The benefit of generating errors during learning: What is the locus of the effect?

Rosalind Potts, Gabriella Davies, and David R. Shanks

University College London

Author Note

Rosalind Potts, Gabriella Davies, and David R. Shanks, Division of Psychology and Language Sciences, University College London.

Parts of the research reported here were presented at a meeting of the Psychonomic Society (2016) and a meeting of the Experimental Psychology Society (2017).

We thank Alyssa Chen, Dina Khatun, Kelly Lee, Alison Mok, Georgina Pendrigh, Minseo Song, Deepa Salvaraj, Christie Tang, Jessie Teoh, Jun Wang, and Peri Wiseman for assistance with data collection.

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

Abstract

Guessing translations of foreign words (*hodei?)*, before viewing corrective feedback (*hodei-cloud*), leads to better subsequent memory for correct translations than studying intact pairs (*hodei-cloud*), even when guesses are always incorrect (Potts & Shanks, 2014), but the mechanism underlying this effect is unknown. Possible explanations fall into two broad classes. One puts the locus of the effect at retrieval: Items studied through a generation process have more potential retrieval cues associated with them, or a more distinctive context, and are therefore more accessible at final test. The other puts the locus at encoding and involves enhanced encoding of the correct answer following the generation of an error compared with passive studying (Potts & Shanks, 2014). In six experiments we found support for the proposal that generating errors benefits memory through stimulating curiosity to learn correct answers following an incorrect guess, leading to enhanced processing of targets following generation. In Experiment 1, generating possible translations *after* seeing correct answers did not produce better memory than studying without generating, suggesting that an element of surprise or anticipation is necessary for generating to benefit memory. Experiments 2a-c found enhanced recognition memory for targets following generating, suggesting increased focus on targets following a guess. In Experiments 3 and 4, participants rated their curiosity to learn correct answers higher when ratings were given after generating than before, suggesting that the act of generation increases curiosity to learn the answers. These findings imply that enhanced processing of feedback is a key consequence of errorful generation.

Keywords: learning, errors, generation, motivation, curiosity

It is well established that successfully retrieving items from memory leads to better subsequent memory for those items than doing nothing or simply restudying them, a phenomenon known as the testing effect (see Roediger & Karpicke, 2006, for a review, and Rowland, 2014, for a meta-analysis), leading to calls for a greater use of testing in educational settings (e.g., Pashler, Bain, Bottge, Graesser, McDaniel, & Metcalfe, 2007). While the testing effect typically involves retrieval from episodic memory, a similar benefit is observed when items are generated from semantic memory in response to a cue according to a given rule (e.g., "opposite of *cold: h__?"*) compared with when targets are presented together with their cues (*cold: hot*): the generation effect (Slamecka & Graf, 1978). Retrieval or generation of correct answers is beneficial even when no feedback is given (Allen, Mahler & Estes, 1969; Carpenter & DeLosh, 2005, 2006; Kuo & Hirshman, 1996) – this is referred to as a direct effect of testing – though feedback can enhance memory for both correct and incorrect responses (Butler & Roediger, 2008).

Furthermore, it has been shown that conditions that make initial retrieval or generation difficult or effortful yield greater subsequent memory benefits. For example, the benefit of testing is greater when the initial test involves recall rather than recognition (Bjork & Whitten, 1974); when retrieval practice is spaced rather than massed (Cull, 2000; Jacoby, 1978; Landauer & Bjork, 1978; Karpicke & Roediger, 2007; Pyc & Rawson, 2009); when fewer retrieval cues are provided at initial test (Carpenter & DeLosh, 2006); when cue-target pairs are weakly rather than strongly associated (Carpenter, 2009); and when retrieval latencies at initial test are longer (Gardiner, Craik, & Bleasdale, 1973). All of these are indicative of more effort causing greater beneficial effects. Since more difficult tests carry a greater risk of producing errors, a factor that could limit the use of testing in educational settings is the concern that errors could interfere with memory for correct information. However, this fear may be unfounded: There is evidence that even pre-testing, that is, having participants generate responses before they have ever been exposed to the relevant material, can be beneficial to memory even when generation produces many errors (Kane & Anderson, 1978; Potts & Shanks, 2014; Pressley, Tanenbaum, McDaniel & Wood, 1990; Richland, Kao & Kornell, 2009). This type of benefit is an *indirect* effect of testing, since the enhancement to memory is not to the tested item itself but to the subsequently studied correct version of the material, and is assumed to occur through more effective processing of the studied material following the generation of a response, even when that response is an error.

There are other ways that testing may benefit memory indirectly, for example by motivating students to study continuously throughout a course rather than cram for a final exam (Leeming, 2002), or by enabling students to identify items that are not yet well learned and focus their subsequent study efforts on those items (Kornell & Metcalfe, 2006), but these are cases where learners have typically already been exposed to the study materials. By contrast, the focus of the current study is on the situation where generating a response to hitherto unfamiliar materials, followed by a study opportunity, leads to memory enhancement relative to passive studying without generation. There are myriad circumstances in which someone may make a guess in response to an unfamiliar stimulus. In educational settings, a student may attempt to answer questions in an exam for which they have failed to prepare, or may find they have to guess in response to a question put to them by a teacher in class. The answering of trivia questions is involved in a variety of popular recreational activities, from internationally-known board games such as Trivial Pursuit to TV quizzes and the pub quizzes that form a well-established part of British social life. On many occasions a contestant will find themselves facing a question to which they do not know the answer. Does guessing incorrectly in these situations make the correct answer more or less likely to be remembered once it is known?

In an early study addressing this question, Kane and Anderson (1978) found that participants who were asked to fill in the missing last word of a sentence before being shown the experimenter-designated "correct" completion showed better subsequent memory for the words than those asked to read complete sentences, even when sentences were constructed such that it was not possible to infer the missing word and participants' initial responses were therefore always incorrect. The authors concluded that the requirement to complete the sentence forced participants to process the material for meaning, while reading the intact sentence did not and this deeper level of processing led to enhanced memory in the more active condition.

Other studies have found that answering pre-questions before studying text passages led to better performance on a post-study test than simply reading the pre-questions (Richland et al., 2009) or evaluating them for comprehensibility without answering them (Pressley et al., 1990), even when the pre-questions had been answered incorrectly, though the difference was not significant in the Pressley et al. study when only the initially incorrect responses were included. Richland et al. (2009) proposed that the benefit in their study was due to elaborative processing stimulated by the question: Even when the question was not correctly answered, the search for the answer would have activated many relevant concepts, facilitating encoding of the answer when the text was studied. Likewise, Kornell (2014) found a benefit of incorrect guessing when participants were pre-tested with trivia questions and proposed a similar explanation for this finding.

In such studies it is never possible to preclude altogether the possibility that participants already know the answers to some of the questions but are unable to retrieve them on the pre-test. Slamecka and Fevreiski (1983) demonstrated how partial retrieval in response to a cue could give the appearance of generation failure on a pre-test while facilitating retrieval on a later test. Kornell, Hays, & Bjork (2009), in an effort to address this issue, put fictional trivia questions to participants and found mixed results, but never any detriment to guessing by comparison with reading. This study also found no benefit, but again no detriment, of guessing for nonfictional trivia questions.

In these studies, the pre-questions or cues involved material that had pre-existing associations that could be linked semantically with the targets once these were presented. What happens when this is not so easy? Traveling abroad throws up many situations in which one has to make one's best guess as to what a foreign word means, for example, when deciphering menus or instructions. Faced with an unfamiliar word, we may search for a similar-looking word in our own language or another language that we know. Unfortunately, sometimes the appearance of the word can mislead. The first author, when an impoverished student, ordered “flan” in a Spanish café, thinking she would get a kind of tart or pie that might do in place of lunch. Imagine her disappointment on receiving a small pot of crème caramel that was most definitely not sufficient to see her through to the evening meal. More than thirty years later, the true meaning of the Spanish “flan” has stuck in a way that it might well not have done had she simply read it together with its translation in her guide book.

Indeed, Potts and Shanks (2014) found just such a benefit of incorrect guessing when cues were novel foreign vocabulary items or very unusual English words that had never previously been seen: Participants showed better memory for Euskara-English word pairs when they had learned them by generating a guess on first presentation of the cue (e.g., *hodei?_),* followed by viewing the correct answer (*hodei-cloud*) than when they had learned the words by reading the intact word pair (*hodei-cloud*). In their procedure, participants either studied the pairs in a Read condition (*hodei-cloud*), were presented with the cue (*hodei?_*) and had to generate a response before being shown corrective feedback (*hodei-cloud*: Generate condition), or were presented with the cue (*hodei?_*) and four possible options from which to choose, again before viewing feedback (Choice condition). Since the cues had never previously been seen, participants nearly always generated incorrect responses for Generate items, yet on a subsequent final multiple choice test of all the items, participants scored higher on Generate than either Read or Choice items. This result was all the more striking because the total trial time was equated in the three conditions so that exposure to the correct answer was much greater in the Read than in the Generate condition.

Since the cues here were novel, previously unseen, items, they would have had no pre-existing associations in participants’ minds and therefore there would have been no opportunity to process them for meaning as in the Kane and Anderson (1978) study, or to stimulate elaborative processing that could facilitate encoding of a semantically related target as in the Richland et al.

(2009) study. Indeed, Potts and Shanks (2014) conducted a latent semantic analysis and found no semantic association between participants' guesses and the correct targets. Under the conditions of the Potts and Shanks (2014) study, participants' initial generations are nearly always both erroneous and unrelated to the targets. The mechanism underlying this "errorful generation" benefit for unfamiliar materials (Potts & Shanks, 2014) is not well understood, but is an important issue to explore at a time when we are seeing a proliferation of self-study systems available on the internet. Popular online language learning platforms such as Memrise and Duolingo make frequent use of tests with feedback. A critical issue for the creators of such systems is determining the point at which tests should be introduced during learning. Testing an item too early increases the risk of error, with the potential for creating confusion in the learner, while excessive caution with regard to testing can lead to precious time wasted, with the risk that users will go elsewhere for their studies. For testing to be put to optimal use it is important to identify the conditions that determine when errors are helpful or harmful to memory. The aim of the current study is to shed some light on this issue by exploring possible explanations for the benefit of generating errors during the learning of unfamiliar material such as foreign language vocabulary. However, in doing so we also aim to speak to the wider issue of when pre-questions – including of more complex materials presented in textbooks or lectures - may be helpful or harmful to learning. First, using novel foreign language vocabulary as cues allows us to eliminate the possibility that participants already know the answers, even if they cannot immediately retrieve them. Second, it allows us to explore mechanisms that may be operating over and above those stemming from the semantic relatedness between cue and target, and which could otherwise be obscured by the effects of semantic relatedness that are likely to occur in other, commonly occurring, pre-testing scenarios.

Possible explanations for the benefit of errorful generation previously observed for novel vocabulary items fall into two broad classes: those emphasising processes that are active during the generation attempt and those concerned with processes that are operating after the feedback has been presented. In the first case, generating errors could benefit memory because the act of generating a response to a cue entails the activation of many concepts which, even if they are unrelated to the target, may serve as retrieval cues at final test, thereby making the encoded memory more accessible (see Grimaldi & Karpicke, 2012, and Kornell, Hays, & Bjork, 2009, for related proposals.) On the other hand, generating errors may be helpful to memory because it brings about more effective encoding of the cue-target pair, for example by arousing curiosity to know the correct answer, increasing motivation to learn that answer when it is presented as feedback (Potts & Shanks, 2014). The aim of the experiments reported here was to attempt to distinguish between these two classes of explanation, though it should be noted that they are not mutually exclusive. Next, we describe these accounts in more detail.

The elaborative generation hypothesis

The first explanation, which we call the *elaborative generation hypothesis,* proposes that the process of generating a guess involves the generation of many other concepts, which become bound to the target during study and can subsequently serve as mediators to aid retrieval of the target. At final test, presentation of the cue activates in memory the generated items that were associated with it at study, including the participant’s response and other ideas generated at the same time, and these can serve as mediators to the target. For example, seeing the word *hodei* at study might lead to the generation of “coal” because the word is reminiscent of “hod”. In turn, this could activate a number of related concepts to do with coal cellars, fires and so forth. When the correct translation, *cloud*, is presented as feedback, any of these concepts can become associated with it, either by design on the part of the participant (e.g., imagining a cloud of coal dust), or simply by virtue of its contemporaneous activation. At final test, presentation of *hodei* once again calls to mind a hod and the related concepts that were evoked at study, one or more of which may, in turn, lead to *cloud.* By contrast, studying the intact word pair (*hodei-cloud*) does not involve this proliferation of cues which can subsequently be used as mediators to the target. Potts and Shanks (2014, Experiment 3), using a procedure in which the final multiple choice test included the incorrect response generated by the participant at study, found that participants were very good at rejecting their own incorrectly generated responses, choosing them significantly less often than other incorrect options. This finding suggests that memory for the incorrect generation persists to final test and could potentially be used as a mediator to the correct answer.

This hypothesis has not yet been directly tested in relation to the errorful generation effect for novel stimuli but findings from the testing effect literature provide some evidence that additional cues generated during testing may serve as retrieval cues at a later test. Pyc and Rawson (2010) investigated the role of mediators in the testing effect. At study, they had participants generate mediators to help them remember associations between Swahili words and their English translations (e.g., "*wing*" for "*wingu-cloud"*) and then repeatedly either restudy the words or take a test with feedback. As well as a standard testing effect, Pyc and Rawson found that, at final test, both mediator retrieval (ability to recall the mediator from a cue) and mediator decoding (ability to identify the target from the cue plus mediator) were better following testing than restudy and concluded that testing benefits memory by enhancing both the retrieval and the decoding of mediators.

In Pyc and Rawson's study, participants were explicitly taught to generate mediators for both studied and tested items and were told to report them on every trial. Carpenter (2009, 2011) explored how additional cues generated during the search for a studied target could act as mediators on a later test, even when participants were not deliberately creating and recalling mediators. She proposed an *elaborative retrieval hypothesis* as an explanation for the benefit of testing on memory for previously studied material (Carpenter, 2009). According to this account, retrieval of studied material at initial test activates items associated with the cue and, by a process of spreading activation (Collins & Loftus, 1975), activates a semantic network of related concepts, any of which can provide a route to retrieval of the target at final test. Carpenter (2009) argued that when initial retrieval was more difficult, more concepts would be activated during the search for the target at initial test, and this would benefit subsequent test performance by providing more retrieval routes to the target on the later test. In support of this proposal, she found that weak cues (*basket – bread*) led to better final test performance than strong cues (*toast – bread*).

In a subsequent study, Carpenter (2011) explored the role of semantic mediators more directly. Participants studied weakly associated word pairs, such as *mother-child*. They then either took a test of the cue (*mother?*) or restudied the pair (*mother-child*). On a final recognition test comprising cues, targets, semantic mediators and unrelated words (e.g., *bread*), false alarms to semantic mediators were higher for tested than restudied items, in support of the proposal that testing was more likely to activate mediators than studying. In a second experiment, the final test involved recalling targets from cues, mediators or unstudied words that were weakly related to the cues (e.g., *birth*). Recall of targets from semantic mediators was better for tested than restudied items, in support of the proposal that testing activates mediators that can be used as routes to retrieval of targets at a later test.

Lehman and Karpicke (2016) cast doubt on the conclusion that testing activates more mediators than restudying. Participants studied weakly related word pairs, as in Carpenter (2011). They then either took a test of the cue plus stem (*mother-ch__*) or restudied the pair (*mother-child*), immediately followed by a lexical decision judgment to an unstudied semantic mediator (*father*), an unrelated word (*banquet*) or a nonword (*clett*). A final cued recall test of the targets showed a typical testing effect but priming of semantic mediators was no greater following test than restudy trials, suggesting that greater activation of mediators during retrieval than restudy may not be responsible for the testing effect. However, the use of a cue plus word stem in the test condition may have constrained the generation of potential mediators, making it less likely that, for example, *father* would have been activated during test of *mother-ch__.* In this case the test condition could still have activated more mediators than the restudy condition, but not the ones that were presented in the lexical decision task.

The experiments by Carpenter (2009, 2011) and Lehman and Karpicke (2016) were designed to investigate the mechanism by which retrieval of *correct* answers at initial test benefits later memory. Grimaldi and Karpicke (2012) explored whether a similar mechanism could account for a benefit of generating *incorrect* responses that they observed for weakly related word pairs (e.g., *frog-pond*), testing the hypothesis that the incorrect response generated by the participant at study provided an additional cue that could be used as a mediator to the target at final test, since, with semantically related pairs, the generated guess is likely to be related to both cue and target. The basic procedure was similar to that used by Potts and Shanks (2014) as described above. However, in Grimaldi and Karicke's (2012) second experiment, in addition to conditions in which participants either read items (*tide-beach*) or generated responses (*tide?__*), they included two new conditions: *study-lure*, in which participants read the cue paired with an incorrect response (*tide-wave*) before being shown the correct pair (*tide-beach*), and *constrained pre-test,* in which participants were encouraged to generate an error in response to a fragment (*tide-wa__?)* before being shown the correct pair. Grimaldi and Karpicke argued that, if generation of responses enhances memory by creating additional cues, then both of these conditions should enhance memory relative to reading, but this was not the case. In fact, for both related and unrelated pairs, the *constrained pre-test* condition led to worse memory than reading the pairs, and the *study-lure* condition was no better than the Read condition. Grimaldi and Karpicke argued that these findings were inconsistent with what they termed the *additional cue theory*.

However, in both of the new conditions the lure was provided by the experimenter: Even in the *constrained pre-test* condition, the participant was not generating their own response to the cue but one that fitted the fragment provided to them, which may have made it less useful at final test than a response freely generated by the participant themselves. Furthermore, generation was constrained to one particular item, the one that fitted the fragment. During unconstrained generation, it is likely that many items will be activated before settling on a response, creating an elaborate network of retrieval cues that can be used to access the target at a later test. Therefore it is possible that the benefit of generating incorrect guesses during novel learning, as observed in the Potts and Shanks (2014) study, could be explained by an *elaborative generation hypothesis*, similar to Carpenter's *elaborative retrieval hypothesis*: Cues activated during a generation attempt could create a distinctive context for that part of the study episode and act as retrieval cues to the target at later test, and this could be the case even when they are semantically unrelated to the target, as they are likely to be when generating responses to previously unseen vocabulary items.

The enhanced encoding of feedback hypothesis

An alternative proposal is that generation, even when it is errorful, benefits subsequent memory not by activating cues that can later be used as retrieval prompts, but by enhancing the encoding of corrective feedback (Potts & Shanks, 2014). There are several reasons why encoding of correct answers might be more effective following generation than during a Read trial, when learning novel material such as previously unseen foreign vocabulary. First, the active process of generating a guess is more effortful than reading and a generation attempt forces the learner to confront their own lack of knowledge about the answer and to realize, by experiencing first-hand, the difficulty of generating a correct response to the cue. This realization then leads to more processing of the target when it is presented as feedback, enhancing learning of the target. By contrast, during Read trials learners may experience a "knew-it-all-along" effect (Fischhoff, 1977), or a "foresight bias" (Koriat & Bjork, 2005) whereby the future difficulty of retrieving the correct target from the cue is underestimated when the cue and target are seen together. Potts and Shanks (2014) found that participants consistently gave lower judgments of learning (JOLs) to Generate items than to Read items while test performance showed the opposite pattern, supporting the proposal that participants perceive Generate items as more difficult to learn than Read items. Similarly, in an errorful generation study using weakly associated word pairs, Yang, Potts, and Shanks (2017b) found that participants allocated more restudy time to items to which they had generated incorrect responses than to those they had learned by reading, despite recall being better for the Generate items, again suggestive of more effort being applied to Generate items.

Furthermore, there is evidence that taking tests is motivating. Weinstein, Gilmore, Spzunar, and McDermott (2014) found that a group that was warned to expect a test following an upcoming study episode performed better when tested on the studied material than a group that did not expect to be tested, and Yang, Potts, and Shanks (2017a) found that taking many interim tests led to participants spending longer encoding subsequent lists, consistent with the idea that tests maintain motivation.

There are other reasons why participants might be more focused on correct targets in the Generate condition. Butterfield and Metcalfe (2001) found that, when being tested on trivia questions, errors made with high confidence were more likely to be corrected than those made with low confidence, a phenomenon they called the *hypercorrection effect*, and suggested that the surprise engendered by the discrepancy between the participant's expectations and the outcome captured attention, leading to enhanced learning. In support of this proposal, Butterfield and Metcalfe (2006) found that participants were less able to detect a tone occurring during presentation of feedback following high-confidence errors than low-confidence errors, suggesting attention was captured by the feedback, while Fazio and Marsh (2009) found that participants were more likely to remember the font colour of feedback presented following high-confidence errors, again suggesting greater attention to the feedback.

In the task used by Potts and Shanks (2014), which entailed generating guesses about the meanings of never previously encountered foreign words or obscure English words, participants were unlikely to be confident that their guesses were correct for any of the items, but there was still a discrepancy between the generated guess and the corrective feedback, and generating guesses for novel items may have aroused curiosity to learn the answer. Curiosity has been described as arising from an information gap, that is, a discrepancy between what an individual knows and what they want to know (Loewenstein, 1994). In the errorful generation task used by Potts and Shanks (2014), the effortful process of searching for a response to an unfamiliar cue (*hodei_?*), may have benefitted memory by making participants aware of this information gap, leading them to focus more on the correct answer when it was presented as feedback. Consistent with this account, Berlyne and Normore (1972) found that recall of pictures of objects was better when the pictures were first presented in a blurred version, followed by a clear version, than when they were clear from the outset (akin to the 'revelation' effect in recognition memory), and concluded that the uncertainty induced by the blurred picture made participants especially attentive to information that relieved the uncertainty.

More direct evidence comes from a study by Kang, Hsu, Krajbich, Loewenstein, McClure, Wang, and Camerer (2009), who had participants answer trivia questions and rate their curiosity to know the answers before being given corrective feedback. On a surprise subsequent recall test, memory for items answered incorrectly at initial test was greatest for questions to which participants had given the highest curiosity ratings. Functional imaging revealed that curiosity ratings were correlated with brain activation in prefrontal cortex and caudate areas associated with reward anticipation. Gruber, Gelman, and Ranganath (2014) also observed better recall of trivia questions associated with high rather than low curiosity ratings and concluded that "curiosity enhances learning, at least in part, through increased dopaminergic modulation of hippocampal activity" (p. 491).

Thus, possible explanations for the benefit of generating errors fall into two broad classes. One puts the locus of the effect at retrieval: Items studied through a generation process have more potential retrieval cues associated with them, or a more distinctive context, and these additional cues can be used as mediators between cue and target at final test. The other class of explanation puts the locus of the effect at encoding and involves greater engagement with the correct answer following the generation of an error compared with during passive studying (Potts & Shanks, 2014).

In Experiments 1 and 2 we attempt to distinguish between these classes of explanation and determine the extent to which either or both contribute to the errorful generation effect.

We used the same Read and Generate conditions as in Potts and Shanks (2014). In Experiment 1, we additionally introduced a modified Read condition in which participants were asked what they would have guessed had they not already seen the target, thus involving the generation of potential retrieval cues but with no information gap, and no opportunity for curiosity to be aroused, since the target is already known. To foreshadow, participants performed no better in the modified Read condition than in the standard Read condition, suggesting that an element of anticipation or surprise is necessary for generation to enhance memory. Experiment 2 aimed to examine enhanced encoding of the feedback more directly by testing participants on their memory for the targets only. Participants recognised significantly more Generate than Read targets in a target-only recognition test. If the benefit of generating arises solely from the use of mediators between cue and target during final test, we would have expected to see no difference in memory for Generate and Read items in the target-only final test. Since the cues were not available to participants at final test, this suggests that at least part of the benefit of generating errors is due to enhanced encoding of the target in the Generate condition, making these targets more accessible at final test.

Experiments 3 and 4 explored whether this enhanced encoding of feedback stemmed from increased curiosity aroused by generation, by having participants rate their curiosity to know the answer on a 7-point scale. In Experiment 3, curiosity ratings were higher when participants generated a response before making a rating than when they simply made a rating without overtly generating. In Experiment 4, curiosity ratings given after generating responses were significantly higher than ratings given before generating.

Experiment 1

The aim of Experiment 1 was to explore whether the errorful generation benefit for novel materials arises because generating activates myriad cues that can be used to access the target at final test or because the act of generation increases curiosity to learn the answer, leading to increased interest in and processing of the feedback at encoding. To distinguish between these two explanations, we created a modified version of the Read condition used in our previous experiments (Potts & Shanks, 2014). This modified version was designed to include an element of generation but without the potential for arousal of curiosity, or for surprise at the discrepancy between response and target. Instead of merely reading the cue and target, in the modified Read condition participants first saw the cue-target pair and were then asked to enter what they might have guessed had they not already seen the answer. They therefore generated a guess based on the cue, just as in the standard condition, but there could be no arousal of curiosity about the answer, or surprise at the discrepancy between the response and the target when the correct answer was presented, as this was already known.

If the benefit of generating arises because the act of generating a response activates cues which can later be used as mediators to the target, the modified Read condition should yield a benefit over reading in the same way as the Generate condition does. If, on the other hand, the effect depends on an element of anticipation or curiosity before the answer is revealed, or surprise on viewing the feedback, the modified Read condition will not yield these benefits and performance will not be significantly different from that in the standard Read condition. It should be noted that, if there is a benefit of Modified Read over Read, there is no reason to suppose that the magnitude of the difference will be the same as that for the Generate advantage over Read: Once the target has been seen, it may be difficult for participants to imagine what they would have guessed had they not already seen it, resulting in “guesses” that are more closely related to the target than in the Generate condition. If this is the case and if the benefit of generation is that it activates cues that can be used as mediators, performance in the Modified Read condition could outstrip that in the Generate condition since the more closely related responses should be more effective mediators. On the other hand, guesses that are close to the target mean that the spread of activation is likely to be smaller in the Modified Read condition, leading to fewer cues being activated and therefore fewer potential routes to retrieval.

The important point, though, is not how the Modified Read condition performs in relation to the Generate condition, but how it performs in relation to the Read condition: If the act of generation is sufficient to produce a memory benefit, then the Modified Read condition, which involves the generation of a response, should produce better final test performance than the Read condition, which does not. The final test was in multiple choice question (MCQ) format, consistent with the previous work we are following up (Potts & Shanks, 2014; see also Potts, 2014). Multiple choice is a very common format in official tests, such as the Medical College Admissions Test (MCAT) in the USA, and the "Life in the UK" test that forms part of the application for British citizenship for foreign nationals. The MCQ format is also commonly used in the online materials accompanying college textbooks, and widely used for undergraduate assessments in many countries. Language learning platforms such as Memrise make extensive use of this type of test during the learning of foreign vocabulary. While much of the pre-testing literature has explored the effect of incorrect guessing on performance in cued recall or free recall tests, the effect on MCQ tests has received less attention, despite the prevalence of this form of testing in real world scenarios such as those described above.

Although our primary aim was to explore the locus of the errorful generation benefit, we also captured judgments of learning (JOLs) at the end of each study trial, and we report these results for completeness. In our previous work (Potts & Shanks, 2014), participants gave higher JOLs to Read items than to Generate items, perhaps reflecting a "knew it all along effect" (Fischoff, 1977) or foresight bias (Koriat & Bjork, 2005). When studying Read items, participants may fail to consider the difficulty of generating a response in the absence of the target. By contrast, the effort of coming up with a response in the Generate condition may lead to a perception of difficulty, reflected in lower JOLs. If this is the case, having participants generate a pseudo-guess in the Modified Read condition may alert them to how difficult generating a response to the cue could be, leading to lower JOLs in this condition than in the standard Read condition. We therefore expected that JOLs for the Modified Read condition, because it involved the effort of generation, would be similar to JOLs for the Generate condition and lower than those in the standard Read condition.

Method

This experiment and the others reported here were approved by the UCL Research Ethics Committee.

Participants

To determine an appropriate sample size, we carried out a power analysis using G*Power (Faul, Erdfelder, Lang, & Buchner, 2007), based on data from a previous experiment using foreign language vocabulary as stimuli (Potts & Shanks, 2014, Exp 2B), which yielded an effect size of d_z = 0.69 for the difference between Read and Generate final scores, and determined that a sample of 30 participants would give us a power of 0.95 to detect a similar size effect in the current experiment (two-tailed), or 35 participants to give power of 0.99 (one-tailed). We expect the Modified Read condition to achieve similar scores either to the Read condition or to the Generate condition. Thirty-six participants were recruited by opportunity sampling. One participant failed to follow the instructions, resulting in their final test data not being recorded. This participant was replaced, yielding data from 36 participants, 27 female, average age 22.64 (*SD* = 7.23). Four participants reported that English was not their native language.

Design

Experiment 1 used a within subjects design with three conditions: Read, Modified Read and Generate. There were two dependent measures: final memory test score and judgment of learning.

Materials

Stimulus materials were 60 Swahili-English word pairs taken from a larger set normed by Nelson and Dunlosky (1994). For each pair, we created three English lures for use in the final multiple choice test. The lures were approximately matched to the targets for word frequency, concreteness and imageability. The set of 60 items was divided into three subsets for counterbalancing purposes.

Procedure

Participants were told that they would study some foreign language vocabulary presented in three different formats and that they would later be tested on the English translations of the words. The total trial time in each condition was 15 s, the three trial types were randomly interleaved and the order of presentation of items was randomized on a per participant basis. In the Read condition the cue-target pair was presented for the full 15 s. In the Generate condition the cue alone was presented for 10 s, and the participant was prompted to type in their response. After 10 s had passed, whether or not the participant had finished typing their response, the correct answer was displayed for 5 s. In the Modified Read condition, the cue-target pair was on screen for the full 15 s, as in the Read condition. It was presented alone for the first 2 s of the trial, then the participant was prompted to enter what they would have guessed for the meaning of the cue had they not already seen the target (the *pseudo-guess*). They had exactly 10 s to enter their response, the same as the time allowed in the Generate condition. The cue-target pair remained visible on screen while they were making their pseudo-guess. This was followed by a further 3 s during which the pseudo-guess disappeared and the cue-target pair remained on screen alone. This condition was designed to be as close as possible to the Generate condition except that, since the correct answer was on screen from the start of the trial, there was no information gap between what the participant knew and what they wanted to know (i.e., the translation). Similarly, it was identical to the Read condition except that it involved the generation of a response to the cue. See Figure 1 for a depiction of the procedure and timings.

At the end of each trial, participants were asked to enter a judgment of learning, that is, to predict how likely they were to remember that item later, by entering a number from 0 (“No chance I’ll remember it”) to 100 (“I’ll definitely remember it”). After a 5 min distractor task involving the solving of arithmetic puzzles, participants took a self-paced, four-alternative multiple choice test with new (unstudied) items as lures.

Results.

Test scores

At study, a correct response was generated just five times (once each for five participants). Since we were interested in the effect of generating errors on memory, these items were excluded from the analysis of the final scores. The mean final test scores for the three study methods are shown in panel A of Figure 2. A repeated measures ANOVA showed that there was a main effect of study method, *F*(2,70) = 3.60, *p* = .032. Final test performance was significantly better for items studied in the Generate condition than in either the Read condition, *t*(35) = 2.47, *p* = .018, mean difference 4.38%, 95% CI [0.75, 7.59], or the Modified Read condition, *t*(35) = 2.36, *p* = .023, mean difference 4.31% [0.64, 7.97], with no difference between Read and Modified Read, *t*(35) = .070, *p* = .945, mean difference 1.39% [-3.89, 4.17]. Eighteen participants remembered Generate items better than Read items, with 6 showing the opposite pattern, and 20 showed better Generate than Modified Read performance, with 10 showing the opposite pattern. Sixteen participants did better for Read than for Modified Read items, with 13 showing the reverse pattern.

Latent semantic analysis

As discussed earlier, the Generate and Modified Read conditions differ critically in that the Generate condition involves the generation of a response in the absence of knowledge about the correct answer (the target), while the Modified Read condition involves generating once the correct answer is known. It seems likely that generated responses in the latter condition might be more closely related to the targets than those in the Generate condition. To explore this, we carried out a latent semantic analysis (LSA) to estimate the degree of semantic relatedness of the responses generated at study to the corresponding targets, using the tools provided at the LSA website hosted at the University of Colorado. Latent semantic analysis involves using statistical computations on large corpora of text to identify the similarity in meaning between pairs of words, with values close to 1 representing a high degree of semantic relatedness and values close to 0 meaning a very low degree of relatedness (Landauer, Foltz, & Laham, 1998). For each participant, we calculated the average LSA value for the generated responses in relation to the targets in each condition and compared these in a paired samples *t* test.

As expected, responses generated to Modified Read items (*M* = 0.116, *SD* = 0.027) were more closely related to the targets than responses generated to Generate items (*M* = 0.103, *SD* = 0.025), *t*(35) = 2.25, *p* = .031, mean difference 0.013 [0.0013, 0.0249]. However, they were generally not very closely related: An example of a set of responses from a typical participant is shown in Table 1.

Judgments of learning

Mean JOLs are shown in Panel B of Figure 2. An ANOVA showed that there was no difference between study methods for the JOLs, *F*(2,70) = 0.27, *p* = .763. Mean times taken to make JOLs in each condition were: Read (*M* = 4039 ms, *SD* = 3030), Generate (*M* = 3799 ms, *SD* = 1562), Modified Read (*M* = 3954 ms, *SD* = 1228). These did not significantly differ, *F*(2,70) = 0.17, *p* = .844.

We had predicted that the Modified Read condition would not yield higher JOLs than the Generate condition, since the act of having to generate a response to the cue would alert participants to the difficulty of doing so, and this prediction was supported. It was a little surprising, however, that there was no difference between JOLs for the standard Read condition and the Generate condition, given our previous findings of a benefit for Read over Generate in participants' JOLs. One possibility is that this was due to the design of the experiment. If, as we hypothesised, the Modified Read condition eliminated any illusion participants were under that the Read condition was easier than the Generate condition, it is possible that this disillusionment transferred also to the standard Read condition since all three conditions were interleaved. In other words, experiencing the difficulty of generating in the Modified Read condition may have made participants less confident of their ability to remember items learned in the standard Read condition.

To explore this possibility, we split the study trials into two time periods (trials 1-30 and 31-60) and calculated mean JOLs in each study condition in each time period. These are shown in Figure 3. A 3 (Study method) x 2 (Time period) ANOVA revealed no main effect of either study method, *F*(2,70) = 0.15, *p* = .865, or time period, *F*(1,35) = 1.62, *p* = .211, but a significant interaction between the two, *F*(2,70) = 6.11, *p* = .004. JOLs for the Generate and Modified Read conditions remained at approximately the same level over the two halves of the study period, *t*(35) = .014, *p* = .989, mean difference 0.032 [-4.61, 4.54] for Generate, and *t*(35) = .141, *p* = .889, mean difference 0.28 [-3.81, 4.38] for Modified Read. By contrast, JOLs for Read items dropped substantially between the first and second half, *t*(35) = 3.64, *p* < .001, mean difference 6.01 [2.66, 9.36], the first half showing somewhat higher judgments for Read items than for Generate items, mean difference 3.44 [-.146, 7.03], *t*(35) = 1.95, *p* = .059, in line with our previous findings. This pattern is consistent with the proposal that participants began the study phase with an “illusion of knowing” or foresight bias for Read items but, over the course of the study phase, experience of the Modified Read condition altered their perception of the memorability of Read items. It would be interesting to confirm this pattern and examine this further in future research. However, since the focus of the current study is on exploring the mechanisms underlying the errorful generation benefit, rather than on metacognitive awareness of the effect, we did not collect judgments of learning in the following experiments reported here.

Discussion

The aim of Experiment 1 was to attempt to distinguish between two classes of explanation for the errorful generation benefit: one where the locus of the effect is at encoding and involves enhanced processing of the target, perhaps due to the arousal of curiosity, and the other where the locus is at retrieval and involves the use of additional retrieval cues activated by the process of generation. Generating a response before seeing the target led to an increase of more than 4% in final test scores compared with reading, replicating our previous work. By contrast, the Modified Read condition, in which participants generated a response *after* seeing the target, fared no better than the Read condition at final test. A latent semantic analysis conducted on the relationship between generated responses and targets showed that, as expected, responses were more closely related to targets in the Modified Read condition than in the Generate condition. This was not surprising since participants had already seen the target at the time of generating their response. Theoretically, therefore, participants could have produced responses in the Modified Read condition that would help them to remember the items on the final test, but their test performance was no better for these items than for items studied under the Read condition. However, the generated responses were semantically only weakly related to targets in both conditions, suggesting that participants attempted to follow the instructions by producing a response that they might have given had they not been aware of the correct answer. It is reasonable to suppose that the attentional load was greater in the Modified Read than the Read condition, since it involved both generating a response and studying the correct pair. However, our procedure was designed to create similar demands in the Modified Read condition and the Generate condition: In both cases, the participant had 10 s in which to enter their response, and the remaining 5 s of the trial time were assigned purely to study of the cue-target pair, yet one condition produced a benefit over reading and the other did not.

The findings of Experiment 1 suggest that the act of generation is not sufficient for a benefit to be observed: Generating responses before seeing the target benefitted memory by comparison with reading, but generating afterwards did not, suggesting that the errorful generation benefit depends on an element of surprise, or an information gap between what the participant knows and what they need to know. These findings provide support for the proposal that the benefit can at least partly be attributed to enhanced encoding of the target following generation, but only when the target is not yet known. Consistent with these findings, Clark, Yan, & Bjork (2013) and Clark

(2016), in several experiments using a slightly different design and materials, also found that generating a guess before seeing the target was better for memory than generating after.[1]

Experiment 2a and 2b

Experiment 2 was designed to follow up the findings of Experiment 1, and the suggestion that the locus of the errorful generation effect is at encoding, by addressing the question of processing of feedback more directly. Is feedback more effectively encoded following generation than during passive study? To explore this, we introduced a new final test format. Participants studied items under Read and Generate conditions in the same way as in Experiment 1, but there was no modified Read condition in Experiment 2. At final test, half the items were tested on a standard multiple choice test and half on a recognition test of targets only (see Figure 4 for an illustration). If generating an incorrect guess leads participants to focus more on the target than reading does, then participants should show better memory for targets alone when they have been studied in the Generate condition than when they have been studied in the Read condition. If, on the other hand, Generate items produce better final test performance solely because more associations have been made with the cue and, during final test, these are used as retrieval cues to the target, then an advantage for Generate items will be evident on a standard MCQ test, in which both cues and targets are shown, but not on a target-only test, since there will be no cues available to trigger associations made during guess generation at study.

To illustrate, imagine that the Swahili word *tumbili* is presented at study and the participant generates the response *tumble.* When the correct answer, *monkey*, appears, *tumble* and *monkey* become associated. On a final test that includes the cue (the standard MCQ test), presentation of the cue *tumbili* once again evokes *tumble* which, in turn, activates *monkey.* By contrast, on a target-only test, *tumbili* does not appear and so there is no opportunity for *tumble* to be activated and, in

[1] We thank Courteney Clark and Bob Bjork for drawing our attention to these studies.

turn, activate *monkey.* In this case, superior recognition of targets for items studied under Generate conditions compared with Read conditions must be a result of enhanced encoding of *monkey* during study. Enhanced encoding of Generate targets could occur for a variety of reasons, including that participants apply more effort to encoding targets because they experience generating as difficult, consistent with the lower JOLs given to these items in our previous work (Potts & Shanks, 2014), or they are more curious about feedback following generation because generation stimulates a desire to close an information gap.

It is also possible that participants use an explicit strategy of linking the cue to the target or their generated guess to the target (e.g., creating a mental image of a monkey tumbling) and this involves greater processing of the target during encoding. In this case, one possibility is that the allocation of attention to the cue and the target is more evenly distributed when studying Generate items, since cue and target are presented sequentially. By contrast, during study of Read items, attention may be divided between cue and target in a less systematic way, leading to stronger encoding of Generate targets by comparison with Read targets.[2] However, whether or not such a strategy is used at study, the findings of Experiment 1 make it unlikely that superior use of such links at final test is, by itself, responsible for the errorful generation benefit, since there is no reason to suppose that participants would not have used the same strategy in the Modified Read condition, yet this showed no benefit over reading. We return to these, and other possibilities, in our discussion. The aim of the current experiment is not to distinguish between these possible reasons why generating incorrect responses might enhance encoding of corrective feedback but, first and foremost, to establish whether it does, in fact, enhance engagement in and encoding of targets relative to reading, whatever the reason.

Of course, it is possible that both increased focus on the target during encoding and the use of responses as mediators at final test contribute to the benefit of generating errors. In this case, we

[2] We are grateful to an anonymous reviewer for this suggestion.

would expect to see an interaction between study method and test type such that both types of test show a benefit of generating over reading but a greater benefit is seen in the MCQ test than in the target-only test. We report Experiments 2a and 2b together as they share a similar design and aim.

Method

Participants

For Experiment 2a, 36 participants, 26 female, average age 27.00 (*SD* = 12.18), were recruited from the UCL participant pool, which comprises both students and non-students, or by word of mouth. For Experiment 2b, 24 participants, 19 female, average age 24.17 (*SD* = 5.90) were recruited from the UCL participant pool. Participants were paid £4. One participant's data were excluded in Experiment 2b as he correctly generated 70% of responses at study, demonstrating prior knowledge of the Swahili vocabulary used in the experiment. As noted in Experiment 1, 30 participants gives power of 0.95 to detect a difference between scores in the Read and Generate conditions. However, in the current experiment, because there are four within-subjects conditions rather than three but the total number of items remains at 60, the final score for each condition is based on only 15 items rather than the 20 items in our previous work. To increase the power of our experiment, we decided to combine data from Experiments 2a and 2b in the analysis, giving a total of 59 participants.

Design

Experiment 2 used a within subjects design with two independent variables, study method with two levels (Read and Generate) and test type, also with two levels (MCQ and target-only).

Materials and procedure

Participants studied 60 Swahili-English word pairs from the same pool as in Experiment 1 either by reading or generating in the same way as in Experiment 1, but with no Modified Read condition. At study, the only difference between Experiments 2a and 2b was in the amount of time participants had to study Read items. In both experiments, the Generate condition consisted of presentation of the cue for 8 s during which time the participant was to enter a response, followed by presentation of the cue-target pair for 5 s. In Experiment 2a, items in the Read condition were presented for 13 s, equating the total trial time, consistent with all our previous experiments. In Experiment 2b, Read items were presented for 5 s, equating exposure to the correct answer. This experiment was run concurrently with Experiment 2a with the intention of exploring whether we would see a different effect if the cue-target pair were presented for the same duration in both conditions, unlike in our standard procedure where the pair is on screen for a much longer duration in the Read than Generate condition. Experiments 2a and Exp 2b were identical in every other respect.

At final test, half of the items (15 Read and 15 Generate) were tested in a multiple choice test consisting of the target and three new (unstudied) items as lures (standard MCQ test), while the other half were tested in a four-option forced choice test of the targets only, without the cues (target-only test). See Figure 4 for an illustration. For example, if the pair *chura-frog* appeared in the MCQ test, the cue *chura* would be shown with four options, consisting of *frog* plus three previously unseen items, e.g., *nun, owl, hoof.* If, however, it appeared in the target-only test, only the four options – *frog, nun, owl, hoof* – would be shown, without the cue word *chura*. In this case, participants were asked to select the translation they had studied earlier. Items were randomly assigned to test type on a participant by participant basis, ensuring that there were always equal numbers of Read and Generate items assigned to each test type. At final test, presentation of items was blocked by test type. In Experiment 2a, due to an error in the program, the MCQ test always came first, followed by the target-only test. In Experiment 2b half of the participants were given the MCQ test first and half were given the target-only test first.

Results

At study, six participants correctly generated one response each (four in Experiment 2a and two in Experiment 2b). These items were excluded from the following analysis.

Figure 5 shows the mean final test scores for Read and Generate items in the MCQ and target-only tests, for all 59 participants in Experiments 2a and 2b combined. An initial ANOVA with Experiment, study method, test type and test order as the factors found no main effects or interactions involving Experiment, so we combined the data from Experiments 2a and 2b for the following analysis. A mixed 2 (Study method: Read vs Generate) x 2 (Test type: MCQ vs target-only) x 2 (Test order: Target-only first vs MCQ first) ANOVA, where the first two factors were within subjects and the last one was between subjects showed a main effect of study method, *F*(1, 57) = 15.31, *p* < .001. Generating yielded better final test performance than reading. There was no main effect of test type, *F*(1,57) = 0.65, *p* = .799. There was a main effect of test order, *F*(1,57) = 4.19, *p* = .045 and an interaction between test type and test order, *F*(1,57) = 9.49, p = .003, in that test performance was higher for whichever test came first. However, there were no other interactions with test order: The expected Generate over Read benefit was evident whichever type of test was taken first. There was no interaction between study and test type, *F*(1,57) = 0.07, *p* = .791, indicating that the benefit of generating over reading was similar in both types of test.[3] Generating led to higher final test scores in both types of test: In the target-only test, mean difference 5.76 [1.77, 9.74], *t*(58) = 2.89, *p* = .005, and in the MCQ test, mean difference 7.25 [3.39, 11.12], *t*(58) = 3.76, *p* < .001. In the target-only test 32 participants recognised more Generate than Read targets, with 14 showing the opposite pattern, while in the MCQ test 27 remembered more Generate than Read items, with 8 showing the opposite pattern.

Discussion

Experiments 2a and 2b add support to the findings of Experiment 1 by demonstrating that participants showed better recognition of targets when they had learned the translation by

[3] Two participants provided final test scores that were more than two standard deviations below the mean. When the analysis was rerun excluding these, the outcomes were broadly similar: There was a main effect of study method, *F*(1, 55) = 8.35, *p* = .006 and no effect of test type, *F*(1,55) = 2.04, *p* = .159. The main effect of test order disappeared, *F*(1,57) < 1, *p* = .599. The interaction between test type and test order remained, *F*(1,55) = 5.20, p = .026. There was no interaction between study method and test type, *F*(1,55) = 0.002, *p* = .963.

generating a response than when they had studied it by reading, suggesting that generating incorrect responses benefits memory by increasing engagement with the target. If the benefit of incorrect generation came entirely from processes operating during the final test - from participants using their guess as a link from the cue to the target - we would have expected to see a benefit for generating in the MCQ test, where the cue is present, but not in the target-only test, where it is not. That there was no interaction between study and test suggests that the errorful generation benefit can be predominantly attributed to enhanced encoding of targets when presented as feedback, with little or no role for use of the generated guess as a mediator at final test. To be clear, this is not to say that participants could not have been using mediating strategies. Rather, these did not produce a Generate benefit over and above the benefit of enhanced encoding of targets during study. However, several participants obtained the maximum score on all measures, producing a ceiling effect, which may have masked a potential interaction. To address this, we ran a third experiment, Experiment 2c, in which we modified the task to make it more difficult by inserting a 24-hour retention interval between study and test.

Experiment 2c

Method

In Experiment 2c, 40 participants, 28 female, mean age 19.87 (*SD* 1.61), studied 60 Swahili-English word pairs taken from the same pool as in the previous experiments. The data from Experiments 2a and 2b combined yielded an effect size of d_z = 0.49 for the difference between Read and Generate in the MCQ test and d_z = .38 in the target-only test. Based on these data we used G*Power to determine that a sample size of 38 was sufficient to achieve a power of 0.90 to detect a similar size effect in the MCQ test, and 0.75 to detect the effect in the target-only test (both one-tailed). Two participants failed to complete the study phase and were replaced. The procedure was the same as in Experiment 2a except that a 24-hour delay was inserted between the study and the test phases and there was no filler activity following the study phase. The order of the final test blocks was counterbalanced across participants.

Results

At study, one participant generated correct responses to 20% of the items and their data were excluded from the analysis. Across all remaining participants, four responses were correct at study and those items were excluded from the following analysis. The use of a longer retention interval fulfilled its purpose: No participant achieved a score of 100% on any measure in Experiment 2c. Figure 6 shows the mean scores for Read and Generate in the two tests.

A mixed 2 (Study method: Read vs Generate) x 2 (Test type: MCQ vs target-only) x 2 (Test order: Target-only first vs MCQ first) ANOVA, where the first two factors were within subjects and the last factor was between subjects, showed a main effect of study method, *F*(1, 37) = 49.50, *p* < .001, no main effect of Test type, *F*(1, 37) = 2.75, *p* = .106, no effect of Test order, *F*(1,37) = 0.15, *p* = .701, and no interactions (between study and test, *F*(1,37) = 3.61, *p* =.065; between test and test order, *F*(1,37) = 2.77, *p* =.105; between study, test and test order, *F*(1,37) = 0.98, *p* = .330).

Follow-up analyses confirmed that the benefit of generating over reading was present in both the target-only test, mean difference 17.1 [11.53, 22.65], *t*(38) = 6.22, *p* < .001, and the MCQ test, mean difference 9.69 [4.55, 14.83], *t*(38) = 3.82, *p* < .001. In the target-only test, 29 participants recognised more Generate than Read items, with 4 showing the opposite pattern. In the MCQ test, 28 participants gave more correct responses to Generate than Read items, with 7 performing better on Read than Generate items.

Discussion

Experiment 2c replicated the findings of Experiments 2a and 2b: Even when the Swahili cue was not present at test, participants recognised more targets for items that they had studied under Generate conditions than under Read conditions, suggesting that the act of generating a guess led to enhanced encoding of targets. Experiment 2c is also the first to demonstrate that the errorful generation benefit for novel material persists with a 24-hour delay between study and test.

It should be noted that the benefit of generating over reading in the target-only test does not indicate that participants generally solve the errorful generation task by recognising targets alone, rather than cue-target associations. In some of our previous work (e.g., Potts & Shanks, 2014, Experiment 3), final test lures were studied items. When lures in a standard multiple choice test are studied items, participants can only solve the task (which they were reliably able to do) by recognising the correct cue-target association. We have also observed the benefit of generating over reading in a cued recall test (Potts, 2014, Experiments 6 and 7), though this has not always been replicated: Clark (2016) observed identical performance for Generate and Read items with cued recall, though performance in that task was very low, which might have obscured differences, and see Potts (2014, Experiment 8) for a similar finding of identical performance. What the findings of the current experiment do show is that the errorful generation benefit is observable in a target-only recognition test, evidence of more effective encoding of targets in the Generate condition. It is worth noting that the interaction between study and test was not far from being significant ($p <$ .065). Figure 6 suggests that, numerically at least, while generating was more effective than reading in both types of test, Read items benefitted more from the presence of the cue at final test. However, this is in contrast to Experiment 2a and 2b, in which Generate items received a greater boost from the presence of the cue than Read items did. An obvious difference between the two studies is that Experiment 2c used a longer (24-hour) retention interval. It would be interesting for future research to explore how the effect of incorrect generation on memory may vary over different retention intervals.

Experiment 2, then, provided evidence that generating led to stronger encoding of targets than reading did. Why might this be? As we have discussed, the findings of Experiment 1 suggest that a critical factor is the information gap created by the requirement to generate a response in the absence of the target. In that experiment, the Generate and Modified Read conditions both offered the opportunity for links to be made between cue, guess and target during study, but only Generate items were better remembered relative to Read items. Put differently, the act of generating a guess was not sufficient to lead to a benefit over reading: A benefit was only evident when there was an information gap between what the participant knew and what they wanted to know. Several studies have found memory to be enhanced by curiosity following incorrect responses (e.g., Kang et al., 2009, Gruber et al., 2014). Taken together, the results of Experiments 1 and 2 support the proposal that the presentation of a cue alone in the Generate condition creates an information gap that the participant is motivated to fill, leading to greater curiosity to learn the correct answer when it appears than in the Read condition, where the target is on screen from the start of the trial and no information gap is created. To explore this possibility further, in Experiments 3 and 4 we investigated whether the very act of generating a response when the target is not yet known increases curiosity to learn the correct answer by comparison with when there is no generation.

Experiment 3

The aim of Experiment 3 was to examine whether generating a response before seeing the correct answer would lead to greater curiosity to know the correct answer than simply reading the cue word without any requirement to generate. Curiosity was measured using self-reported ratings on a 7-point scale. Here we were not concerned with curiosity about particular items: It is reasonable to suppose that some items will elicit greater curiosity than others by virtue of certain characteristics they feature, for example, the extent to which they remind participants of other, known, words. Instead we were interested in whether overall level of curiosity to gain information is affected by study method: Does the very act of generating a guess increase participants' curiosity to acquire some information, and could this at least partly explain the memorial advantage of generating over reading in our errorful generation task? Experiments 3 and 4 were designed to address these questions.

In Experiment 3, in one condition (Rate Only), the cue alone was presented and participants were simultaneously asked to rate their curiosity to learn the correct answer, on a scale from 1 (not at all curious) to 7 (very curious). As soon as they had done so, they were shown the correct answer. In a second condition (Generate-Rate), the cue was presented and participants had to generate a guess as to its meaning, following which they were asked to rate their curiosity to learn the answer. Again, once the curiosity rating had been given, the correct answer was displayed for them to learn. If, as we propose, the very act of generating a response to an unfamiliar cue arouses curiosity, then we would expect curiosity ratings given after generating a response to be higher than curiosity ratings given on immediate presentation of the cue, with no requirement to generate a response. Of course, we could not include a Read condition in this comparison, since there is no information gap when cue and target are presented together and thus no opportunity for curiosity to be aroused. In view of this, and of our finding in Experiment 1 that the information gap is critical to the errorful generation benefit, we did not expect to see differences in memory performance between the Rate Only and Generate-Rate conditions in Experiment 3.

We were unsure whether we would be able to elicit a range of curiosity ratings in response to vocabulary items. Although an association between curiosity and memory has been demonstrated many times (e.g., Kang et al., 2009, Gruber et al., 2014), stimuli in those studies have typically been trivia questions, which by their nature are likely to elicit a range of degrees of curiosity. The same is not necessarily true of vocabulary items – there is no intrinsic reason why someone should be much more curious about one word than another, particularly for vocabulary in a foreign language that they are never likely to use. Participants may therefore find it unintuitive to rate their curiosity for this kind of item, leading to an inclination to give the same rating throughout the task and making it difficult to detect a difference in level of curiosity between the two rating conditions if such a difference exists. In this experiment and in Experiment 4, we therefore used unusual English words as stimuli in the expectation that participants would be more interested in learning new words in their own language and hence might be more likely to engage seriously with the curiosity rating task.

We hoped that this would allow any difference in curiosity level between the two rating conditions to emerge.

Method

Participants

Twenty-four participants were recruited through opportunity sampling. Two were excluded because they failed to follow the instructions. Of the remaining 22 participants, 16 were female and the average age was 21.91 (*SD* = 7.99). As we had no similar previous experiment on which to base power calculations, we used the precision planning approach advocated by Cumming (2012). Assuming a correlation of 0.9 between our two within-subject conditions (Rate Only and Generate-Rate), we calculated that 19 participants would be sufficient to give 99% assurance that our margin of error for the difference between the two rating conditions would be no more than 0.3σ.

Materials

Stimuli were 60 unusual English words with their more usual synonyms (e.g., *hispid-bristly*) taken from a pool of word pairs used by Potts and Shanks (2014). Three lures were associated with each word pair for use in a final multiple choice test: These consisted of targets from other studied items.

Design

The experiment employed a within-subjects design with one independent variable, rating condition, with two levels (Rate Only and Generate-Rate).

Procedure

Participants studied 60 unusual English words (e.g., *hispid-bristly*). For half of the words, participants were shown the cue and simultaneously asked to rate, on a 7-point scale, how curious they were to know its meaning (*Rate Only* condition). Curiosity ratings were self-paced and were immediately followed by presentation of the target, which remained on screen for 5 seconds. For the other half of the words, participants saw the cue and were asked to generate a one-word synonym and type it in, which they had 10 s to do (*Generate-Rate* condition). They were then asked to give their self-paced curiosity rating in the same way as for the Rate Only condition, following which the correct answer was displayed for 5 s for the participant to learn. The two conditions were randomly interleaved and the order of presentation was randomised on a per participant basis. The final test was in self-paced multiple choice format. Lures were all synonyms which had been seen at study.

Results

At study, four responses, one each from four participants, were correctly generated. These items were removed from the following analyses.

The aim of Experiment 3 was to explore whether generating a guess increases curiosity to know the answer. Curiosity ratings given to Generate-Rate items (*M* = 4.72, *SD* = 1.29) were higher than those given to Rate Only items (*M* = 4.50, *SD* = 1.26), *t*(21) = 2.49, *p* = .021, mean difference 0.23 [0.04, 0.41] supporting the hypothesis that generating a guess made participants more curious to know the correct answer. Fourteen participants gave higher ratings after generating than with no generation, while six gave higher ratings in the Rate Only condition, with two ties. For completeness we report, in the Appendix, gamma correlations between curiosity ratings and memory performance.

As expected, there was no difference in final test performance between the Rate Only (*M* = 60.00, *SD* = 16.87) and Generate-Rate (*M* = 61.61, *SD* = 17.65) conditions, *t*(21) = 0.63, *p* = .537, mean difference 1.61 [-3.72, 6.94].

Discussion

Experiment 3 showed that participants rated their curiosity to learn correct answers significantly higher when they were asked to generate and type in a guess on presentation of the cue than when no response was required, suggesting that the act of overtly generating a guess increased curiosity to know the answer by comparison with making a curiosity rating without the requirement to produce an overt response. There was no difference in final test performance between the two conditions. This was as expected, since the benefit of generating incorrect responses that we are exploring in the current study is a benefit *by comparison with reading the cue together with the target*. In Experiment 3 we did not include a Read condition because it would make no sense to ask participants how curious they were to know the answer in a case where the answer had already been presented to them: Curiosity level would presumably be zero in this situation. There is therefore an information gap in both the Rate Only and the Generate-Rate conditions.

Although participants were not explicitly asked to generate responses in the Rate Only condition, asking them to reflect on their curiosity to know what the word meant is likely to have encouraged participants to search their memory to see whether they already knew the meaning, leading to a similar outcome for Rate Only items as for Generate-Rate items in terms of final test performance (see Metcalfe & Kornell, 2007, for a related finding where simply presenting a cue alone was sufficient to eliminate a generation effect). Put differently, both searching memory for a possible meaning and overt generation of a response boost memory for targets sufficiently to put them above the threshold for recognition in the final MCQ test. Since the Generate-Rate condition involves both memory search and overt generation, items in this condition are likely to exceed the recognition threshold by a greater margin than the Rate-Only items, which involve a search only. Since answers at final test are either correct or incorrect, the test cannot reveal differences in memory strength between items that are correctly identified. Overt generation of a response can therefore increase curiosity to know the answer without this necessarily being reflected in final test scores. In this experiment, both conditions are likely to arouse curiosity, whereas in the standard task comparing generating and reading, only the Generate condition has the potential to do this. The difference in curiosity levels between the Read and Generate conditions of our standard task will therefore be much greater than the difference in curiosity between the Rate-Only and Generate-Rate conditions of Experiment 3. Despite this, we still observed a significant difference in self-reported curiosity levels between our Rate-Only condition, which involved no explicit generation, and the Generate-Rate condition, which required the active generation of a response to the cue, consistent with our proposal that the act of generating a response to a cue in the absence of the target arouses curiosity to learn the answer.

However, the two conditions were not ideally equated. In the Rate-Only condition, the participant sees the correct answer immediately after making their curiosity rating. By contrast, in the Generate-Rate condition, the participant has to generate an overt response before seeing the correct answer. Experiment 4 was designed to address this more directly by having participants generate a guess on every trial, capturing their curiosity ratings either before or after the generation. We expected that, in line with the findings of Experiment 3, curiosity ratings given *after* generating a response would be higher than curiosity ratings given *before* generating a response. Again, we did not expect a difference in final memory test performance since participants were generating on every trial.

Experiment 4

In Experiment 4, all items were studied in the Generate condition. For half of the items, participants gave curiosity ratings before generating a response, while for the other half they gave ratings after generating a response. In both cases, participants had 8 s in which to generate their response and 5 s to view corrective feedback. Curiosity ratings were self-paced. If it is the act of generating a response which increases curiosity, then ratings given after generating should be higher than ratings given before generating, but final test performance should be similar (since participants are generating in both cases).

Method

Participants

We aimed to recruit 32 participants, to confirm the findings of Experiment 3 in a larger sample. One extra participant was tested due to experimenter error, making a total of thirty-three participants, 19 female, average age 29.18 (*SD* = 14.45), recruited through opportunity sampling. Using G*Power, and based on the data from Experiment 3, which yielded an effect size of d_z = 0.51, we determined that this sample size would give us power of 0.8 to detect a similar effect as in the previous experiment.

Materials, design and procedure

The materials, design and procedure were as in Experiment 3, except that the Rate Only condition was replaced with a Rate-Generate condition. In this condition participants saw the cue and were simultaneously asked to rate their curiosity to know the answer. Once they had given this self-paced rating, they had 8 s in which to generate their guess as to the word's meaning, after which the correct answer was displayed for 5 s. The Generate-Rate condition was as in Experiment 3.

Results

Curiosity ratings given after generating a response (*M* = 4.63, *SD* = 1.18) were significantly higher than ratings given before generating (*M* = 4.43, *SD* = 1.16), *t*(32) = 2.61, *p* = .014, mean difference 0.20 [0.04, 0.35], providing further support for our hypothesis that the generation of a response increases curiosity. Twenty participants gave higher ratings after generating than before, while eight participants showed the opposite pattern. As expected, there was no difference in memory performance between the Rate-Generate (*M* = 55.50, *SD* = 19.69) and Generate-Rate (*M* = 58.48, *SD* = 18.53) conditions, which both involved generating responses, *t*(32) = 1.20, *p* = .238, mean difference 2.99 [-2.08, 8.05]. For completeness, we carried out gamma correlations between curiosity ratings and final test performance and these are reported in the Appendix.

Discussion

Experiment 4 provides further support for the proposal that the act of generating a response increases curiosity to know the correct answer. As in Experiment 3, participants gave higher curiosity ratings after they had generated a guess. Taken together with the findings of Experiments 1 and 2, these results suggest that errorful generation benefits memory by creating an information gap and arousing participants' curiosity to learn correct answers, leading to more effective encoding of targets following generation of a guess than when cue-target pairs are simply read. Final test performance in Experiment 4 was similar for both rating conditions. If generating increases curiosity to learn the answer, as is suggested by the findings of Experiment 4, then in the Rate-Generate condition the actual level of a participant's curiosity to learn the answer will have increased between the time of rating and the time of viewing corrective feedback, due to the intervening generation, but this will not be reflected in the recorded rating, since this was given before the act of generation occurred. If curiosity to learn the answer leads to enhanced encoding of corrective feedback, as we propose, then participants should be just as curious, and encoding just as effective, by the time the corrective feedback appears, whether the item is studied in the Rate-Only or the Generate-Rate condition. Put differently, curiosity ratings taken after generating reflect actual curiosity levels immediately before the viewing of corrective feedback. Curiosity ratings taken before generating reflect the state of curiosity before the act of generation takes place to increase that state.

General Discussion

We have previously observed an advantage of generating errors over reading during the learning of novel vocabulary items (Potts & Shanks, 2014), but the mechanism underlying this benefit is unknown. The aim of the current study was to examine whether the benefit occurs predominantly because generating activates many potential retrieval cues which act as mediators to the target at final test or because generating leads to increased engagement with targets at encoding. In four experiments we found support for the second explanation, that errorful generation benefits the learning of novel materials by enhancing target processing. Furthermore, we found evidence that this benefit arises due to a need to fill an information gap. In Experiment 1, generating only benefitted memory when there was an information gap between what the participant knew and what they wanted to know. In Experiment 2, generating yielded better memory performance on a target-only recognition test, suggesting that more effective encoding of targets had taken place for Generate than for Read items. Experiments 3 and 4 showed that generating a guess in response to an unfamiliar cue increased curiosity to learn the correct answer.

The roles of generation and an information gap in the errorful generation benefit.

Our previous work (Potts & Shanks, 2014) showed that generating errors benefitted memory but choosing an incorrect option from a choice of two (Experiment 1) or four (Experiments 2 and 3) options did not, suggesting that generating a response oneself, rather than merely selecting one, is necessary for the benefit of making errors to be observed. Our first experiment in the current study showed that it is not, however, sufficient: For generating an error to benefit memory, there must be an information gap between what the participant knows and what they desire to know. Thus it appears that generating will only benefit memory in the presence of an information gap.

In our modified Read condition of Experiment 1, participants generated responses to a cue just as they did in the Generate condition except that, in the modified Read condition, they had already seen the correct answer, thus eliminating the information gap. Final test performance in the modified Read condition was no better than performance in the standard Read condition, and significantly worse than performance in the Generate condition, suggesting that generating incorrect responses may only benefit memory when the correct answer is not yet known. In this situation, an information gap is created between what the participant knows and what they want to know. Our proposal is that attempting to generate a response to a novel cue motivates the participant to close that gap, leading to more interest in, and more effective encoding of, the target following generation than during study of a Read item.

Enhanced encoding of feedback

Experiment 2 was designed to examine this idea more directly, exploring whether encoding of feedback was in fact enhanced following generation, by testing participants on their memory for the targets only. Participants recognised significantly more Generate than Read targets in a target-only test, suggesting that at least part of the benefit of generating errors arises from enhanced encoding of the target in the Generate condition. Why does generating enhance target encoding? In our previous work (Potts & Shanks, 2014), we found that participants consistently gave lower judgments of learning to Generate than to either Read or Choice items, suggesting that they perceived items as more difficult to learn when they had experienced incorrect generation. One possibility, then, is that generating an error made participants aware of the information gap and created a motivation to close it, leading to more engagement with the feedback when it appeared.

An alternative possibility is that the discrepancy between the participant’s generated response and the correct answer, when it appears as feedback, elicits surprise and triggers an error correction process. Butterfield and Metcalfe (2001) found that, when participants gave an incorrect response that they were confident was correct, the discrepancy between the participant’s expectations and the outcome led to a hypercorrection effect, and Fazio and Marsh (2009) found that participants were more likely to remember the font colour of feedback presented following high-confidence errors, suggesting that more attention was paid to that feedback as a result of the discrepancy between expectation and outcome. Our findings offer support to the first proposal: Experiments 3 and 4 showed that the very act of generating a guess increased curiosity to learn correct answers: Self-reported curiosity ratings were higher when participants generated a response before making a rating than when they simply made a rating without overtly generating (Experiment

3) or when they made a rating before generating (Experiment 4). However, we cannot rule out a role for surprise, perhaps in conjunction with curiosity, and this would be worth exploring in future research, for example using a font colour manipulation as in Fazio and Marsh (2009), or a tone detection task as used by Butterfield and Metcalfe (2006).

Generated response as mediator

Previous attempts to explain the benefits of generating incorrect guesses have typically focused on situations where the cue is a familiar item or a meaningful question, and have highlighted the importance of a semantic association between the cue and the target (e.g., Grimaldi & Karpicke, 2012, Richland, Kao, & Kornell, 2009). The idea is that, when cue and target are related, any response to the cue is likely also to be related to the target and can therefore act as a helpful mediator between the two at final test. Presentation of the cue at study activates a semantic network associated with both the cue and the target, eliciting a response that is also related and enabling the target to be more elaborately encoded when it appears. For example, Kornell, Hays, and Bjork (2009) had participants study weakly associated word pairs, such as *pond-frog.* On presentation of *pond*, participants might picture a pond and call to mind associations such as water, fish and reeds, and perhaps select *water* as their response. Then, when *frog* is presented as the correct response, all these activated associations are easily integrated with *frog* to form an elaborate memory that is likely to be remembered.

This is a very plausible explanation for the benefit of generating incorrect guesses in situations where the cue is a familiar item or meaningful question which is capable of activating a network of semantic associations but it cannot explain our finding of an errorful generation benefit when the cue is a novel foreign vocabulary item or unfamiliar English word, since guesses are typically semantically unrelated to both cue and target. In the current study, we considered the possibility that a generated response, even when unrelated to the target, could act as a retrieval cue at later test, perhaps by creating a more distinctive context at study, but our Experiment 1 provided no support for this account: When participants generated a response in the Modified Read condition, final test performance was no better than in the standard Read condition. It may be the case, then, that even in the more typical situation where a cue evokes pre-existing semantic associations, the benefit of generating is not solely due to the ability to create a mediating link between cue, generated response and target. It is likely that motivation to fill an information gap also contributes to the advantage of generating incorrect responses over reading. Furthermore, our findings have relevance to study strategies that involve the explicit use of keywords as mediators (as opposed to incidental effects of generating responses to the cue). Evidence for the effectiveness of keyword strategies has been mixed (see Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013, for a useful review). Studies exploring this type of strategy typically have participants generate keywords while viewing the cue and target simultaneously (e.g., Pyc & Rawson, 2010). Our findings raise the possibility that a keyword strategy may be more effective when participants are instructed to generate keywords in response to a cue alone, based on the cue's orthographical or phonological features, and then make an explicit link between the keyword and the target when it appears. For example, for the pair *mashua-boat*, the participant, when presented with *mashua,* may generate the keyword *mash,* as in *mashed potato.* When *boat* is subsequently presented, the participant creates a mental image of mashed potato in a boat. This approach harnesses the power of creating a semantic association while taking advantage of the interest aroused by the presence of an information gap. Testing empirically the effectiveness of this approach relative to other strategies involving mediation would be a fruitful avenue for future research.

Some studies have not yielded a benefit for errorful generation over reading, and it is important to understand the conditions which determine when errorful generation is helpful and when it is ineffective. Kornell, Hays, and Bjork (2009) found a benefit with weakly associated word pairs but not with fictional trivia questions when total trial time was equated (Experiment 2). With the latter stimuli, participants typically generated no response at initial study, rather than an erroneous one. If generating arouses curiosity and this leads to enhanced interest in feedback, as our study suggests, the failure to generate a response to the fictional trivia questions could explain the absence of an errorful generation benefit for these stimuli. We strongly encouraged our participants to generate a response to every cue and, in most cases, they did. In addition, attempting to learn a set of answers to disparate general knowledge questions may be intrinsically less motivating than learning a coherent set of vocabulary items from the same language, which could potentially be useful to the participant in the future. However, it is encouraging to note that, while errorful generation was not positively beneficial to memory for fictional trivia questions in the Kornell et al. (2009, Experiment 2) study, neither was it harmful.

Grimaldi and Karpicke (2012) found an errorful generation benefit for semantically related word pairs (e.g., *tide-beach*) but not unrelated word pairs (e.g., *pillow-leaf*). In this case, the absence of a benefit for the unrelated pairs may have occurred because the cue had strong pre-existing associations which were likely to elicit a generated response that was highly related to the cue but not to the target. At final test, this response is likely to have come to mind, interfering with memory for the “correct”, unrelated, target. In this case, the effect of interference from a strong semantic associate of the cue may have cancelled out a benefit of generating arising from a desire to close the information gap. Fortunately, this type of scenario, where the generated response is strongly related to the cue but unrelated to the target, is unlikely to occur in a typical classroom pre-testing situation.

As noted earlier, a benefit of generating over reading with novel stimuli, similar to those used in the current study, has not always been found when the final test is in cued recall format (see Clark, 2016; Potts, 2014, Experiment 8), though it is encouraging to note that there was no detriment to generating incorrect responses in these cases. This in itself is striking, since in the task used in the current study and by Clark (2016), exposure to correct answers was much briefer in the Generate than in the Read condition. Other studies using cued recall format have shown a benefit of generating (Potts, 2014, Experiments 6 and 7). Furthermore, in this study we have not considered individual differences in memory ability or strategy use. Brewer and Unsworth (2012) found that testing provided more benefit to students with low memory ability, as measured by performance on a series of episodic memory tests, than to those with high memory ability. It would be interesting for future research to explore the role of individual differences in the errorful generation effect, perhaps looking at test performance in relation to scores obtained on the Need for Cognition Scale (Cacioppo, Petty, & Kao, 1984).[4] Overall, these disparate outcomes from errorful generation studies highlight the fact that the more we can find out about the different factors that affect learning and the effects of combinations of factors, the better we can identify effective learning régimes and tailor instruction accordingly.

It is important to note that our study, like many studies on the effect of making errors during learning, used single trial learning. In real life learning situations, especially where foreign vocabulary is concerned, it is unlikely that durable learning will occur on just a single exposure, particularly where many new items are being learned in the same session, as we had our participants attempt to do. The set-up in our study is therefore a somewhat artificial one, designed to elucidate mechanisms involved in the effect of generating errors during learning, rather than to simulate a real world learning situation. In real world learning scenarios, each item may be encountered many times before it is fully acquired. An important question, then, is whether there is any benefit or detriment to testing early on in learning, if testing results in the making of errors.

Our findings suggest that there is no reason to avoid the early introduction of tests during learning. Not only is there no detriment to generating incorrect responses but, with early testing, as soon as the correct answer to an item is acquired, memory for that item can start to benefit from a traditional testing effect, being strengthened on every subsequent test trial. We recently ran a large online experiment in collaboration with the creators of the learning software platform, Memrise (Potts & Shanks, 2017) following a research competition in which research groups submitted their best "recipe" for learning some novel vocabulary items. Several learning methods were pitted

[4] We are grateful to an anonymous reviewer for this suggestion.

against each other in an hour long study session and performance was measured in a cued recall test a week after study. Methods involving the generation of an incorrect response on first encounter performed better than or as well as those involving study and then test, and considerably better than a pure repeated study method. It will be important for future research to test systematically the effect of generating errors early in multiple trial learning. This is particularly important for the design of algorithms underpinning online learning tools, where the objective is to achieve the most effective learning with the most efficient use of the learner's time.

In conclusion, the practice of pre-testing, or of testing items before it is certain that they have been encoded, during the learning of foreign vocabulary may seem a risky strategy due to the concern that incorrect answers may persist and interfere with memory for correct answers. Our findings, however, suggest that the active process of generation is not harmful to memory even when it produces errors. The current study adds to our understanding of this issue by showing that generation alone is insufficient to produce a benefit: The act of generating a response to an unfamiliar cue, where the correct response is not yet known, stimulates a desire to close an information gap, leading to enhanced motivation to encode corrective feedback. The practice of pre-testing could therefore be a useful tool in the learning of foreign vocabulary items as long as it is designed to involve both active generation and an information gap. Online learning platforms often have learners study vocabulary items several times in paired associate format (e.g., *mashua – boat*) before introducing a test, typically beginning with a multiple choice test. Our findings suggest that there can be substantial benefit in introducing tests earlier, including tests that require learners to engage in active generation before viewing feedback. It will be useful for future research to explore the optimal mix and sequence of different types of test in the early stages of vocabulary learning.

References.

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

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., & Whitten, W. B. (1974). Recency sensitive retrieval processes in long-term free recall. *Cognitive Psychology*, **6**, 173-189.

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. (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.

Cacioppo J. T., Petty, R. E., & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment, 48,* 306–307.

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

Carpenter, S. (2011). Semantic information activated during retrieval contributes to later retention: support for the mediator effectiveness hypothesis of the testing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37 (6)*, 1547 – 1552.

Clark, C. M. (2016). When and why does learning profit from the making of errors? *Doctoral dissertation, UCLA,* retrieved from https://escholarship.org/uc/item/6zv5867p.

Clark, C. M., Yan, V. X., & Bjork, R. A. (May, 2013). Examining the mediator explanation of error-enhanced encoding: Does it matter whether the target is present or absent? Poster presented at the Annual Convention of the Association for Psychological Science, Washington, DC.

Collins, A. M., & Loftus, E. F. (1975). Spreading activation theory of semantic processing. *Psychological Review, 82 (6)*, 407 – 428.

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

Cumming, G. C. (2012). *The New Statistics.* New York: Routledge.

Dunlosky, J., Rawson, K.A., Marsh, E.J., Nathan, M.J., & Willingham, D.T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological science in the public interest,* 14 (1), 4-58.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2), 175-191.

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

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

Gardiner, J. M., Craik, F. I. M., & Bleasdale, F. A. (1973). Retrieval difficulty and subsequent recall. *Memory and Cognition, 1 (3),* 213 – 216.

Gruber, M. J., Gelman, B.D., and Ranganath, C. (2014). States of curiosity modulate learning via the hippocampus-dependent dopaminergic circuit. *Neuron, 84 (2),* 486 – 496.

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

Jacoby, L. L. (1978). On interpreting the effects of repetition: solving a problem versus remembering a solution. *Journal of Verbal Learning and Verbal Behavior,17,* 649 – 667.

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.

Karpicke, J. D., & Roediger, H. L. III. (2007). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33,* 704–719.

Koriat, A., & Bjork, R. A. (2005). Illusions of competence in monitoring one's knowledge during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(2), 187 - 194.

Kornell. N. (2014). Attempting to answer a meaningful question enhances subsequent learning even when feedback is delayed. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40(1),* 106-114.

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.

Kornell, N., & Metcalfe, J. (2006). Study efficacy and the region of proximal learning framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32 (3)*, 609-22.

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

Landauer, T. K., & Bjork, R. A. (1978). Optimum rehearsal patterns and name learning. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), Practical aspects of memory (pp. 625-632). London: Academic Press.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes, 25,* 259–284.

Leeming, F. C. (2002). The exam-a-day procedure improves performance in psychology classes. *Teaching of Psychology, 29 (3)*, 210-212.

Lehman, M., & Karpicke, J.D. (2016). Elaborative retrieval: Do semantic mediators improve memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42 (10),* 1573-1591.

Loewenstein, G. (1994). The psychology of curiosity: A review and reinterpretation. *Psychological Bulletin, 116 (1),* 75 – 98.

Metcalfe, J., and Kornell, N. (2007). Principles of cognitive science in education: The effects of generation, errors and feedback. *Psychonomic Bulletin and Review, 14 (2),* 225 – 229.

Nelson, T.O., & Dunlosky, J. (1994). Norms of paired-associate recall during multitrial learning of Swahili-English translation equivalents. *Memory, 2(3),* 325 – 335.

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 Publication 2007–2004). Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education.

Potts, R. (2014). Memory interference and the benefits and costs of testing. (Doctoral dissertation, University College London).

Potts, R., & Shanks, D. R. (2014). The benefit of generating errors during learning. *Journal of Experimental Psychology: General*. 143(2), 644-667.

Potts, R., & Shanks, D. R. (2017). The Memrise Prize: An international optimal learning research competition. Poster presented at the Psychonomic Society, Vancouver, Canada, November.

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.

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. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives of Psychological Science, 1,* 181-210.

Rowland, C. A. (2014). The effect of testing versus restudy on retention: a meta-analytic review of the testing effect. *Psychological Bulletin, 140 (6),* 1432-1463.

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

Weinstein, Y., Gilmore, A. W., Szpunar, K. K., & McDermott, K. B. (2014). The role of test expectancy in the build-up of proactive interference in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(4), 1039-1048.

Yang, C., Potts, R., & Shanks, D.R. (2017a). The forward testing effect on self-regulated study time allocation and metamemory monitoring. *Journal of Experimental Psychology: Applied.* Advance online publication, 10.1037/xap0000122.

Yang, C., Potts, R., & Shanks, D.R. (2017b). Metacognitive unawareness of the errorful generation benefit and its effects on self-regulated learning. *Journal of Experimental Psychology: Learning, Memory and Cognition. 43(7)*, 1073-1092.

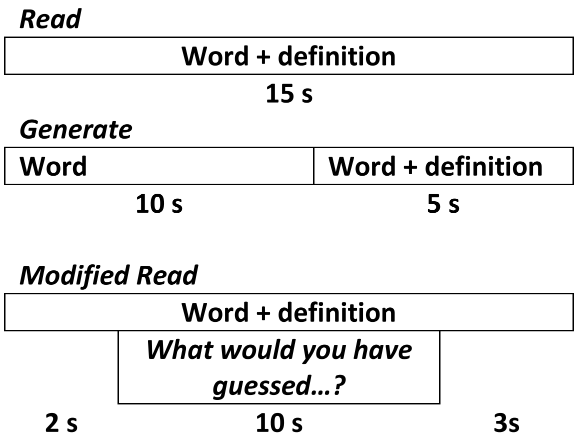

Figure 1: Study procedure in Experiment 1.

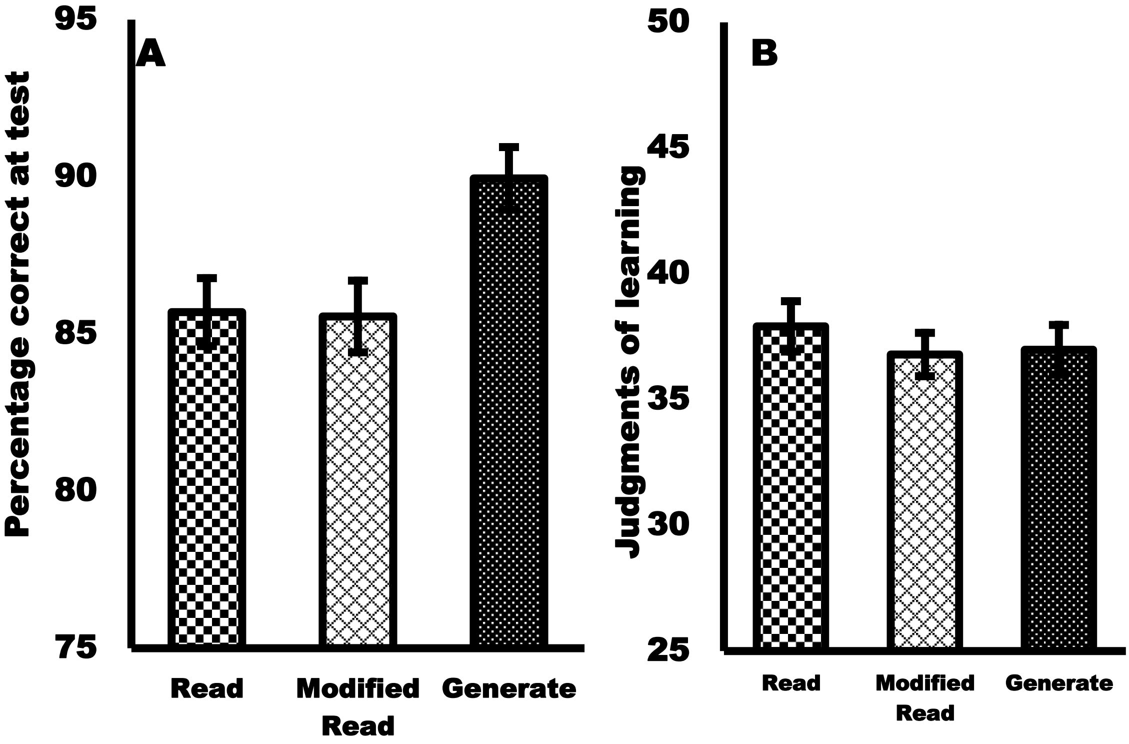

Figure 2. Final test performance (panel A) and judgments of learning (Panel B) in Experiment 1. Error bars show the within-subjects standard error.

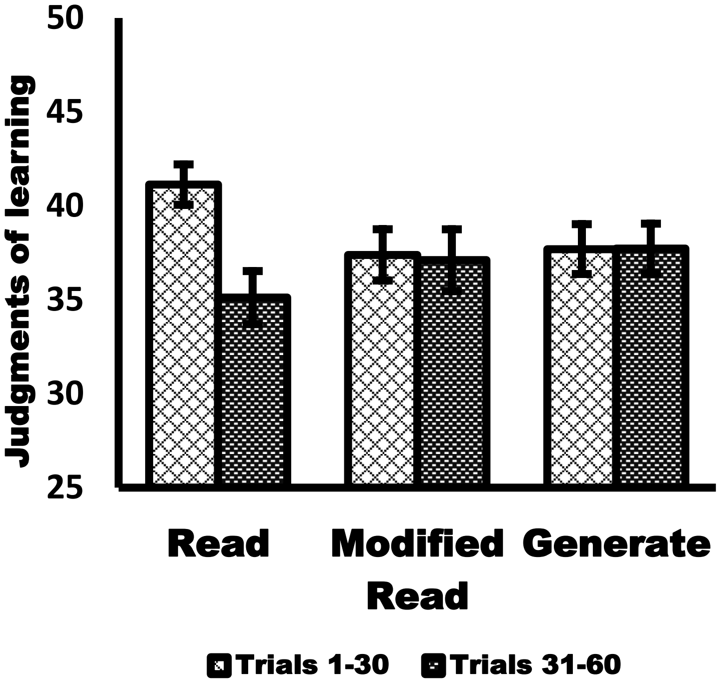

Figure 3 JOLs by study condition and time period. Error bars show the within-subjects standard error.

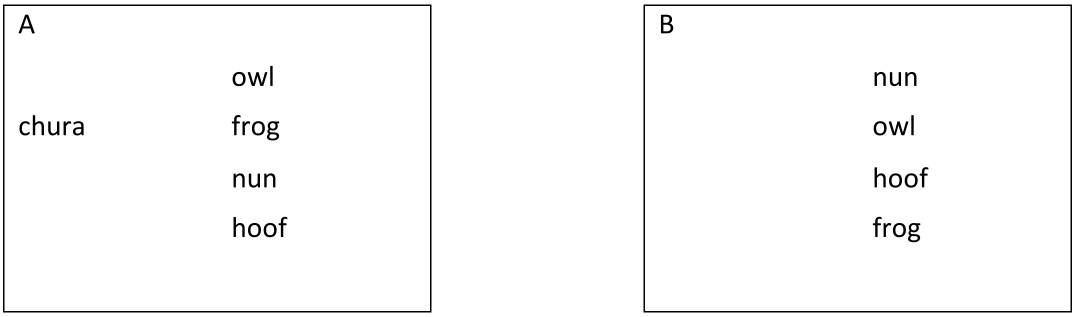

Figure 4: Test formats in Exp 2. A: Multiple choice test. B: Target-only test. In each case, the participant's task is to choose the previously studied item.

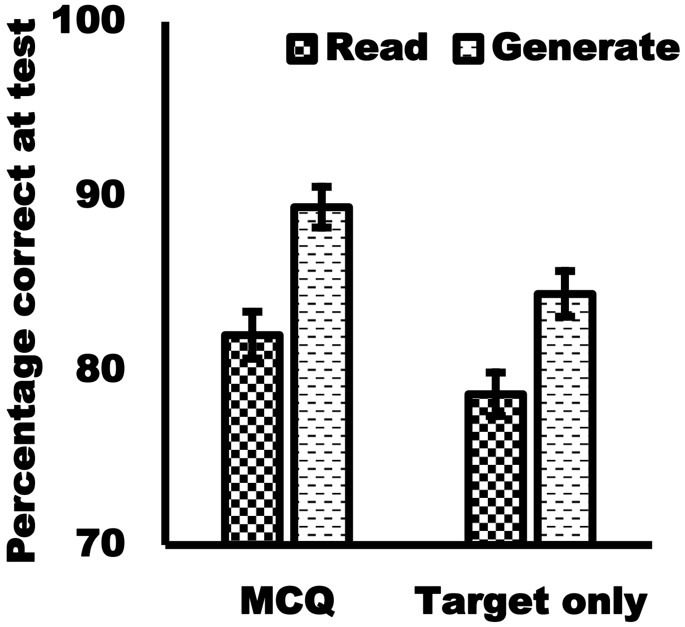

Figure 5. Final test performance in Experiment 2a and 2b combined. Error bars show the within-subjects standard error.

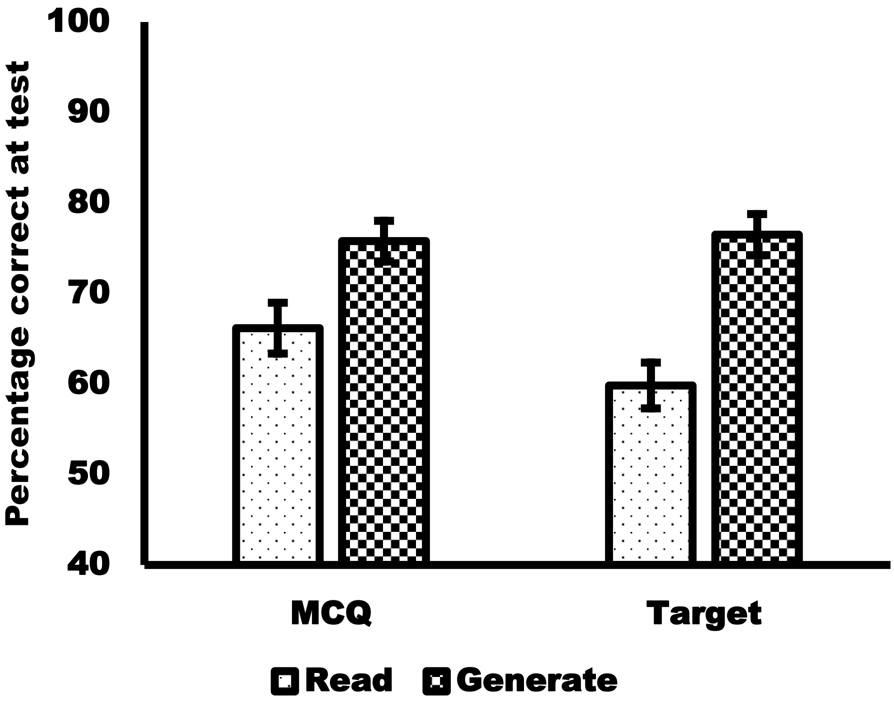

Figure 6: Mean percentage scores in Experiment 2c. Error bars show the within-subjects standard error.

Table 1: The generated responses of a typical participant during study in the Modified Read and Generate conditions of Experiment 1.

Modified Read responses				Generate responses			
Swahili cue word	Target	Modified Read response at study	LSA value	Swahili cue word	Target	Generate response at study	LSA value
nanga	anchor	hunger	0.08	pipa	barrel	seed	0.09
mfupa	bone	sleeve	0.08	zuila	carpet	water	0.14
rushwa	bribe	fast	0.02	pamba	cotton	panda	0.02
jibini	cheese	life	0.09	adui	enemy	uncle	0.05
dalasini	cinnamon	seed	0.20	bustani	garden	fire	0.06
desturi	custom	land	0.09	kaburi	grave	home	0.20
vumbi	dust	broom	0.24	ziwa	lake	open	0.14
bahasha	envelope	king	0.01	jani	leaf	know	0.08
paji	forehead	father	0.24	pafu	lung	powerful	0.04
chura	frog	cheese	0.05	samadi	manure	sand	0.15
lango	gate	freedom	0.04	tumbili	monkey	roll	0.16
tajiri	merchant	curry	0.01	rembo	ornament	rumble	0.04
fumbo	mystery	ridiculous	0.26	sala	prayer	prayer	1.00
chaza	oyster	male	0.10	nabii	prophet	messenger	0.31
sahani	plate	honey	0.10	elimu	science	igloo	0.01
inda	spite	open	0.26	usingizi	sleep	jingle	0.06
adha	trouble	ginger	0.16	theluji	snow	oil	0.03
duara	wheel	roll	0.34	chama	society	chicken	0.00
jeraha	wound	dust	0.14	vuke	steam	create	0.09
nira	yoke	nearer	0.18	hadithi	story	book	0.17

Appendix

Within participant gamma correlations between curiosity ratings and test performance.

Experiment 3

In Experiment 3 we were interested in whether mean curiosity ratings would be higher when given after generating a response than on immediate presentation of the cue, rather than in the relationship between ratings and test performance for individual items. Indeed, many uncontrolled factors will contribute to the likelihood that an individual item will be successfully retrieved from memory, so we did not necessarily expect to see such a relationship. However, for completeness, we carried out a within participant gamma correlation. This analysis showed positive but weak correlations between curiosity ratings and memory for both the Generate-Rate items (γ =.171, *SD* = 0.39), 95% CI [-0.012, 0.353], marginally significant at *t* (19) = 1.96, *p* = .065, and the Rate Only items: γ =.126, *SD* = 0.51, [-0.112, 0.364], *t* (19) = 1.11, *p* = .281. When curiosity ratings were analysed pooled across conditions, this analysis failed to reach significance, γ =.157, *SD* = 0.40, 95% CI [-0.027, 0.341], *t* (20) = 1.78, *p* = .090. Ratings were distributed evenly across the seven-point scale, with a modal value of 6 for Generate-Rate items and 5 for Rate Only items.

Experiment 4

In Experiment 4, within participant gamma correlations again showed a positive but weak relationship between curiosity ratings and memory for Rate Before items, γ = .063, *SD* = .38, 95% CI [-0.076, 0.202], *t* (31) = 0.93, *p* = .360, but not for Rate After items, γ =.002, *SD =* .41, 95% CI [-0.145, 0.149], *t* (31) = 0.03, *p* = .980. Pooling the data also revealed no significant correlation between ratings and recall, γ =.008, *SD* = .23, 95% CI [-0.075, 0.092], *t* (31) = .207, *p* = .838. This is not altogether surprising, since our findings show that curiosity ratings were higher after generating than before. Since, in Experiment 4, participants generated on every trial, actual levels of curiosity at the time of viewing feedback will have been higher than the ratings given in the Rate-Generate condition, since generation will have taken place between the rating and the viewing of feedback.

Ratings were distributed evenly across the seven-point scale, with a modal value of 5 for both Generate-Rate and Rate-Generate items.