadrous	elegant	protective
levisomnous	observant	nocturnal
mechlin	lace	shrine
plangent	noisy	clumsy
rebarbative	repellent	acoustic
recreant	cowardly	flamboyant
stanchion	post	watch
subduce	withdraw	enslave
surquedry	arrogance	diligence
tressilate	quiver	grapple
valinch	tube	bird
balatron	joker	phantom
complect	intertwine	overwhelm
droze	drip	coax

frisket	mask	glove
gallionic	careless	knightly
hauberk	garment	banquet
hispid	bristly	artful
kedge	anchor	razor
maunder	ramble	nurture
minatory	threatening	extensive
morigerous	obedient	inflexible
mundify	wash	hurt
quant	pole	dish
recondite	hidden	thorough
senescent	elderly	forthcoming
sejant	seated	junior
sodality	fellowship	mystery
sprauncy	neat	lean
stentorian	loud	stern
tantivy	quick	warm

Table A8: Stimulus materials used in Experiment 7.

Cue	Target
leggiadrous	elegant
spoffish	fussy
recreant	cowardly
hispid	bristly
esurient	hungry
infandous	horrible
sprauncy	smart
coriaceous	tough
clancular	secret
surquedry	arrogance
desman	mole
mechlin	lace
kedge	anchor
quidnunc	gossip
valinch	tube
gadoid	fish
succursal	branch
objurgate	rebuke
conspue	despise
subduce	withdraw
morigerous	obedient
levisomnous	observant
compendious	concise
insulse	dull
gallionic	careless
minatory	threatening
recondite	obscure
glabrous	smooth
tantivy	quick
hauberk	garment
falchion	sword
frisket	mask
carcanet	necklace
roke	mist
kaross	jacket
sodality	fellowship
bistoury	knife
droze	drip
descry	detect
mundify	wash

Table A9: Stimulus materials used in Experiment 8

Cue	Target
upela	cask
bakarti	recluse
zarata	fuss
ondare	legacy
ohoin	thief
gazta	cheese
sagar	apple
leize	cave
artile	wool
untxi	rabbit
ohitura	custom
sardesca	fork
tipula	onion
atezain	porter
gorespen	praise
hodei	cloud
ikasgai	lesson
opari	gift
gosari	breakfast
baserri	farm
aran	plum
debeku	embargo
arbuio	scorn
atsekabe	dismay
kaxoi	drawer
ahate	duck
eskularru	glove
tximino	monkey
ukondo	elbow
mahuka	sleeve
belatz	hawk
oinetako	shoe
orratz	needle
bizitasun	vitality
saski	basket
bidaia	journey
gerriko	belt
hezur	bone
zelai	grass
landare	plant

Table A10: Stimulus materials used in Experiment 9

English	**Swahili (List 1)**	**Finnish (List 2)**
Shared cues		
cat	paka	kissa
house	nyumba	talo
bird	nyuni	lintu
flower	ua	kukka
stone	jiwe	kivi
book	kitabu	tilata
child	mwana	kapsi
water	maji	vesi
bread	mkate	leipa
chair	kiti	tuoli
Unique cues - Swahili		
bed	kitanda	
hair	nywele	
food	chakula	
broom	fagio	
head	kichwa	
jug	chombo	
horse	farasi	
fire	moto	
bag	gunia	
tree	mti	
Unique cues - Finnish		
fish		kala
window		ikkuna
potato		peruna
table		payta
apple		omena
deer		hirvi
spoon		lusikka
dress		puku
sun		aurinko
train		Juna

Table A11: Stimulus materials used in Experiment 10

English	Swahili (List 1)	Finnish (List 2)
boat	mashua	soutuvene
lake	ziwa	jarvi
doctor	tabibu	laarkari
garden	bustani	puutarha
fish	samaki	kala
potato	kiazi	peruna
jug	chombo	ruuku
spoon	kijiko	lusikka
friend	rafiki	ystava
curtain	pazia	verho
child	mwana	kapsi
scarf	leso	huivi
snow	theluji	lumi
bird	nyuni	lintu
dust	vumbi	tomu
cat	paka	kissa
cloud	wingu	pilvi
bread	makate	leipa
beer	pombe	olut
horse	farasi	hevonen

Appendix B – 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 Experiment 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; 2.58s (*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 C - 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 (0-100) indicating how accurate they thought their JOL was. As shown in Figure C1, 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.

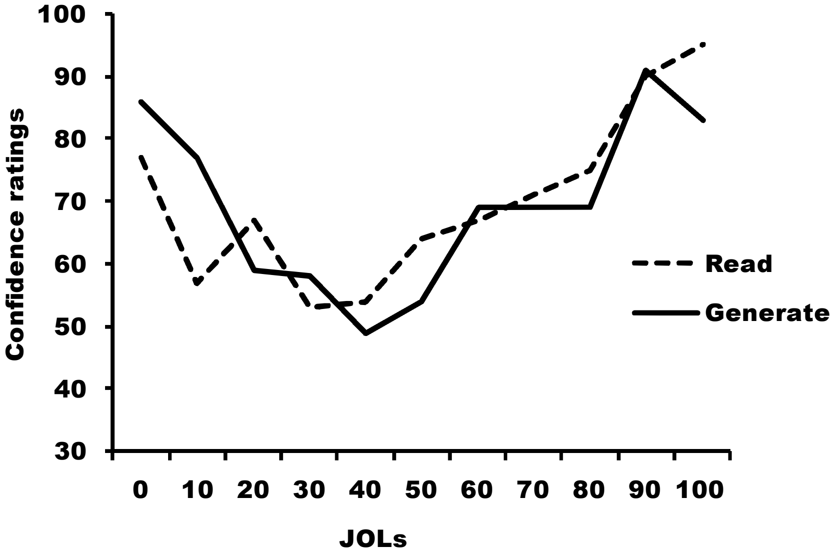


Figure C1: Mean confidence ratings as a function of judgments of learning (JOLs).

Appendix D - JOLs for items correct and incorrect at test

For completeness, we report data concerning the relationship between JOLs and test accuracy in Experiments 2-4, 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. In Experiments 2A, 2B, 3 and 4, and within each condition, JOLs for items subsequently answered correctly at test were numerically higher than for items answered incorrectly.

We computed 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 D1. 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.

Experiment 4 showed a similar pattern: The mean gamma correlation between JOLs and test accuracy was significantly higher than zero in all conditions, *t*(28) = 4.12, *p* < .001 for Read; *t*(28) = 4.13, *p* < .001 for Generate; and *t*(28) = 2.08, *p* = .047 for Choice, with no difference in resolution between conditions, *F* (2,56) = 1.75, *p* = .183.

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 in any experiment. 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.

Table D1: Mean (*SD*) gamma values for the correlation between JOLs and final test performance in Experiments 2A, 2B, 3 and 4.

	Read	Generate	Choice
Exp 2A	.34 (.38)	.34 (.43)	.27(.45)
Exp 2B	.09 (.54)	.20 (.53)	.13 (.53)
Exp 3	.39 (.33)	.33 (.38)	.18 (.46)
Exp 4	.27 (.36)	.34 (.44)	.14 (.37)

Appendix E - 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, η_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* = 75.4, *SD* =17.6) led to better recall than reading (*M* = 67.5, *SD* = 20.7), *t*(105) = 4.18, *p* < .001, *d* = .41. Choosing (*M* = 72.8, *SD* = 21.3) was better than reading, *t*(105) = 2.47, *p* = .015, *d* = .25, with no difference between choosing and generating, *t*(105) = 1.20, *p* = .234, *d* = .13.

Appendix F - 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 hazardous to draw any firm conclusions from 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 may therefore not 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).