UNIVERSITY OF CALIFORNIA

Los Angeles

# When and Why does Learning Profit from the Introduction of Errors?

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Psychology

by

Courtney McCauley Clark

2016

ABSTRACT OF THE DISSERTATION

When and Why does Learning Profit from the Introduction of Errors?

by

Courtney McCauley Clark

Doctor of Philosophy in Psychology

University of California, Los Angeles, 2016

Professor Elizabeth Ligon Bjork, Co-Chair

Professor Robert A. Bjork, Co-Chair

Generating information from memory not only gives a read out of the contents of memory—it makes those memories stronger than they would have been otherwise (Bjork, 1975). Many researchers have explored the impact of using tests as learning events with a wide range of materials (from word pairs to expository texts) and the effects are largely positive, with tested information faring better than restudied information (Roediger & Karpicke, 2006a). In the current dissertation, I explore a surprising related finding—that guessing answers, even if you are always wrong, can also aid memory when the correct answer is studied afterward (e.g., Kornell, Hays & Bjork, 2009). These results are counterintuitive, as generating guesses takes time out of study, and the guesses could very well interfere with memory for the correct answer. This phenomenon has implications for pretesting in the classroom, where students are likely to generate many errors.

The goal of my research was to explore possible mechanisms of the errorful generation effect. In Chapter 2, I assessed the role of how changing the way incorrect guesses are made affects learning from subsequent feedback. Results showed that giving semantically related guesses improved learning of related word pairs, but making guesses based on another generation rule—rhyme—was not beneficial. In Chapter 3, I report results of experiments testing whether learners are using their incorrect guesses as mediators, or helpful cues to the correct response. While other correlational evidence suggests that mediation is a plausible mechanism, the current experimental evidence does not support a mediation account. Finally, in Chapter 4, I evaluated recent evidence that suggests errorful generation benefits may be hidden in some cases by interference. Results were mixed, and suggested an evaluation of what types of associative and item information are strengthened by errorful generation.

The dissertation of Courtney McCauley Clark is approved.

Alan Dan Castel

Gordon Henderson

Barbara Knowlton

Elizabeth Ligon Bjork, Committee Co-Chair

Robert A. Bjork, Committee Co-Chair

University of California, Los Angeles

2016

*For my family*

## Table of Contents

## LIST OF FIGURES

## LIST OF TABLES

## ACKNOWLEDGMENTS

This research was supported by a grant from the James S. McDonnell foundation. Portions of this work were presented at Annual Meetings of the Psychonomic Society and the Association for Psychological Science Conventions. Research in Chapter 3 was part of a collaboration with Dr. Veronica Yan.

A huge thank you to Bob and Elizabeth Bjork for all of your guidance and creativity over the years, and for fostering such an engaging lab environment. It has been an honor to work with you. And to Alan Castel, Barbara Knowlton, and Gordon Henderson, thank you for all of your thoughtful questions and suggestions.

To all of the current and past members of CogFog, thank you for the spirited (is that the word?) discussions and thousands of calories worth of bagels. From those who have been gone for years now to the current crew, you have made the day-to-day work of grad school so much more fun—Jeri, John, Colin, Monica, all the Mikes, Carole, Veronica, Doe, Toshi, Victor, Erin, Saskia, Tara, Nick, KP, Adam, Catherine, Mary, Alex, Tyson, and many others. And to Emma, thanks for the coffee shop company that led to the words on paper part of this dissertation!

To my family, thanks for the many “school to Trader Joes” phone chats and your continued support of my (extended!) academic pursuits. To Michael, thanks for showing me that writing a dissertation is no big deal—especially when you have a pup around and treat yourself to fancy coffee. To Malcolm, thanks will be given in treats and cheese. And finally, to all of the new friends we've picked up along the way in LA and San Diego, thanks for all the support and adventures.

## VITA

| | |
|---|---|
| 2008 | B.A., Cognitive Science and Philosophy<br>Northwestern University<br>Evanston, Illinois |
| 2011-2016 | Teaching Assistant<br>Department of Psychology<br>University of California, Los Angeles |
| 2012 | M.A., Psychology<br>University of California, Los Angeles |
| 2014 | C. Phil, Psychology<br>University of California, Los Angeles |

### PUBLICATIONS AND PRESENTATIONS

Yan, V. X., **Clark, C. M.,** & Bjork, R. A. (in press). Memory and metamemory considerations in the instruction of human beings revisited: Implications for optimizing online learning. J. C. Horvath, J. Lodge, & J. A. C. Hattie (Eds). *From the Laboratory to the Classroom: Translating the Learning Sciences for Teachers.*

**Clark, C. M.,** Bjork, R. A., **(**2014). When and why introducing difficulties and errors can enhance instruction. In V.A. Benassi, C.E. Overson, & C.M. Hakala (Eds.). *Applying science of learning in education: Infusing psychological science into the curriculum.*

**Clark, C. M.,** Bjork, E. L., & Bjork, R. A. (November, 2015). On the benefits of generating anticipatory errors: Is semantic activation a necessary condition? Poster presented at the 56th annual meeting of the Psychonomic Society, Chicago, IL.

**Clark, C. M.** (June, 2014).The role of retrieval and errors in optimizing learning. Invited talk at the Center for Education Innovation in Life Sciences, UCLA, Los Angeles, CA.

**Clark, C. M**., Yue, C. L., Kraslieva, K. E., Johl, A. G., & Bjork, E. L. (November, 2014). How does relatedness affect the benefits of interleaving? Poster presented at the 55th annual meeting of the Psychonomic Society, Long Beach, CA.

Sana, F., **Clark, C. M**., Soderstrom, N. C., Bjork, R. A., Bjork, E. L., & Kim, J. A. (November, 2014). Pretesting directs attention to structure-based features of statistics concepts. Poster presented at the 55th annual meeting of the Psychonomic Society, Long Beach, CA.

**Clark, C. M**., Yue, C. L., Kraslieva, K. E., Johl, A. G., & Bjork, E. L (May, 2014) Effects of interleaving and relatedness on the learning of text passages. Poster presented at the

Annual Convention of the Association for Psychological Science, San Francisco, CA.

**Clark, C. M.**, Bjork, R. A., & Castel, A. D. (November, 2013). Errorful categorization does not impede inductive learning. Poster presented at the 54th annual meeting of the Psychonomic Society, Toronto, ON.

**Clark, C. M.**, Yan, V. X., & Bjork, R. A. (November, 2013). Why does predicting, incorrectly, an upcoming to-be-learned response enhance learning? Talk presented (by Robert Bjork) at the 54th annual meeting of the Psychonomic Society, Toronto, ON.

**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.

**Clark, C. M.** Bjork, R. A., & Bjork E. L. (November, 2012). Why does trying, and failing, to generate an upcoming to-be-learned response facilitate its later recall? Poster presented at the Annual Meeting of the Psychonomic Society, Minneapolis, MN.

**Clark, C. M**. & Bjork, R. A. (May, 2012). Why does spacing study opportunities enhance retention? Support for an accessibility principle. Poster presented at the Annual Convention of the Association for Psychological Science, Chicago, IL.

**Clark, C. M**. & Bjork, R. A. (April, 2012). An accessibility-based explanation of the spacing effect. Poster presented at the Annual Convention of the Western Psychological Association, San Francisco, CA.

**Clark, C. M.** & Bjork, R. A. (November, 2011). Interference dynamics and spacing effects, or the lack thereof: Support for an accessibility principle. Poster presented at the Annual Meeting of the Psychonomic Society, Seattle, WA.

**Clark, C. M.**, Scott, N. W., Levy, B. J. & Anderson, M. C. (September, 2009). Measuring the intrusion of unwanted memories and the regulation of mnemonic awareness. Poster presented at the Annual Meeting of the British Association for Cognitive Neuroscience, London, U.K.

**Clark, C. M.**, Safron, A., & Paller, K. A. (March, 2009). Controlling mental imagery: Investigating brain activity underlying mental image generation and inhibition. Poster presented at the 16th Annual Meeting of the Cognitive Neuroscience Society, San Francisco, CA.

Rudoy, J. D., **Clark, C. M.**, Mesulam, M. M., Weintraub, S., & Paller, K. A., (April, 2008). Detail retrieval for remote episodic memories can be impeded by improper retrieval orientation: Implications for consolidation theories. Poster presented at the 15th Annual Meeting of the Cognitive Neuroscience Society, San Francisco, CA.

## Chapter 1: Introduction and Overview

Errors, often marked in slashes of red by an evaluative pen, are feared by students and leave teachers disappointed. Students associate success with breezing through study activities then doing well on their exams. Making mistakes on tests, then, happens "after learning is over" and is cause for concern. Teachers, too, interpret fast, errorless learning as evidence that their lessons are effective, and use tests to assess learning at the end of lessons.

Decades of research have revealed that fast improvement during learning, however, can be misleading. Conditions that make learning more difficult, against all intuition, are often ones that encourage more effective processing and yield lasting learning. These study conditions, in general, provide opportunities for the current accessibility of information in memory to decrease (or remain low), so that repetition can become more effective. Termed 'desirable difficulties' (Bjork, 1994), such conditions include spacing out study sessions instead of massing them together, interleaving or mixing practice instead of blocking practice, and testing one's memory for to-be-learned materials instead of restudying those materials.

There is overwhelming evidence that tests can be used not only as assessments, as they are often used in the classroom, but as learning events (for a review, see Roediger & Karpicke, 2006a). For instance, in one experiment, participants either studied a text passage once and took three free-recall tests on it (never receiving feedback), or they studied the passage four times. On a final test, the tested group outscored the group who had only studied (Roediger & Karpicke, 2006b). Retrieving an item from memory not only serves to measure whether or not it can be remembered, but to modify it as well—making it more accessible than it would have been otherwise (Bjork, 1975).

**What Are the Consequences of Producing Errors When Tested?**

Must retrieval during practice tests be perfect to be beneficial? Difficult but successful retrieval can be especially beneficial for later memory (e.g., Carpenter & DeLosh, 2006). On most tests, however, one risks making errors. Errors could be of omission, for example, if a learner is only able to reproduce part of a studied text about the solar system. Tests may also introduce errors of commission, when information is recalled imperfectly. When asked to identify the largest planet, a student may incorrectly recall Saturn, when the correct answer is Jupiter. So in addition to producing an incomplete repetition of the original lesson, testing can lead to the retrieval of incorrect information.

How might errors retrieved during tests affect later learning? Given that retrieval strengthens what is retrieved, producing testing effects, it is conceivable that generating an error should strengthen it, and the error, then, could potentially interfere with the encoding of a later to-be-learned response. On a later test, the learner might experience confusion between the generated error (Saturn) and the correct response (Jupiter).

There is a surge of recent evidence, however, that generating guesses before studying the correct answer—even when your guesses are always incorrect—can actually be better than studying the correct answer for an extended time (e.g., Grimaldi & Karpicke, 2012; Kornell, Hays & Bjork, 2009; Potts & Shanks, 2014). In one example of such effects, and using a paradigm that has been used often by other researchers, Kornell and colleagues (2009) had participants learn weakly related word pairs through either *generate* trials or *read* trials. On generate trials, participants were presented with a cue word, "*whale*", and given 8 s to type in their guess of what the to-be-learned associate would be (e.g., "*dolphin"*). The correct to-be-learned pair ("*whale – mammal"*) was then presented for 5 s. Participants only rarely guessed the

correct associate, and when they did, those pairs were removed from further analyses. When generating versus reading was manipulated within participants, a benefit for generating a guess was found over reading the intact pair for 5 s (to match exposure time for the correct answer between generate and read conditions) on a cued-recall test after a retention interval of 5 minutes. A generation benefit was also found when read trials were extended to 13 s to equate total time on task. This effect was maintained when read versus generate was manipulated between instead of within participants. Finally, the generation advantage was found to persist at a delayed test, taken on average 38 hours after completing the learning phase. Other researchers have replicated the advantage of errorful generation with weakly related associates (e.g., Hays, Kornell & Bjork, 2013; Huelser & Metcalfe, 2012). See Figure 1.1 for a schematic of a typical trial structure and results.

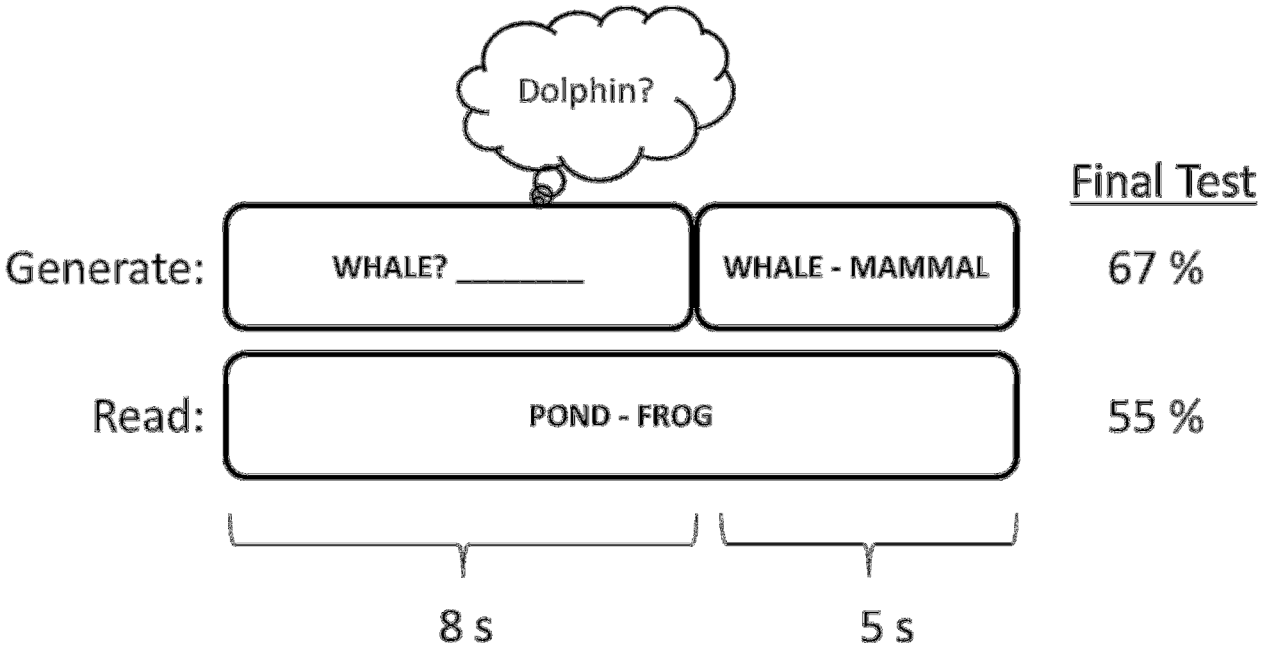


*Figure 1.1.* Schematic of typical trial structure and results for an errorful generation experiment. Participants learned 30 generate pairs and 30 read pairs. Results are from a cued-recall test from Kornell et al. (2009) Experiment 4.

Given that participants' guesses were almost always incorrect, and guessing takes time out of studying the correct pairing, benefits of incorrect guessing are surprising. As an aside, when someone guesses that the correct response to *whale* is *dolphin*, they are not wrong, and have not erred per se. There is no correct response to *whale;* there is only a response that is assigned to be correct for the purposes of the experiment. To use a term such as "errorful generation" to describe a benefit of "guessing that turns out to be inappropriate" is aspirational, as we seek to investigate what happens when mistakes are made while answering meaningful questions about meaningful materials with real answers. The relationship between the guess and the to-be-learned information is one of conflict, however, and we can say that the error is there.

Errorful generation is not always beneficial for learning. Changes to the types of materials and delays to feedback, among other things, can attenuate benefits. In this dissertation work, we hope to clarify the mechanisms by which generating errors improves learning and explore boundary conditions of the benefits of errorful generation. Knowing the boundary conditions of when errorful learning is beneficial has both theoretical and practical applications. From a theoretical perspective, an analysis of errorful generation can inform our understanding of successful generation and testing effects more generally. From a practical perspective, this work can be used to inform teachers of best practices for using pretests and practice tests in their classrooms. Students, as well, tasked with regulating their own study either in a traditional class or through online modules, can benefit. If errorful learning is shown to be broadly applicable, then students should have different attitudes about errors—seeing erring as a learning technique and not just an indication of failure.

## Characteristics of Errorful Generation Effects

I now detail properties of the errorful generation effect that have implications for theorizing about explanations. In this section I focus primarily on instantiations of the paradigm using paired associate learning to allow for easier comparison across experiments. To preface, I will focus here on the properties of the effect that are readily interpretable by the semantic activation hypothesis. Explanations relating to semantic activation are commonly offered to account for errorful generation effects.

**Semantic relationship.** Several research groups have found evidence that benefits of errorful generation can be found for related, but not unrelated word pairs (Grimaldi & Karpicke, 2012; Huelser & Metcalfe, 2012; Knight, Ball, Brewer, DeWitt & Marsh, 2012). The related word pairs used in these experiments are generally weak associates, where the chance of guessing the target on the first try is approximately 5%. The pairs have some relationship, then, but it is rare that learners generate the to-be-learned response.

When the to-be-learned associates are word stems with full word completions of those stems, for instance *bro-brother,* which share many surface characteristics but do not share a semantic relationship, guessing first has been found to be worse than studying intact (Bridger & Mecklinger, 2014).

With materials other than word pairs, schematic support may also play a role in errorful generation benefits. McGillivray and Castel (2010) taught both older and younger adults face-age pairings through either generation, where the participant saw a face and guessed the person's age before receiving feedback, or reading, where the correct face-age pairing was presented the entire time. Despite guessing incorrectly over 90% of the time, when the to-be-learned ages were consistent with the appearance of the face, both younger and older adults benefited from prior

generation. When the ages assigned to the faces were inconsistent with the appearance of the face, however, generation was no longer as helpful. Whether or not the appearance of the face could be used as a reliable cue about the age of the face seemed to determine whether there was a benefit of errorful generation.

**Timing of feedback**. To benefit from errorful generation, it may also be important that the correct response is provided shortly after the error is generated (Grimaldi & Karpicke, 2012; Hays et al., 2013; Vaughn & Rawson, 2012). Guesses with feedback provided immediately were more effective than when feedback was provided after an average of 9.5 minutes (Hays et al., 2013). In fact, generating with delayed feedback was no better than just studying the item when items were tested (in the forward direction).

**Retrieval dependence**. Benefits of errorful generation seem to rely on participants generating guesses themselves, not just studying additional words related to the targets. Grimaldi and Karpicke (2012) assessed which guesses participants were most likely to give in response to a given cue. When presented with *tide*, for instance, *wave* was the most frequently given response (22% of the time on average for all of the guesses). In a *study-lure* condition, participants studied a cue paired with its normative guess. For instance, participants studied "*tide-wave"* for 7 s before studying the cue with its to-be-learned response for 5 s, "*tide-beach."* Participants in the study-lure condition performed numerically worse than participants who studied only the correct pairing for 12 s. Participants in the study-lure condition were also more likely to have an intrusion on the final test than were participants in an error generation condition. Intrusions were defined as the studied lure for the study-lure participants and guesses generated during the study phase for participants in the generation condition. The study-lure condition, therefore, showed what we might think of as the expected costs of proactive

interference from studying alternative responses, while participants who generated their own responses did not show these costs.

**Constraining retrieval.** Cases that constrain the possibilities of responses that can be generated may also eliminate benefits of errorful generation. Using the same normative guesses as in their study-lure condition, Grimaldi and Karpicke (2012) presented participants with a cue and a two letter stem for one of the lures ("*tide-wa__*"). Instead of being able to guess any possible response to *tide*, participants had to generate a guess that completed the word stem, here, *wave*. The correct pairing ("*tide-beach"*) was then presented. Performance on the final test was significantly worse for pairs encoded after constrained generation than it was for pairs studied intact. This result is especially surprising because learners were generating a response from memory and that response was one they might have generated anyways. See Table 1.1 for a reproduction of results from Grimaldi and Karpicke (2012).

Table 1.1

*Final test performance by condition for related pairs. Results from Grimaldi and Karpicke (2012) Experiment 2.*

| Study Condition | Final Test Performance |
|---|---|
| No Pretest | .57 |
| Pretest | .68 |
| Study Lure | .47 |
| Constrained Pretest | .39 |

Bridger and Mecklinger (2014) considered cuing guesses with word stems such as *bro-* as leading to highly constrained retrieval (their real materials were in German). When constructing

materials, word stems were chosen that had two likely completions of similar strength (e.g. *broom* and *brother*). The word-stem case was compared to generating from full words where the to-be-learned pairings were weak semantic associates. Generating from the full word was considered low constraint, as participants could complete the pairing with any word that was semantically related to the cue, and participants were instructed to generate a guess that would be related but not too related to the cue. With high constraint stimuli, no benefits were found for errorful generation. In fact, performance was significantly worse than studying alone. Typical benefits of error generation over study were found for the low constraint stimuli.

**Possible Theoretical Explanations for Errorful Generation Effects**

Taken together, what can the above evidence about the constraints of errorful generation tell us about possible underlying mechanisms? There are many plausible general mechanisms for why having to guess before studying the correct answers may improve learning. For instance, perhaps having to guess increases curiosity about what the correct answer will be, and leads to increased attention to subsequent feedback. Or maybe generated responses act as mediators between the cue and target, and promote formation of triples in memory. Perhaps guesses, as they may be quite distinctive, enrich memory for the episode of encoding and could act as an additional cue at later test.

Many proposed explanations for benefits of errorful generation rely on guessing inducing helpful semantic activation. The idea is that when given a cue, perhaps the act of generating a response activates semantic information related to the cue. When feedback (the target word) is related to that activation, as in the case of related pairs, then there will be better encoding of the response. When feedback is not related to that activation, as in the case of unrelated pairs, it is less likely to be relevant and facilitate encoding of the response. For

instance, when asked to generate what goes with *whale,* a learner may activate a network including *dolphin, ocean, big, Sea World,* etc., and then type "*dolphin."* If *whale-mammal* is presented as the to-be-learned pairing, that network activation may facilitate encoding. The feature of whales being mammals could have even been activated during the generation phase. If, however, *whale-spoon* is presented as the to-be-learned pairing, it is less likely that the *whale* relevant activation will facilitate encoding of *spoon,* and exceedingly unlikely that anything about spoons was activated during the generation of *dolphin.*

Delaying feedback has also been found to diminish benefits of error generation. If generation produces short lived activation of related information, immediate feedback may be needed to reap the benefits (e.g., Vaughn & Rawson, 2012).

Grimaldi and Karpicke (2012) conceptualize relevant activation as a potential search set. Retrieval attempts, not merely studying, may be necessary to activate related information (search set). With constrained retrievals, where participants have to complete word stems instead of freely generating from memory, activation of the search set may be restricted and less likely to reach the to-be-encoded target.

**Overview of the Dissertation**

In my dissertation, I seek to reconcile and extend explanations about errorful generation benefits. First, I examine the role of types of generation rules in errorful generation by having participants make guesses based on another generation rule—rhyme. Second, I investigate the potential role of generated guesses as mediators. Third, I assess recently discovered challenges to the plausibility of the semantic activation hypothesis, the commonly-posed type of explanation of errorful generation benefits.

## Chapter 2: On the Role of Generation Rules in Moderating the Effects of Errorful Generation

Generating information from memory improves later recall of that information more than does reading the information intact, a finding typically referred to as the *generation effect* (Jacoby, 1978; Slamecka and Graf, 1978). In a typical generation effect paradigm, participants are given a cue and letter stem to complete (*rapid-f___*) for generation trials, and an intact cue-target pairing for study only trials (*rapid-fast*). On a later test, generated items tend to be remembered better than are studied items.

Generation of the target from memory does not always have to succeed in order to improve memory (e.g., Kane & Anderson, 1978; Slamecka & Fevrieski, 1983). Slamecka and Fevrieski (1983) created conditions where participants were likely to succeed (*pursue- av__d*) or less likely to succeed (*pursue- a_____*) while generating antonyms. In both cases, participants were shown the correct answer as feedback. On a final free-recall test, target words that had been correctly generated during the encoding phase were as likely to be recalled as those that were not correctly generated during the encoding phase—and both were more likely to be recalled than those items that had been merely studied.

### Errorful Generation

In the past several years, researchers have regained interest in a variation of Slamecka and Fevrieski's paradigm, and have used such studies to investigate how generating errors affects learning. In one representative study, Kornell et al. (2009) had participants learn weakly associated cue-target pairings in two ways. For some pairs, participants read the cue and attempted to generate the correct associate for 8 s (*whale- ?*) before studying the correct pairing for 5 s (*whale-mammal*). For other pairs, participants studied the correct pairing intact for 13 s.

In the generation condition, participants almost never guessed correctly, and when they did, those pairs were removed from further analysis. On a final cued-recall test, pairs for which participants had tried—and failed—to generate a response during the study phase were remembered better than pairs for which they had studied the correct answer the entire time.

Such errorful generation effects seem inconsistent with findings from AB-AD interference paradigms. The first, incorrect, response (A-B) could potentially interfere with the correct answer (A-D) on a later test (e.g., Osgood, 1949). Moreover, in studies where time on task is controlled, participants in the guess-first condition spend time generating alternative responses (that will turn out to be wrong) that they could have spent studying the correct answers.

Researchers have proposed several possible mechanisms for why errorful generation improves learning. Many suggest that the processes underlying successful generation and testing effects also produce generation effects when an alternative—errorful—response is generated (e.g., Knight et al., 2012; Kornell & Vaughn, *in press*; Potts & Shanks, 2014) For instance, there are perhaps direct effects of generation, whereby activation of a semantic network that includes the to-be-learned target enhances encoding of the target. There may also be indirect effects, where retrieval attempts can potentiate encoding of the feedback even in the absence of relevant semantic activation (Potts & Shanks, 2014). These mechanisms are discussed at length in Chapters 4 and 5.

Our main focus for the current experiments is on how the type of task (or rule) used for generation affects the way people learn from errorful generation. How does the type of generation affect the size of the benefit?

**Generation Rules**

In successful generation tasks, many types of generation rules produce memory benefits. For instance, in Slamecka and Graf's (1978) seminal paper on the generation effect, generating based on associates, category, opposites, synonyms, and rhyming all produced generation benefits. In a meta-analysis, Bertsch, Pesta, Wiscott and McDaniel (2007) assessed the role of generation difficulty as a moderator of the generation effect. They found that generation rules that they classified as easy (e.g., transposing letters) did not differ substantially from moderate (e.g., rhyming) or more difficult (e.g., mental multiplication) generation activities. All but one of the generation rules assessed in the meta-analysis produced generation effects (only unscrambling anagrams did not). The rhyming generation rule, for instance, produced a substantial generation benefit with an effect size of $d$ = .46, antonyms produced an effect size of $d$ = .28, and synonyms $d$ =.41. Other aspects of the generation task did moderate the size of the benefit. One, for instance, was how much of the target participants had to generate; effect sizes were larger when the entire target was generated ($d$ =.55) as compared to only part of it ($d$ =.32).

For errorful generation tasks, the role of generation rule has been less explored. Sometimes the task is left unspecified—participants are given a cue and asked to type in what they think the target might be (e.g., Kornell et al., 2009; Huelser & Metcalfe, 2012). In these cases, it is reasonably natural to guess an associate, especially when the to-be-learned materials are weakly related pairs. In other studies, participants are explicitly asked to guess an answer that is semantically related to the cue (e.g., Knight et al., 2012), or a member of a provided category (Cyr & Anderson, 2015).

Some efforts have been made to get participants to make guesses based on a basis other than semantic association. For instance, Knight and colleagues (2012) instructed participants that the cue-target pairings they would be learning had the same number of letters in each word, so

they should guess accordingly. Across two experiments, participants encoded semantically related or unrelated pairings by either using the same letter or a semantic guessing strategy. For unrelated pairs, guessing on either basis (number of letters or semantic association) did not produce an advantage over studying—in fact, generating produced worse performance than studying did. In contrast, for related pairings, participants remembered more generate pairings than read only pairings, regardless of whether those guesses were based on number of letters or semantic association. However, even when participants were instructed to make their generations on the basis of number of letters, those generations were often also semantic in nature. After the experiment, participants wrote out the strategy that they had used in the guessing phase and raters classified each strategy as relying either on orthographic/phonologic characteristics or semantic characteristics. When instructed to use a semantic strategy, participants reported doing so over 80% of the time. When instructed to use the number of letters strategy, participants still reported using a semantic strategy 67% of the time. When you see a cue word, even if you are trying to make your guess based on the number of letters it contains, it is perhaps natural to begin by identifying the first things that come to mind and then assess whether those words have the right number of letters. So in effect, generation type was not fully manipulated in this investigation.

In other studies, participants have been given word stems and asked to complete them—so the nature of the generation is quite different from studies where guesses are generated from more meaningful, complete word cues. When given word-stem cues that were constructed to elicit two possible (and roughly equally common) word completions, Bridger and Mecklinger (2014) found that pairs for which learners guessed first were remembered worse than those that were studied intact (except for several catch trials, if participants initially guessed correctly, they

were given the other alternative as their target pairing). Cyr and Anderson (2015) also had participants generate guesses from word stems (two guesses for each cue), what they referred to as making lexical errors. They found that making lexical guesses was worse than studying intact pairs. In both of these investigations, however, the to-be-learned pairing becomes the word stem and its completion, not complete word pairs.

Grimaldi and Karpicke (2012) used a word stem completion task in a different way. From one of their experiments, they tabulated the most common guesses participants made in response to a set of cues. In a subsequent experiment, they presented the cues and word stems for those most common guesses, and asked participants to complete the stems before presenting the correct, intact pairing. So the guesses were semantic in nature, in that the word to be filled in was related to the cue, but participants did not have freedom in their guessing (except for those participants for whom the word stem completion was their first response for that cue). This constrained guessing condition produced worse performance than did studying intact pairs.

**Goals of the Present Research**

In the current experiments, we manipulated the type of generation learners were doing in a different way—by having participants generate either rhymes or semantic associates. We took this as an example case of a generation rule for which successful generation effects have been found in the past, to assess whether it extends to errorful generation. We also manipulated the type of pairs used in the generation—either related by rhyme or semantic relationship—to assess whether the ideal type of generation depends on what you are trying to learn. In most other studies on errorful generation effects, stimuli have been semantically related or unrelated word pairs. Here, we also use pairs that are related by another rule—rhyme.

Furthermore, in Experiments 2-4, we are able to distinguish between the nature of the generation and the nature of the to-be-learned pairing, whereas in other investigations these are one in the same. That is, for studies where participants generate from word stems, the association they learn is between a word stem and the complete word, and if participants generate from cues of related or unrelated pairs, they are typically generating semantic associations. In contrast, we used semantic pairs and rhyme pairs each with two different types of generations—rhyme and semantic.

If semantic guessing appropriate to the to-be-learned target is needed for errorful generation to enhance learning, then we might expect only to find benefits where both pairs and generation are semantic in nature. But if the crucial dynamic is whether the generation type is consistent with the materials, then we would expect semantic guesses to benefit semantic pairs and rhyming guesses to benefit rhyming pairs. It could also be the case that the relationship of the cue and the target is important, but that the generation type is not as crucial, so we may see benefits of guessing rhymes for semantic pairs. This effect could also occur if participants are naturally using a semantic strategy (as they did in Knight et al., 2012), even when instructed to base their guesses on rhyme.

In addition to memory dynamics, we also investigate participants' metacognitive awareness of performing rhyming and related generation tasks. Prior work has shown that even after taking a final test where performance was better in the guessing condition, participants judged study trials as more effective for learning than guessing trials (Huelser & Metcalfe, 2012).

## Experiments 1, 2, and 3

In Experiments 1, 2, and 3, participants learned semantically related pairs or rhyming pairs by either guessing first or studying intact. After each encoding trial in Experiments 1 and 2, participants gave judgments of learning (JOLs). Participants were told what type of pairs that they were learning, so they could make the appropriate type of guesses. We expected that in the semantic condition, we would replicate prior work and find a benefit of semantic generation. For the rhyming condition, outcomes were uncertain. If generation that is consistent with the pair type benefits learning, then we would expect rhyming guesses to benefit rhyming pairs. But if semantic processing that is appropriately related to the target is needed, then we would not expect rhyming guesses to benefit rhyming pairs.

### General Method for Experiments 1-3

**Participants**. For Experiment 1, 70 participants ($M_{age}$ =33.73) were recruited using Amazon's Mechanical Turk. Participants were paid $1.50.

Participants for Experiments 2 and 3 were students from the University of California, Los Angeles (UCLA). For Experiment 2, 82 students participated ($M_{age}$ = 20.74), for Experiment 3, 69 students participated ($M_{age}$ =20.72).

**Materials.**

***Semantic pairs***: We used sixty semantically related pairs with a forward association strength between .050-.054 (Nelson, McEvoy & Schreiber, 1998). These pairs were used by Kornell and colleagues (2009).

***Rhyming pairs****:* Twenty-seven rhyming pairs were constructed. They did not have overlapping cues or targets with the semantic pairs. The goal was to have the target words very infrequently guessed, (to mimic conditions in the semantic case), and to avoid rhyming pairs that

would rhyme with other rhyming pairs or with words from the semantic pairs. Pairs were taken from Libkuman (1994) rhyming norms. Targets were the first response to a rhyme cue .25% of the time, and included in 11.5% of norming participants' response sets (Libkuman, 1994).

Experiments 1, 2 and 3 used 26 pairs of each type, while Experiments 4 and 5 used 27 pairs of each type, and Experiment 6 used all 60 semantically related pairs.

**Procedure.** Participants learned either rhyming pairs or related pairs (manipulated between participants). Generating versus studying was manipulated within participants. In Experiments 1 and 2, participants made JOLs after each trial, but in Experiment 3 they did not.

All instructions began in the same way. Participants were told that they would learn some word pairs intact, but for other pairs they would see the first word from a word pair and have to guess what the second word might be before studying the correct answer. Instructions then diverged for the two groups of participants. Participants were told either that all of their pairs would rhyme, or that they would all be related, so whenever there was a guessing trial, their guess should be of the appropriate type. Additionally, participants were told that on the final test, they would be provided with the first word from a word pair, and they would have to provide the second.

On generate trials, participants were given 8 s to guess followed by 5 s to study the intact pairing. Study only trials were 13 s presentations of the intact pairing. After each trial in Experiments 1 and 2, participants made JOLs—"How likely are you to correctly recall this item on a later test? (On a scale from 0-100)." JOLs were self-paced. In sum, participants completed 26 trials—13 generate and 13 study intact. Generate and study intact trials were randomly intermixed, and items were randomly assigned to either the generate or study intact condition for each participant. After the encoding phase, participants played Tetris for five minutes. Then, they

took a final cued-recall test where they were given the first word from a word pair and were instructed to type in the second. The final test was self-paced and no feedback was given.

## Experiment 1

For all studies, only responses that were completely correct were counted as correct. When appropriate, estimated marginal means were used for JOL and memory results.

Occasionally, participants would correctly anticipate a to-be-learned pair. This happened 5.07% of the time in the semantically related condition, and 0.48% of the time in the rhyming condition. These pairings were excluded from analyses on the final test. The JOLs for those items were also excluded from analysis.

Participants made JOLs immediately after each trial in the study phase. As shown in Figure 2.1, JOLs did not differ between rhyming ($M$ = 57.96, $SD$ = 17.10) and related ($M$ = 63.14, $SD$ = 18.67) pairings, $F(1,68) = 1.55$, $p = .22$. Overall, JOLs were higher for studied pairs ($M$ = 63.68, $SD$ = 17.18) than for generated pairs ($M$ = 57.42, $SD$ = 18.52), $F(1,68) = 47.91$, $p < .001$. These main effects are qualified, however, by a significant interaction between pair type (rhyming or semantically related) and encoding activity (generating or studying), $F(1,68) = 5.84$, $p = .018$. Paired t-tests revealed that participants had higher JOLs for studied pairs ($M$ = 67.36, $SD$ = 17.39) than for generated pairs ($M$ = 58.92, $SD$ = 19.17) for semantically related pairs, $t(37) = 6.40$, $p < .001$. JOLs were also higher for studied ($M$ = 59.99, $SD$ = 16.30) than generated ($M$ = 55.92, $SD$ = 17.89) rhyming pairs, $t(31) = 3.41$, $p = .0018$. The relative predicted benefit of studying over guessing, however, is larger for related pairs than for rhyming pairs.

As shown in Figure 2.1, on the final cued-recall test, participants in the related condition ($M$ = .79, $SD$ = .21) remembered substantially more pairs than participants in the rhyming condition ($M$ = .38, $SD$ = .28), $F(1,68) = 54.69$, $p < .001$. Overall, there was a trend toward better

performance for generated pairs ($M$ = .60, $SD$ = .32) than for studied pairs ($M$ = .57, SD = .31), $F(1,68)$= 2.99, $p$ = .089. There was no interaction between pair type and encoding type, $F(1,68)$ = 1.48, $p$ = .23. However, as the separate effects of generation on rhyming and related pairs were of prime interest for the current study, we conducted follow-up paired t-tests. For semantically related pairs, we replicate prior work in finding a significant benefit of generating ($M$ = .82, $SD$ = .20) over studying ($M$ = .76, $SD$ = .23), $t(37)$= 2.10, $p$ = .043. There was no difference between generating ($M$ = .39, $SD$ = .28) and studying ($M$ = .38, $SD$ = .26), however, for rhyming pairs, $t(31)$= .37, $p$ = .72.

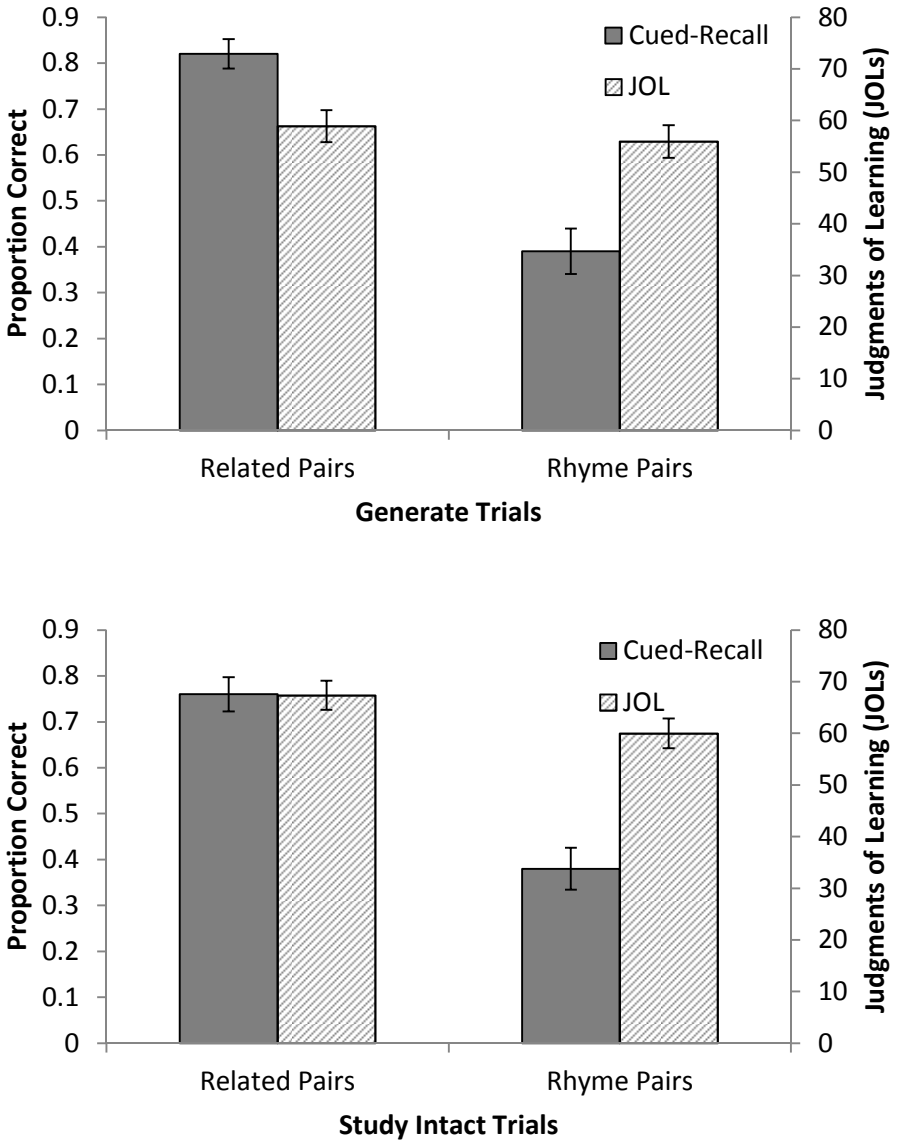


*Figure 2.1*. Mean cued-recall performance on the final test and study-phase item-by-item JOLs from Experiment 1, plotted separately for generate trials and study intact trials. Error bars (here and throughout the dissertation) are standard errors.

In Experiment 1, we found that guessing first while learning related pairs helps learning, but that benefit does not extend to rhyming pairs. Participants in both cases, however, predicted that items that were studied intact were better learned than items for which they had to guess first.

There were several procedural differences between the related condition in Experiment 1 and other relevant studies (e.g., Kornell et al. 2009), namely, we used a shorter list of pairs (26 pairs vs. the 60 pairs common in other studies) and included JOLs. Performance was numerically better than in prior work, and the size of the generation benefit was numerically smaller. To establish whether JOLs were affecting the generation benefit, we ran two follow-up experiments. Experiment 2 is a replication of Experiment 1 but conducted in the lab using participants from UCLA (instead of Mechanical Turk). Experiment 3, which was also conducted in the lab, was identical to Experiment 2 except participants did not make JOLs.

## Experiment 2

In the encoding phase, participants correctly anticipated the correct pairing 2.76% of the time in the related condition, and 1.35% of the time in the rhyming condition. These items and their corresponding JOLs were removed from further analysis.

As displayed in Figure 2.2, participants gave higher JOLs for related pairs ($M$ = 65.46, $SD$ = 16.52) than for rhyming pairs ($M$ = 56.83, $SD$ = 18.41), $F(1, 80) = 5.55$, $p = .021$. JOLs were also higher for study intact trials ($M$ = 63.71, $SD$ = 18.42) than for generate trials ($M$ = 58.58, $SD$ = 17.18), $F(1,80) = 20.31$, $p < .001$. There was no significant interaction between pair type and encoding type, $F(1,80) = 1.94$, $p = .17$.

Final cued-recall test performance is shown in Figure 2.2. Performance for related pairs ($M$ = .81, $SD$ = .19) was again substantially better than for rhyming pairs ($M$ = .40, $SD$ = .24), $F(1,80) = 88.51$, $p < .001$. Performance was better overall for study intact pairs ($M$ = .63, $SD$ = .31) than generate pairs ($M$ = .58, $SD$ = .29), $F(1,80)= 5.76$, $p = .019$. There was no significant interaction between pair type and encoding activity, $F(1,80)= 1.30$, $p = .26$. In fact, we fail to replicate the typical benefit for generating ($M$ = .80, $SD$ = .19) over studying intact ($M$ = .82, $SD$

= .18) with related pairs, $t(41)=.97$, $p = .34$. And for rhyming pairs, generated pairs ($M = .37$, $SD = .19$) were remembered worse than those that were studied intact ($M = .43$, $SD = .29$), $t(39) = 2.32$, $p = .026$).

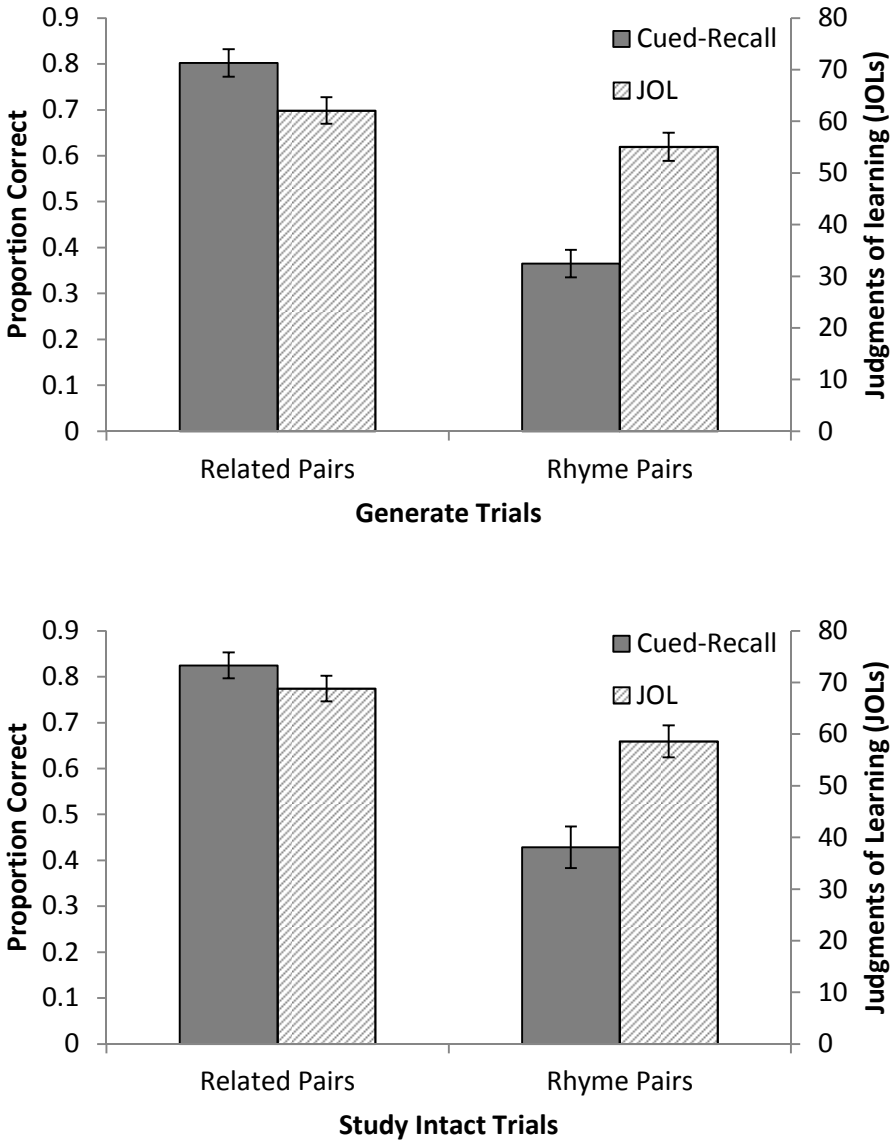


*Figure 2.2*. Mean cued-recall performance on the final test and study-phase item-by-item JOLs from Experiment 2, plotted separately for generate trials and study intact trials.

Unlike in Experiment 1, participants in Experiment 2 accurately gave higher JOLs for semantically related pairs than rhyming pairs. Just as in Experiment 1, in Experiment 2 we

found that participants anticipated better performance for pairs they studied intact as compared to pairs for which they had to guess first. However, because the memory results were different—with performance actually better overall for study intact pairs than generated pairs—these JOLs were in line with performance.

## Experiment 3

Experiment 3 was identical to Experiment 2 except participants did not give JOLs. In the study phase, participants correctly guessed 4.40% of the time in the semantically related condition and 1.13% of the time in the rhyme condition.

On the final test, as shown in Figure 2.3, related pairs ($M$ = .76, $SD$ = .17) were remembered better than rhyming pairs ($M$ = .35, $SD$ = .21), $F(1,67) = 111.37$, $p < .001$. Overall, there was no significant difference between generated (M = .58, SD = .29) and studied pairs ($M$ = .54, $SD$ = .27), $F(1, 67)= 2.68$, $p = .11$. There was, however, an interaction between pair type and encoding activity, $F(1,67)= 4.86$, $p = .031$. For related pairs, generated pairs ($M$ = .81, $SD$ = .15) were remembered better than studied pairs ($M$ = .71, $SD$ = .18), $t(34) = 2.92$, $p = .0062$, while for rhyming pairs, performance for generated pairs ($M$ = .34, $SD$ = .20) and studied pairs ($M$ = .36, $SD$ = .22) did not differ, $t(33)= .38$, $p = .71$.

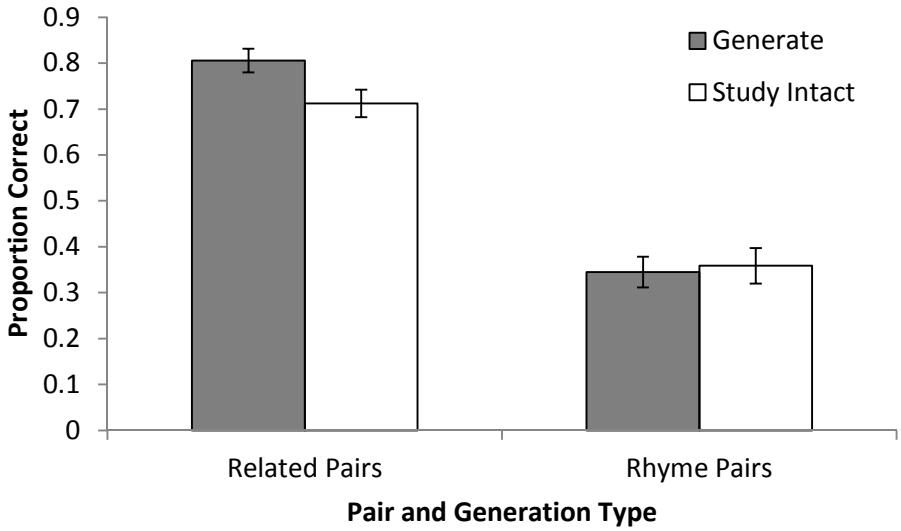


*Figure 2.3*. Mean cued-recall performance on the final test from Experiment 3.

Taken together, Experiments 2 and 3 suggest that making JOLs is not only a way for learners to communicate how they are monitoring their learning in the errorful generation task, but it also affects performance on the task. For the semantically related pairs, there was no benefit to guessing in Experiment 2, but there was a substantial benefit in Experiment 3. These results are in line with those of Soderstrom, Clark, Halamish, and Bjork (2015) who showed that making JOLs improved memory for items in the read condition of a generation effect paradigm, resulting in an attenuation of the generation effect.

Moreover, we find that while as in prior work, making guesses helped learners encode semantic pairs more than studying intact, making guesses for rhyming pairs did not help learners encode rhyming pairs more than studying intact. It is unclear, however, whether the failure to find a generation benefit for the rhyming pairs is due to the type of pair not being amenable to errorful generation, or whether the nature of the generation—rhyming—is not amenable to errorful generation. To disentangle these effects, in Experiment 4, we crossed generation type (guess a rhyme or guess a related word) and pair type (rhyme or related) in a factorial design.

## Experiment 4

### Method

Participants were 64 UCLA students ($M_{age}$ = 20.41). Type of processing (semantic generation or rhyme generation or study intact) and type of pair (semantically related pair or rhyme pair) were manipulated within participants in a 3 x 2 factorial design. Nine rhyming pairs and 9 related pairs were studied intact. For the 18 cues for which participants were prompted to do semantic generation, 9 of the to-be-learned pairings were related and 9 were rhymes. For the 18 cues for which participants were prompted to do rhyme generation, 9 of the to-be-learned pairings were related and 9 were rhymes.

Participants were instructed that they would be learning pairs of words for a later test. For some pairs, they would study the pair intact from the outset. For other pairs, before receiving the correct pairing, they would be given the first word and have to make a guess of what the second word might be. For each guessing trial, they were told that they would be given a cue about what type of guess to make, either "rhyme" indicating that they should guess a rhyming word, or "related" indicating that they should guess a word related by meaning. For guessing trials, participants had 8 s to give a guess before seeing the correct pairing for 5 s. For study only trials, intact pairs were presented for 13 s.

After all 54 pairs were presented, participants played Tetris for five minutes as a distractor task. For the final cued-recall test, each cue was presented (alone on the screen) in a random order and participants were instructed to give the correct right-hand target word. The test was self-paced.

**Results and Discussion**

Overall, participants correctly guessed semantically related pairs 1.74% of the time and rhyming pairs .35% of the time. These items were removed from analysis.

Data were analyzed using a 2(pair type: rhyme or related) x 3(study activity: rhyme guess, related guess, or study intact) repeated measures ANOVA. As shown in Figure 2.4, related pairs were remembered better than rhyming pairs, $F(1, 63) = 448.91$, $p < .001$. There was also a main effect of study activity $F(2,126) = 13.33$, $p < .001$. These main effects were qualified by an interaction between pair type and study activity, $F(2,126) = 30.94$, $p < .001$.

Follow-up tests showed that for related pairs, there was a main effect of study activity, $F(2,126) = 29.02$, $p < .001$. For trials where participants were instructed to generate a related guess ($M = .65$, SD = .22), performance was better than studying intact ($M = .49$, $SD = .26$), $t(63) = 4.69$, $p < .001$. Generating a rhyming guess ($M = .43$, $SD = .23$), in contrast, was worse than studying intact, $t(63) = 2.40$, $p = .019$.

For rhyming pairs, there was also a main effect of study activity, $F(2,126) = 6.69$, $p = .003$. Performance was marginally better for pairs studied intact ($M = .19$, SD = .19) than for pairs encoded after rhyming guesses ($M = .16$, $SD = .17$), $t(63) = 1.89$, $p = .063$. And performance was lowest for rhyming pairs encoded after related guesses (M = .12, SD = .13)—which was significantly worse than rhyming pairs encoded after rhyming guesses, $t(63) = 2.12$, $p = .038$.

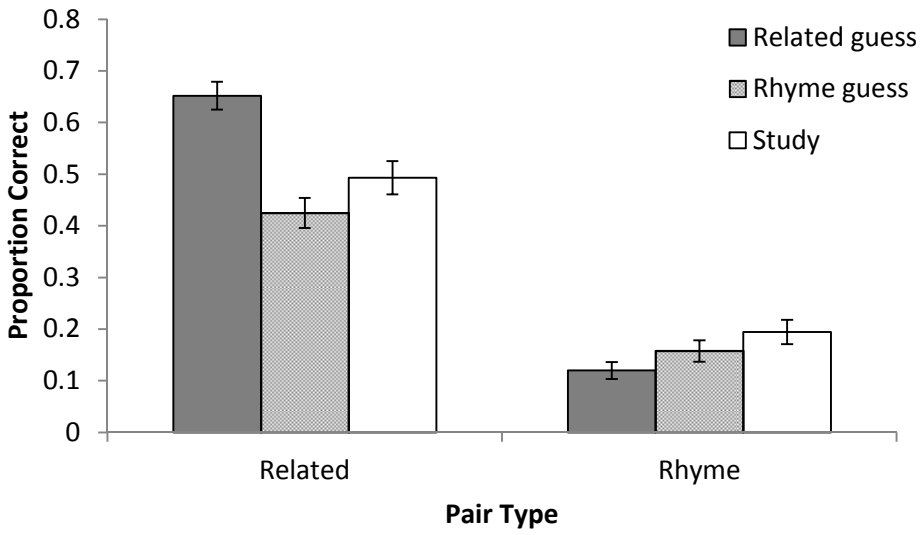


*Figure 2.4*. Mean cued-recall final test performance from Experiment 4.

The results of Experiment 4 were consistent with results from Experiments 1, 2 and 3. A generation benefit was only found for cases where pairs were semantically related and participants were instructed to make semantically related guesses. In Experiment 4, we extended this finding to a within-participants paradigm where both rhyme and related guesses were generated for both rhyme and related pairs.

One possible issue with the procedure of Experiment 4 is that participants felt deceived, because they often (for a third of all trials) were cued to do a type of guess that was not consistent with the to-be-learned pairing. To compensate, they may have covertly made both types of guesses (semantic and rhyming). Some evidence against this possibility, however, is the presence of a generation benefit when cued to do a semantic guess, but the lack of a generation benefit when cued to do a rhyming guess with related pairs.

## Experiment 5

In Experiment 5, we replicated the methods of Experiment 4 with one change. Instead of being cued with the type of guess to make in the study phase, we allowed participants to choose

what type of guess to make. Perhaps if participants are allowed to choose their most intuitive processing type for each cue, we will see benefits for both semantic and rhyme guessing. Furthermore, it is possible that participants felt deceived in Experiment 4, when we told them to complete a generation of the type inconsistent with the subsequent to-be-learned pairing. In Experiment 5, there should be no deception, as participants would be aware that the type of guess they chose to make could be wrong.

**Method**

Design details were identical to Experiment 4, except instead of being cued with what type of guess they should make, participants were instructed that for guessing trials, half of the pairs were related by rhyme and half by meaning, so they could choose whether to guess a related word or a rhyme word. Participants were 49 UCLA students, ($M_{age}$ = 20.33). One additional participant, who otherwise completed the experiment successfully, was not included in the analyses because of an absence of any rhyme guesses for related pairs.

**Results and Discussion**

Participants correctly anticipated the correct to-be-learned pairing 2.95% of the time for semantic pairs and .45% of the time for rhyming pairs. These pairs were removed from analysis.

In this experiment, participants decided whether to give a rhyme or semantic guess on an item-by-item basis. Two research assistants classified guesses by type, and the first author reconciled discrepancies. The pattern of responses participants made in the encoding phase is shown in Table 2.1.

Table 2.1

*Mean number of guesses given by type for rhyming and semantic pairs in Experiment 5.*

| | Pair Type | |
|---|---|---|
| Guess Type | Rhyming | Related |
| Rhyming | 8.65 | 4.47 |
| Semantically related | 7.41 | 10.67 |
| Rhyming and related | 0.02 | 0.14 |
| Partial or no guess | 1.92 | 2.71 |

Rhyming and semantically related guesses were the primary guess types of interest, so we focused analyses on just these pairs for which either a rhyming or related guess was given. Some guesses were both related by rhyming and meaning, and for other pairs, either only a partial guess (not a full word) or no guess was given.

A 2(pair type: rhyme or related) by 3(encoding activity: rhyme guess, related guess, or study) repeated measures ANOVA showed that performance was much better for related ($M$ = .64, $SD$ = .16) than rhyming pairs ($M$ = .15, $SD$ = .10), $F$( 1,48)= 526.30, $p < .001$. Overall, there was a main effect of encoding activity, $F(2,96) = 5.68$, $p = .005$. There was also a significant interaction between pair type and encoding activity, $F(2,96) = 26.1$, $p < .001$. Performance is shown in Figure 2.5.

For related pairs, there was a main effect of encoding activity, $F(2,96) = 18.43$, $p < .001$. Pairs with related guesses ($M$ = .78, $SD$ = .18) were remembered better than those with rhyme guesses ($M$ = .54, $SD$ = .30), $t$(48)= 5.36, $p < .001$. Pairs with related guesses were also remembered better than study intact pairs ($M$ = .59, $SD$ = .21), $t(48) = 6.01$, $p < .001$. Pairs with rhyme guesses did not differ from study intact pairs, $t(48) = 1.01$, $p = .315$.

For rhyming pairs, there was a trend toward a main effect of encoding activity, $F(2,96)$ = 2.88, $p = .061$. Pairs with related guesses ($M$ = .11, $SD$ = .12) were remembered worse than pairs

with rhyming guesses ($M$ = .17, $SD$ = .16), $t$(48) = 2.1, $p$ = .04. Pairs with related guesses were remembered marginally worse than those studied intact ($M$ = .16, $SD$ = .17), $t$(48) = 1.83, $p$ = .07. Pairs with rhyming guesses did not differ from those studied intact, $t$(48) = .19, $p$ = .85.

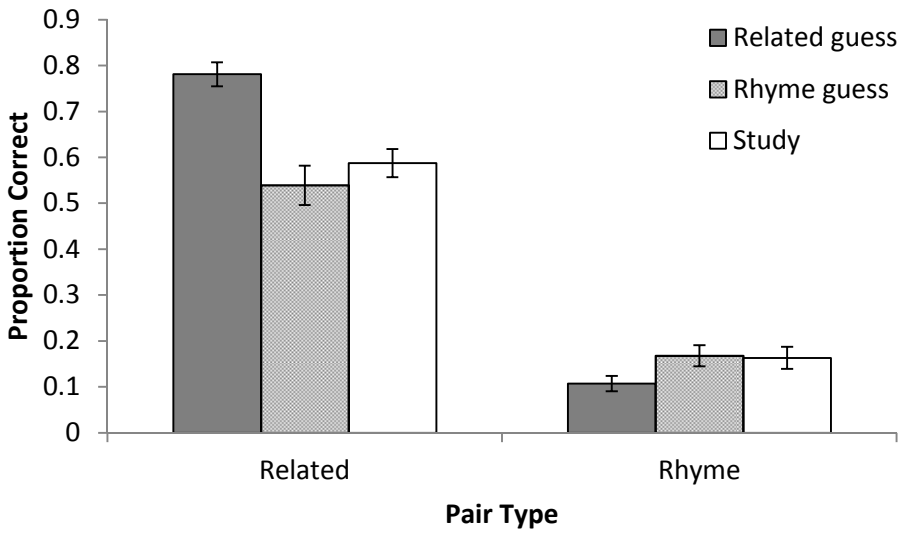


*Figure 2.5*. Mean cued-recall final test performance from Experiment 5.

## Experiment 6

Experiment 6 was conducted to evaluate a possible reason why making rhyming guesses was not as effective as making related guesses for related pairs in Experiments 4 and 5. In Experiments 4 and 5, participants were making rhyming guesses for semantic pairs in a context where they were also learning rhyming pairs. At the time of the final test, then, there is uncertainty about both the type (rhyme or related) and specific identity of the correct target. Perhaps in a context where rhyming pairs can be categorically excluded from the final response set, generation benefits will emerge.

## Method

All 60 pairs were semantically related in this experiment. For 20 pairs, participants were instructed to make rhyming guesses, for 20 pairs, participants were instructed to make semantically related guesses, and 20 pairs were studied intact. All pairs were randomly assigned to conditions for each participant, and pairs were presented in a random order. The distractor task and final test were the same as in Experiments 1-5. Participants were 57 UCLA students, ($M_{age}$ = 19.81).

## Results and Discussion

Participants correctly guessed the to-be-learned response 2.63% of the time, and these pairs were removed from analysis.

A one-way ANOVA comparing encoding activity (rhyming guess, related guess, and study intact) revealed a significant main effect, $F(1,112) = 26.35$, $p < .001$. Related guess pairs ($M = .72$, $SD = .15$) were remembered better than those studied intact ($M = .58$, $SD = .20$), $t(56) = 5.35$, $p < .001$. Related guess pairs were also remembered better than rhyming guess pairs ($M = .54$, $SD = 20$), $t(56) = 7.33$, $p < .001$. Rhyming guess pairs did not differ, however, from those studied intact, $t(56) = 1.18$, $p = .242$. Performance is plotted in Figure 2.6.

Even though all pairs were semantically related, so participants should have been able to immediately disregard any rhyming guesses that came to mind, only related, but not rhyming guesses improved performance over studying intact.

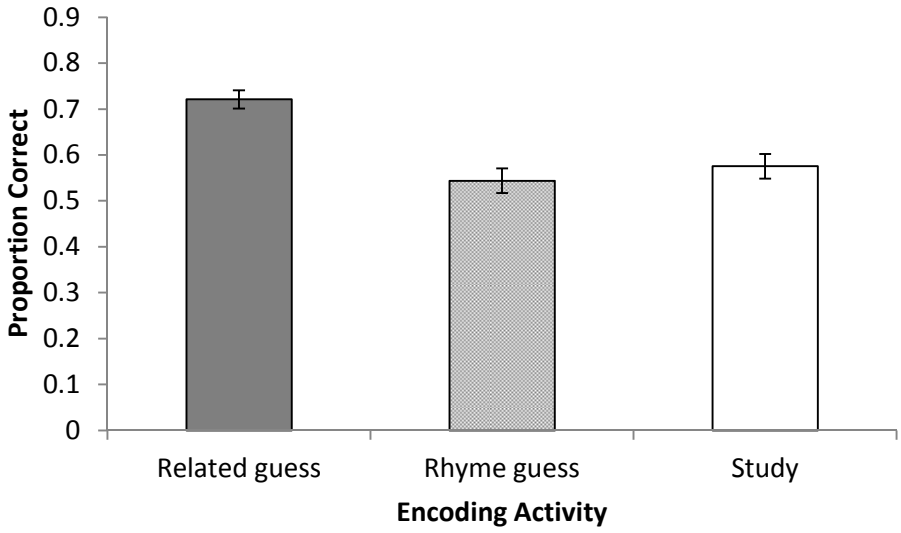


*Figure 2.6*. Mean cued-recall final test performance from Experiment 6. All pairs are semantically related.

## Aggregate Metacognitive Judgments

In each experiment, after participants completed the final test, they made several aggregate metacognitive judgments. Results are presented in Appendix A.

Across all experiments but one, participants thought that studying intact was better than guessing first. These results are in line with findings from Huelser and Metcalfe (2012) and Potts and Shanks (2014). When asked what type of generation helped them learn best, however, more participants correctly chose related generation over rhyming generation.

For the between participants studies, Experiments 1-3, participants who studied rhyming pairs were relatively more likely than those who studied related pairs to think that rhyming pairs were easier than related pairs. In Experiments 4 and 5, where both types of pairs were learned, participants were somewhat more likely to judge (correctly) that related pairs were better learned.

## General Discussion

In the current experiments, we found that while making related guesses enhanced learning for related pairs, making rhyming guesses did not enhance learning for rhyming pairs. Moreover, related guesses did not enhance learning for rhyming pairs, nor did rhyming guesses enhance learning for related pairs. This pattern of effects was apparent both between and within participants, and both when participants were told what type of guesses to make and when they were allowed to choose.

The current results are interesting when compared to the generality of typical (usually successful) generation effects. Benefits of generation are found largely regardless of the rule that guides the generation, but here, we fail to find benefits with a rhyming generation rule. In a meta-analysis, the rhyming generation rule produced an effect size of $d = .46$ (Bertsch et al., 2007). In the current investigation, generating rhymes never produced a positive effect, and in some cases produced a significantly negative effect when compared to studying intact.

Prior studies have also failed to find errorful generation benefits when learners were generating from word stems (Bridger & Mecklinger, 2012; Cyr & Anderson, 2015). Bertsch and colleagues (2007) also compared cue-based generation rules, such as rhyming, where processing of the cue using a rule produces the target, to target based rules, where no cue is needed—for instance, word fragment completion. On cued-recall tasks, both cue-based ($d = .55$) as well as target-based rules ($d = .73$) produced substantial benefits. These word fragment completion tasks are akin to the errorful generation tasks in work by Bridger and Mecklinger (2012) and Cyr and Anderson (2015) which did not enhance subsequent encoding.

It is tempting to call upon the levels of processing framework to explain why related generations help learning more than rhyming generations do for related word pairs. Maybe

rhyming generation is not as deep as related generation. Slamecka and Graf (1978) reject the levels of processing account as a suitable explanation for generation benefits. On their logic, if participants were processing the targets at a deeper level in the generate conditions, there should also be an advantage for cue memory. A shallowly processed cue, they argue, cannot be an adequate platform for generating a semantically processed target. When they tested memory for cues, however, there was no benefit of having generated the corresponding target. Additionally, the rhyming generation rule still produces generation benefits, in their studies, even though processing a word acoustically is a rather shallow level of processing. In fact, the size of the generation benefit did not reliably differ between the rhyming rule and more semantically based rules such as the antonym rule.

On the whole, it is puzzling why seemingly deeper, more demanding (successful) generation tasks do not produce bigger learning benefits than more shallow generation tasks. Testing effects, in contrast, do depend on depth, or elaboration, of retrieval. Carpenter and DeLosh (2006), for instance, had participants study word lists. As an intervening activity, participants then either read the words again, or took a recognition, cued-recall, or free-recall test on them. On the intervening tests, recognition performance was better than cued or free-recall (which did not differ). On the final test, if the format was a recognition test, intervening test type did not matter. For a cued-recall final test, however, intervening free recall was better than either cued recall or recognition. For a free-recall final test, the intervening free-recall test was again the most beneficial. So more elaborative processing on an intervening task was, on the whole, desirable.

## Chapter 3: Evaluating a Mediation Explanation of Errorful Generation

Retrieving information from memory does not merely provide a read out of what you do or do not know (e.g., Bjork, 1975). Instead, successfully retrieving information makes it more accessible later than it would have been otherwise, even as compared to additional study (see Roediger & Karpicke, 2006a, for a meta-analysis). Because of the imperfect nature of our memory systems, however, every time we take a test of some kind there is a chance that we will be incorrect. Even absent any knowledge of the benefits of testing in strengthening what is retrieved, a learner may, upon realizing that a correct answer does not immediately come to mind, either think that retrieving incorrect answers is harmful for learning the correct answer later, or just not see the point of producing an error. Instead, she might choose to seek out the correct answer over trying to respond. Retrieving incorrect information, by either intuition or extrapolation of the testing effect literature, would seem to be either a waste of time, or, by strengthening an incorrect response, make it more difficult to encode the correct answer as feedback.

In the current investigation, we seek to explain the surprising finding that under some circumstances, generating a guess before encoding the correct to-be-learned pairing—even if your guess is always incorrect—can be more effective for learning than is studying the correct pairing intact. Recent investigations of such dynamics have used English-English word pairs, obscure English or foreign language-English pairs, trivia questions, and text materials (e.g., Kornell et al., 2009; Potts & Shanks, 2014; Richland, Kornell & Kao, 2009). Here, we focus on a word pair paradigm used by Kornell et al. (2009) and others. In this paradigm, learners encode word pairs in two ways. For some pairs, they study the pair intact (*whale-mammal*) for 13 s. For other pairs, the cue is presented alone (*whale-*____?) for 8 s and learners are told to guess what

they think the target word might be. The intact, to-be-learned pairing is then presented for 5 s. Pairs for which learners guessed first were remembered better than those studied intact, even though the learners' guesses were almost always wrong (and any correct guesses were excluded from analysis). In this chapter, we focus our analysis on whether generated errors act as mediators that serve as additional retrieval routes to enhance memory for feedback.

**The Fate of Guesses in Errorful Generation**

When a correct target is encoded, what happens to the generated guess, if anything? Perhaps we see benefits and not interference of prior guessing because the guess is suppressed, or made less accessible in memory to facilitate the encoding of the correct to-be-learned response. A suppression hypothesis might predict that less accessible guesses lead to more accessible targets.

To investigate this possibility of guess suppression, memory for generated errors has been tested in several studies. Results do not support a suppression hypothesis. Guesses are remembered quite well, and participants are more likely to recall target items if they remembered their associated guess (Knight et al., 2012; Vaughn & Rawson, 2012; Yan et al, 2014). Yan et al., (2014) found that guesses were even remembered at a higher rate than were targets after a two day delay.

As errors are available in memory at the time of the final test, we can consider whether they are being used as mediators at the time of test. That is, if a learner sees "*whale-_____?"* during the study phase and guesses "*big*" before encoding "*whale-mammal,"* when the cue *whale* is presented again during the final test, *big* may come to mind, which then aids in the retrieval of *mammal*. Many researchers who have investigated errorful generation have offered guess mediation as one possible mechanism underlying benefits of guessing (e.g., Huelser & Metcalfe, 2012; Kornell et al., 2009; Vaughn & Rawson, 2012; Yan et al, 2014).

**Mediation in Successful Retrieval**

Consideration of the potential role of mediation in errorful generation and pretesting is of interest in part because mediation accounts have been offered to help explain mechanisms underlying successful retrieval. Pyc and Rawson (2010) proposed a mediator effectiveness hypothesis for testing effects. On their view, testing is beneficial because it supports the use of mediators, which they define as words, phrases, or concepts that link the cue to the target. To be effective, mediators must be retrieved at the final test, and the learner has to be able to decode the mediator—that is, find the link from the mediator to the target. After studying the pairs once, participants either restudied the pairs or took a cued recall test with immediate feedback (for 3 cycles). For every study trial, participants reported keyword mediators. The tested groups performed better on a final cued-recall test than the study only groups. Providing the cue and the mediator at final test improved performance for the study only groups, but did not help the test/restudy groups (perhaps because they had better memory for the mediators already). Mediator shifting—changing to a better mediator—was more common in testing groups than in study only groups. Pyc and Rawson suggest that the retrieval practice attempts allowed participants to realize when they had bad mediators and allowed them to pick better ones.

Carpenter (2011) provides further evidence for a mediator effectiveness explanation of testing effects: Tests are more likely than restudy opportunities to enhance activation of mediating information between cue and target. In Pyc and Rawson's study (2010), participants were instructed to generate mediators. However, in typical studies of the testing effect, there is no such instruction. How often do people generate mediating information spontaneously, and what kind of information do they generate? In Carpenter (2011) participants learned weak associates (*mother - child*) and then took an initial test (without feedback) or restudied each pair.

The final recognition test included cues, targets, and new words either related to one of the cues or unrelated to any word in the experiment. There were more false alarms for words related to the cues in the testing condition, suggesting that the potential mediators were more activated in the testing than in the restudying condition. In a second experiment, the final test was cued-recall from the original cue (*mother*), from a weak associate to the target (*birth*), or from the presumed semantic mediator (father). Overall, for both restudy and initial testing, the semantic mediators were better at eliciting the targets than were the new related cues, but this difference was larger in the tested group.

The role of mediators or additional cues has also been explored in generation effect experiments (e.g., Soraci et al., 1999). Generation cued by multiple meanings, (e.g., B__ cued by "a winged mammal" and "a wooden club"), was more effective than generation cued by two instantiations of the same meaning (e.g., B__ cued by "a winged mammal" and "a nocturnal flyer") or a single cue (e.g. B__ cued by "a winged mammal").

**The Possible Role of Mediation in Errorful Generation**

Thinking of generated guesses as mediators may also be a helpful exercise in explaining boundary conditions of errorful generation effects. For instance, semantically related, but not unrelated, word pairs show benefits of errorful generation, (e.g., Huelser & Metcalfe, 2012). (An exception, from Potts and Shanks (2014), is discussed at length in Chapter 4). Are guesses remembered differentially for related vs. unrelated pairs, which might explain benefits of one but not the other? Knight et al. (2012) tested participants on their generated guesses after the participants had learned both related and unrelated pairs using generate trials. Again, memory for guesses (for both related and unrelated pairs) was quite high, actually surpassing recall rate of to-be-learned targets. Guesses in the related case, however, were remembered better than guesses in

the unrelated case. And mediator decoding, or probability of generating the correct target given retrieval of the guess, was also higher in the related condition. Notably, though, the nature of the generation was slightly different in the unrelated as compared to the related case. In the unrelated case, participants were instructed to generate a target with the same number of letters as in the cue word. For the related pairs case, participants were instructed to guess an associate. From other experiments reported by Knight et al, when participants were instructed to give guesses with the same number of letters, they often reported also using a semantic strategy, so we might assume that the same is happening in this case as well.

But even if there was something odd about the task, and mediator retrieval did not differ between the related and unrelated pair cases, we might still expect that mediator decoding—the ability to get from the mediator to the target—might differ. We would expect that for related pairs, the mediator would be a good clue as to what the target might be, whereas for unrelated pairs the mediator would not be a good or appropriate clue (Huelser & Metcalfe, 2012). Cyr and Anderson (2015) use similar logic to explain why guessing related words, but not completing word stems, creates errorful generation effects. Conceptually related guesses, on their view, can act as stepping stones to the target, whereas completing word stems might just add retrieval noise, as the guess is not necessarily semantically related to the to-be-learned target. In sum, then, differential mediator effectiveness can plausibly help us explain the relatedness boundary condition of errorful generation.

Another notable boundary condition of the errorful generation effect is that with weakly related word pair materials, feedback must be given immediately to show benefits of generation. Vaughn and Rawson, (2012), manipulated whether feedback was given immediately or after a delay and then tested participants on their generated guesses before testing them on the to-be-

learned target. They posited that perhaps when feedback is immediate, the cue, guess, and target are incorporated into one memory trace. But when feedback is delayed, two separate traces may be formed in memory—one with the cue and guess, the other with the cue and target. On a final test, then, there might be a higher probability of source confusion, with guesses being confused with targets. However, both when feedback was given immediately and when feedback was given after a delay during learning, participants were able to remember a majority of their guesses, and guess recall did not depend on timing of feedback. Vaughn and Rawson argue that their findings make the source confusion account—where participants get their guesses confused with the correct targets—less likely.

When materials are more elaborate, several studies have shown that feedback does not have to be given immediately (Kornell, 2014; Richland et al., 2009). Kornell (2014) found that with trivia questions, generation effects appeared even if feedback was delayed by a day. Kornell argued that real trivia questions could have elicited better, more memorable mediators (as compared to word pairs), perhaps in the form of wrong answers. A mediation account, then, is perhaps helpful when explaining why benefits occur with elaborate, but not simple to-be-learned materials, but it is not especially helpful when explaining the differential effects within simple materials (why the generation effect is found with immediate but not delayed feedback).

As another piece of evidence that suggests that guesses might be used as mediators in errorful generation paradigms, Huelser and Metcalfe (2012) showed that final-test reaction times were slower for items for which an error had been generated during the learning phase. This delay, they argue, could reflect the process of thinking through one's error to get to the target.

**Goals of the Present Research**

In all of the above experiments, memory for guesses and memory for targets are measured and assessed with correlational techniques. Here, we seek to manipulate the likelihood of mediation. The premise of the current experiments is that if generated guesses act as mediators, giving participants the opportunity to generate better, more relevant guesses should improve performance on the final test. To pursue this possibility, we compared the usual errorful generation task, where participants see the cue and guess a possible target before seeing the correct pair, to a case where guesses (alternative responses) are given after the correct pair is shown. In other words, guesses are generated in the presence or absence of the to-be-learned target. We refer to the case where learners see the cue alone and come up with a guess of what the target could be as the *absence condition* and we refer to the case where learners see the cue and target together and come up with an alternative response to the cue as the *presence condition*.

Our basic logic is that in the presence condition, a response is still being generated, but it is generated in the presence of the to-be-learned cue-target pairing, which should increase its potential as a mediator. That is, insofar as the alternative response is a mediator, it might be more relevant in the presence case, because the critical information is already available. For instance, if the complete pairing is *jelly-bread*, when asked to give an alternative response, the participant might type *"peanut butter."* On a later test, when the participant is given *jelly*, *peanut butter* might come to mind as an additional retrieval route to bread. In the absence condition, however, in the encoding phase the participant might have seen "*jelly- ____?"* and guessed *"fish,"* which might not be as helpful on the later test as *peanut butter*, given that it is likely to be less related to the to-be-learned target.

Interpreting this manipulation instead in the context of a semantic-activation or search-set account, we might also imagine even bigger benefits of the presence as compared to the absence condition, as it is more likely that in the presence condition participants will activate relevant information.

To consider competing hypotheses, however, there may be some worry that by potentially limiting the scope of semantic activation, one might create a condition similar to constrained retrieval case in Grimaldi and Karpicke (2012) where guiding participants to generate lures with letter stems before encoding the correct target hindered memory. Moreover, if in the presence condition, participants are generating alternative responses that are more similar to the target, those responses might be more likely to interfere with the correct target. In addition, one might imagine that mediator retrieval (given cue, ability to generate guess) could be better in the absence case, whereas mediator decoding (given guess, ability to generate target) could be slightly better in the presence case.

In five experiments, we assessed the presence and absence cases both within and between participants, and with related and unrelated word pairs. In Experiment 1, we compared presence and absence conditions in a within participants design. In Experiment 2, we incorporated another test of the mediation hypothesis by evaluating memory of the target with and without the guess or alternative response given as an additional cue at the final test. If errors are acting as mediators, and the participant fails to generate that mediator at the time of the cued-recall test, providing them with the generated response should improve their recall of the target response. In Experiment 3 we used a between-participants design to determine how the absence and presence conditions differ from an active copy control. Experiment 4 assessed the effect of absence versus

presence trials for unrelated pairs, and finally, in Experiment 5 we addressed a possible procedural issue with the trial structure of presence trials.

## Experiment 1

### Method

**Participants, design, and materials.** Participants were 60 students from UCLA. Encoding activity—study intact, absence, or presence—was manipulated within participants. Seven other participants completed the experiment but were excluded for recopying the target word instead of giving an alternative response for more than half of all trials.

The to-be-learned word pairs were 60 weak associates (as used in Kornell et al., 2009) with forward association strength ranging between .05 and .054 (Nelson et al., 1998). These word pairs were used in Experiments 1, 2, 3, and 5.

**Procedure**. Word pairs were encoded in one of three ways. For 'study intact' trials, participants studied the cue and target together for 13 s. For 'absence' trials, participants were presented with the cue word and given 8 s to generate a guess. Then their guess was removed from the screen and they studied the correct pairing for 5 s. For 'presence' trials, the intact pairing was presented in the middle of the screen. Directly below it, the cue word was printed again, alongside an empty field. Participants were given 13 s to study the correct pairing and type in an alternative response. Word pairs were randomly assigned to one of the three encoding activities for each participant, and all trials were presented in a new random order for each participant.

***Instructions.***

*You will be learning pairs of words in three different ways.*

*For some pairs, you will just study the pair intact (word-pair).* ***You should try to remember the pair for your test.***

*For other pairs, you will see the first word from a word pair (word-_____), and you should type in your best guess of what the second word will be. After you guess, you will be given the correct pair-* ***this is the pair that you should remember for your test.***

*For the third type of pairs, you will see a pair of words on the screen (word-pair).* ***This is the pair you should remember for your final test.*** *On the line below, you will see the first word from the word pair again (word- _____), and you should type in your best guess of what another response (right hand) word could have been.*

Participants were then given an example for each trial type, and their attention was drawn to which was the important pair to remember in each case.

After the study phase was complete, participants played Tetris for five minutes as a distractor task. The final test had two parts. First, they were given each cue in a random order and asked to recall the correct pairing, and cautioned not to give their initial guess or alternative response. After recall of all 60 pairs had been attempted, they were shown each cue again, and instructed to try to remember what their guess or alternative response had been during the study phase.

**Results and Discussion**

Participants correctly anticipated the to-be-learned pairing on 1.42% of absence trials. These pairings were removed from analysis on the final test. Additionally, after removing the 7 participants who recopied the correct target instead of providing an alternative response on presence trials over 50% of the time, the remaining participants only recopied the target 4.50% of the time. These trials—which were the result of responses from only 6 participants—were not removed from analysis because it was unclear why participants were responding that way.

**Target memory.** Performance on the final cued-recall test for target memory (shown in Figure 3.1) was analyzed using a one-way ANOVA with three levels of types of encoding—

absence, presence, and study intact. The ANOVA revealed a significant main effect of encoding activity, $F(2,118) = 13.52$, $p < .001$. Guessing before seeing the correct target (the absence condition, $M = .74$, $SD = .18$) produced better performance than guessing after seeing the correct target (the presence condition, $M = .67$, $SD = .20$), $t(59) = 3.59$, $p = .001$. The absence condition also outscored the study only condition ($M = .64$, $SD = .19$), $t(59) = 4.76$, $p < .001$. While the presence condition produced numerically better performance than the study only condition, there was no reliable difference between the two conditions, $t(59) = 1.55$, $p = .13$.

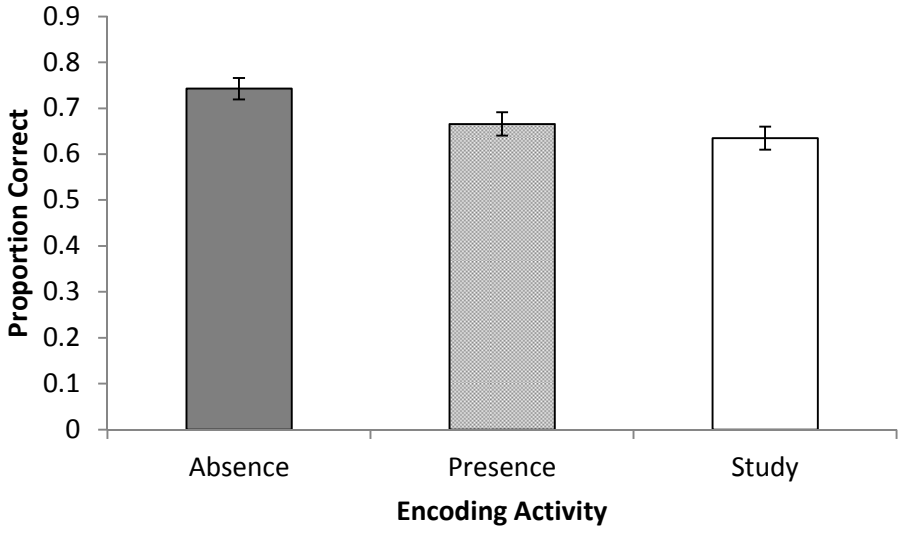


*Figure 3.1*. Cued-recall performance for target memory in Experiment 1.

**Guess/alternative response memory**. After participants attempted to recall all 60 targets, they were tested on their memory for their guesses and alternative responses that they provided during the study phase. These data were analyzed in two ways. In the first way, guesses that were correct during the study phase were excluded from analysis. In that case, participants remembered marginally more of their guesses ($M = .58$, $SD = .31$) than their alternative responses ($M = .55$, $SD = .31$), $t(59) = 1.78$, $p = .081$. In the second analysis, guesses that were correct during the study phase were not excluded from analysis. In that case, participants

remembered significantly more of their guesses ($M$ = .59, $SD$ = .30) than their alternative responses, $t(59) = 2.58$, $p$ = .012. Both methods suggest that when provided with the cue, participants have slightly better access to their guesses from the absence condition than they do their alternative responses from the presence condition.

## Experiment 2

Experiment 2 was identical to Experiment 1 except for one change to the final test. On the final test, instead of seeing the cue alone, participants were shown the cue word along with either the guess or alternative response that they had given during the study phase. Participants were told that they were given these responses to help them recall the targets. For study-intact pairings, only the cue was presented. (Given that the structure of the test for target memory involved presenting the guesses and alternative responses, we did not test memory for those responses after testing target memory). This experiment was designed to show whether better mediators (in terms of mediator decoding) were in fact generated, but deficiencies in mediator retrieval obscure the benefits. Participants were 54 students from UCLA.

### Results and Discussion

For this experiment, participants sometimes gave guesses or alternative responses that were either partial words, nonwords, or blanks. Two raters classified the responses and I resolved discrepancies. Participants gave these flawed responses for an average of 2.52 items ($SD$ = 1.80) in the absence condition and an average of 1.24 items ($SD$ = 1.61) in the presence condition. When partial guesses or typos were presented on the final test as an additional hint, then, it was unclear how helpful they would be. For example, if a participant had seen "*whale*:____?" during the encoding phase and guessed "b" before time elapsed (even though "*big*" was the intended guess), during the final test, "*whale b"* would have been presented (as compared to "*whale big*"

if the participant had typed out the entire word). Furthermore, it is unclear whether there was something different (such as perceived task difficulty) between pairs for which participants could and could not come up with guesses. For completeness, and because there are some small differences in the results depending on whether all responses are considered or only those with complete hints are considered, I present both sets of analyses below.

In the encoding phase, participants correctly guessed on 4.44% of absence trials and these items were removed from analysis. As in Experiment 1, participants were removed from the analysis if they did not understand the instructions in the presence condition, which we arbitrarily defined as responding with the correct target, instead of an alternative response, on over half of presence trials. For the remaining participants, targets were recopied on 1.67% of presence trials, and were not removed from analysis.

Considering all responses in the final test phase, a one-way ANOVA revealed a significant main effect of encoding activity (absence vs. presence vs. study only), $F(2,106) = 15.27$, $p < .001$. There was a trend for the absence condition ($M = .74$, $SD = .18$) scoring higher than the presence condition ($M = .70$, $SD = .21$), $t(53) = 1.62$, $p = .11$. The absence condition outperformed the study only condition ($M = .61$, $SD = .22$), $t(53)= 5.32$, $p < .001$. The presence condition also outperformed the study only condition, $t(53) = 3.74$, $p < .001$. Performance is shown in Figure 3.2.

Considering only word pairs with complete guesses/alternative responses, an ANOVA also showed a main effect of encoding activity, $F(2, 106) = 16.28$, $p < .001$. The absence condition ($M = .75$, $SD = .18$) showed significantly better performance than the presence condition ($M = .70$, $SD = .22$), which differs slightly from the results looking at all responses, $t(53) = 2.02$, $p = .048$. But as before, the absence condition outperforms the study only condition

($M$ = .61 $SD$ = .22), $t(53) = 5.50$, $p < .001$, and the presence condition outperforms the study only condition, $t(53) = 3.68$, $p = .001$.

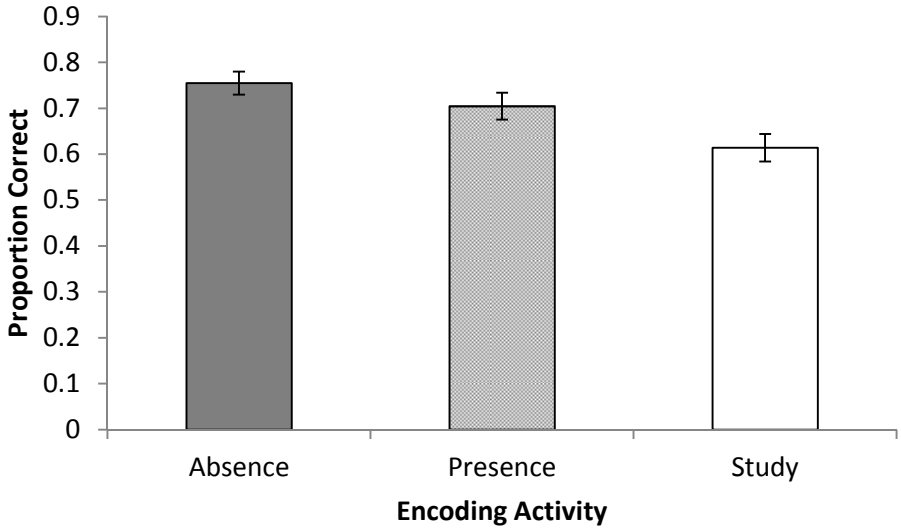


*Figure 3.2.* Performance in Experiment 2 on trials where guesses and alternative responses were entire words.

## Experiment 3

In Experiment 3, we manipulated encoding activity between participants. Furthermore, in addition to the absence, presence, and study intact conditions, we added a control condition that involved copying the to-be-learned target. That is, compared to the study intact condition, where participants merely read the pair, in the copy condition participants were instructed to type out the to-be-learned target. Copy trials appeared to be the same as presence trials on the screen, but instead of providing an alternative response, participants typed out the correct target word.

In this experiment, we sought to determine whether the absence versus presence effect persisted in a between participants design, and also to assess whether the slight benefit of the presence condition over studying observed in earlier experiments was due to the participants having to provide a response, making those trials more active.

Participants were 132 students from UCLA.

**Results and Discussion**

Participants in the absence condition correctly anticipated the to-be-learned target on 3.81% of trials, and these trials were removed from analysis.

Final test performance is shown in Figure 3.3. There was a significant effect of encoding condition (presence vs. absence vs. copy vs. study) on final test performance, $F(3, 128) = 3.20$, $p = .026$. The absence condition ($M = .71$, $SD = .15$) performed marginally better than the presence condition ($M = .63$, $SD = .22$), $t(69) = 1.79$, $p = .078$. The absence condition performed significantly better than both the copy ($M = .57$, $SD = .24$), $t(62) = 2.71$, $p = .0086$, and the study intact conditions ($M = .56$, $SD = .24$), $t(65) = 2.98$, $p = .0040$. The presence condition showed only a numerical benefit over copying, $t(63) = .95$, $p = .35$, or studying intact, $t(66) = 1.20$, $p = .24$. Finally, the copy and study intact conditions did not differ, $t(59) = .21$, $p = .83$.

The results of Experiment 3 suggest that the advantage for absence over presence learning conditions transfers to a between participants paradigm. The numeric benefit of the presence condition over the copy condition suggests that there may be some benefit of the generation in the presence case beyond any benefits of an active control condition. Lastly, the active copy control did not differ from the passive study intact condition, though we do not know whether a different effect would occur in a within-participants paradigm.

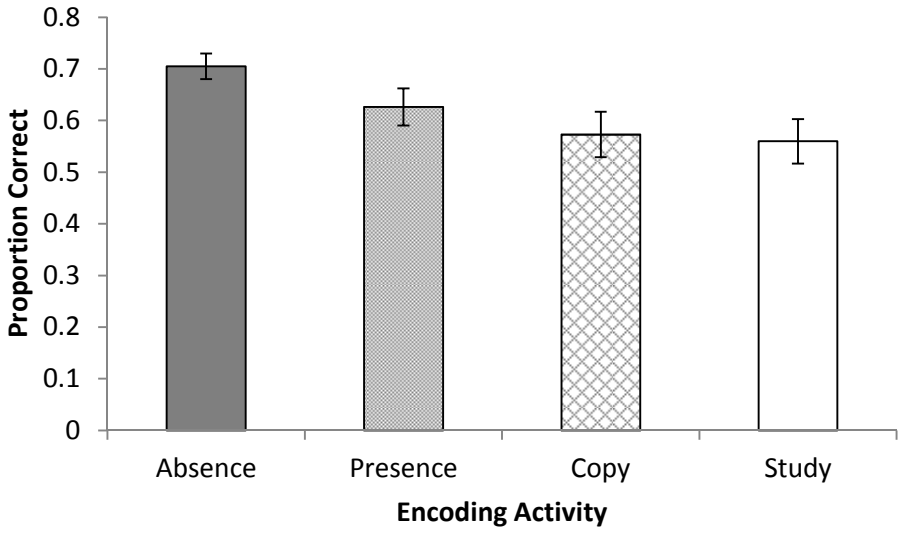


*Figure 3.3*. Final test performance for Experiment 3.

## Experiment 4

Experiment 4 was identical to Experiment 1 except that instead of using weak associates, the to-be-learned pairs were semantically unrelated. Encoding activity—absence, presence, and study intact—was manipulated within participants.

The impetus for doing Experiment 4 came out of the intuition that for novel associations—between unrelated English pairs—seeing the entire pair before generating an alternative response might be especially helpful. Unlike the absence condition, where the generated guess is likely to be related to the cue word, but unrelated to the target, in the presence condition, the learner has the ability to generate a response that is relevant to both the cue and the target.

Participants were 49 UCLA students. Five additional students completed the experiment but were excluded for providing the correct response instead of an alternative response over half of the time.

**Results and Discussion**

In the absence condition, participants did not correctly anticipate any responses. In the presence condition (after removing 5 participants from analysis for writing in the correct response instead of an alternative response over 50% of the time) the remaining participants only responded with the correct target 1.63% of the time.

For final test performance, a one-way ANOVA revealed a significant effect of encoding activity (absence vs. presence vs. study), $F(2,96) = 12.33$, $p < .001$. Performance is plotted in Figure 3.4. The absence condition ($M = .19$, $SD = .17$) did not reliably differ from, but was numerically worse than the presence condition ($M = .22$, $SD = .18$), $t(48) = 1.25$, $p = .22$. The absence condition was reliably worse than the study only condition ($M = .30$, $SD = .22$), $t(48) = 4.37$, $p < .001$. The presence condition was also reliably worse than the study only condition, $t(48) = 3.49$, $p = .001$.

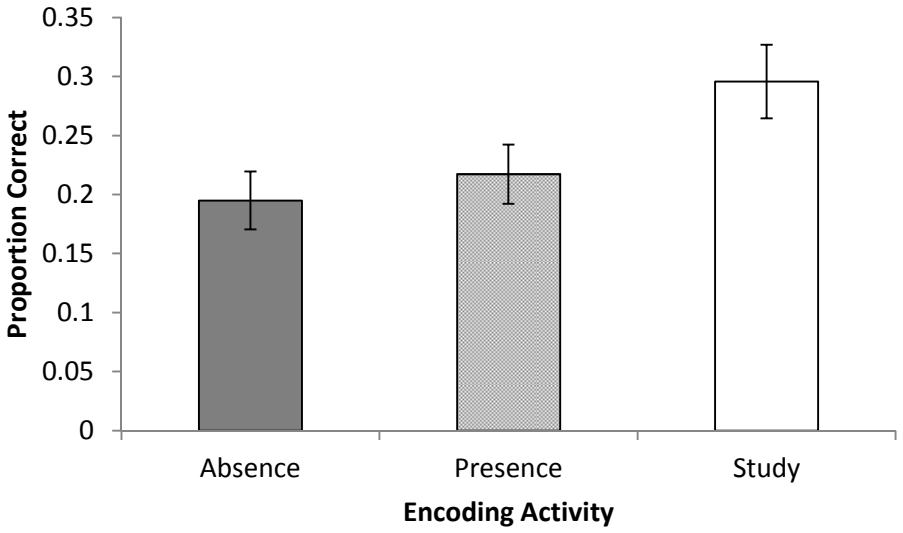


Figure 3.4 Final test performance for Experiment 4. Materials are unrelated English pairs.

## Experiment 5

In Experiment 5, we assessed the role of the particular trial structure of presence trials. For presence trials in Experiments 1-4, the alternative response was typed in while the correct pairing was on the screen. In this experiment, however, the learner first studied the pair intact and then on the next screen saw the cue again and generated an alternative response. If learners have to come up with a guess after the original target has gone off of the screen, not only is there time for uninterrupted encoding of the correct pairing, but we might imagine that when the trial progresses to the alternative response screen they would have to do some (minimal) retrieval of the target, benefiting memory, and perhaps, additionally, there is better differentiation of the correct targets and alternative responses. Absence (guess first) versus presence (guess second) was manipulated between participants. Each group of participants also completed study intact trials.

Participants were 76 UCLA students.

### Results and Discussion

In the absence condition, participants correctly anticipated the to-be-learned responses 5.40% of the time, and these responses were excluded from analysis.

Data were analyzed both for the complete set of participants, as well as excluding those participants who showed evidence of not understanding the instructions in the presence condition—namely, these people, instead of providing alternative responses, just retyped the target word more than half of the time. Unlike the within participants designs of Experiments 1, 2, and 4, in a between-participants design there is worry that excluding participants biases the results—so here, I present analyses with all participants, as well as with the subset of those in the presence condition who showed evidence of understanding the instructions.

First, I present the full sample, with n = 35 for the absence condition and n = 41 for the presence condition. These results are plotted in Figure 3.5. Data were analyzed with a 2 (generate vs. study) x 2(absence vs. presence) mixed model ANOVA. Overall, there was a main effect of encoding activity, $F(1,74) = 22.57$, $p < .001$. Moreover, there was a main effect of absence vs. presence, $F(1,74) = 4.75$, $p = .032$. These main effects were qualified by an interaction between encoding activity and absence versus presence, $F(1,74) = 6.27$, $p = .014$.

Follow-up tests revealed the nature of the interaction to be such that for study only items, there was no reliable difference between those in the absence condition ($M = .52$, $SD = .30$) and those in the presence condition ($M = .45$, $SD = .24$), $t(74) = 1.19$, $p = .24$. For the generate items, though, items in the absence condition ($M = .66$, $SD = .24$) were remembered better than items in the presence condition ($M = .49$, $SD = .23$), $t(74) = 3.10$, $p = .003$.

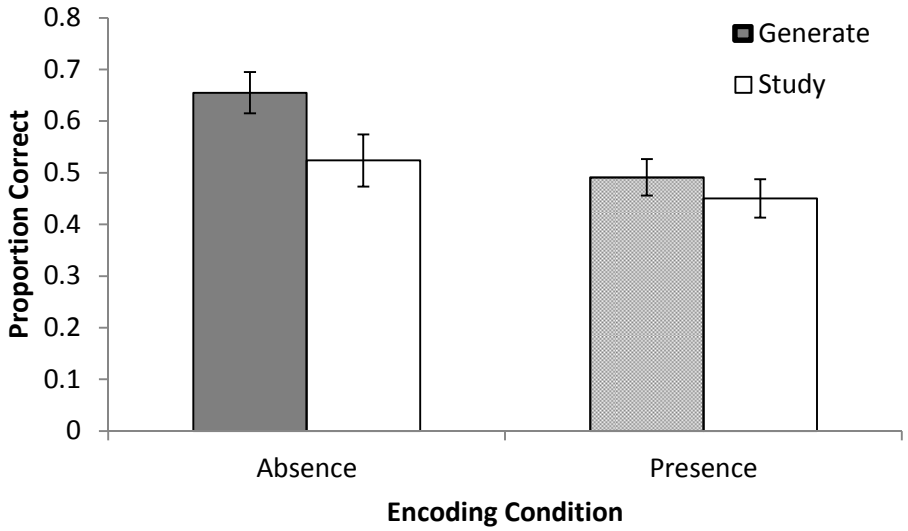


*Figure 3.5*. Final test performance in Experiment 5 for full sample of participants.

In the following analysis, participants were excluded from the presence condition if (instead of providing an alternative response) they retyped the target word more than half of the time. The sample sizes are now n = 35 (the same as before) for the absence condition and n = 29

for the presence condition. As with the full sample, we find a main effect of encoding activity, $F(1,62) = 16.07$, $p < .001$. The main effect of absence vs. presence is no longer significant, $F(1,62) = 2.28$, $p = .14$. The interaction, however, between encoding activity and absence versus presence remains significant, $F(1,62) = 5.48$, $p = .022$.

Follow-up tests reveal that as before, for study only items, there was no reliable difference between the absence condition ($M = .52$, $SD = .30$) and the presence condition ($M = .48$, $SD = .22$), $t(61) = .63$, $p = .532$. Levene's test indicated unequal variances ($F = 4.57$, $p = .037$) so degrees of freedom were adjusted from 62 to 61. For generate items, however, the absence condition ($M = .66$, $SD = .24$) was better than the presence condition ($M = .52$, $SD = .23$), $t(62) = 2.37$, $p = .021$.

Taken together, the conclusions from either analysis are largely consistent, with the only difference being the main effect of overall performance (considering both generate and study trials) in the absence condition as compared to the presence condition that goes away with only the subset of data. The rest of the story is more clear—with both analyses, the study only conditions did not differ from one another, but for the generate items, it was better to guess first than it was to provide an alternative response. So even though the presence trial structure differs from earlier experiments, the results do not.

## General Discussion

In five experiments, we found that having participants generate their guesses after, rather than before, encoding the to-be-learned pairing did not improve performance over the typical errorful generation task. Even when we removed the factor of guess/alternative response retrieval by giving participants both cues and guesses/alternative responses at the final cued-recall test, the typical guess first condition outperformed the presence condition. We expected to see a

substantial benefit of the presence condition over the absence condition while encoding unrelated pairings, as this is a case where the absence condition has failed to help encoding in the past, and perhaps after seeing the intact unrelated pairing, participants could generate a helpful link. Even in this case, performance did not differ reliably from the absence case. There was only a slight numerical advantage, which was the only one found in this series of experiments comparing presence to absence trials.

Our results cast some doubt on a mediator explanation of errorful generation benefits, or at least suggest that perhaps because of the dynamics of the task, alternative responses in the presence condition were less helpful than guesses were in the absence case. One such dynamic might have been due to the timing of encoding of the correct pairs. In the typical case, the absence condition, participants guess one associate before learning another. In the presence condition, participants generate alternative responses after encoding the to-be-learned pair, either with the correct pairing remaining on the screen or, in Experiment 5, after the correct pairing leaves the screen. As memory for the to-be-learned target was assessed at a reasonably short delay (only 5 minutes after the encoding phase ended), the effects of retroactive interference may have been relatively stronger than the effects of proactive interference (Bjork, 2001). For the absence case, more retroactive interference favors memory for the correct response, whereas in the presence case, it favors stronger memory for the (generated) alternative response. At a longer delay, the effects of proactive interference become relatively more apparent than retroactive interference, so perhaps at a longer delay, the presence case—where the correct response is now favored—would appear differentially better than the absence case.

Prior work, as well as the current investigation, suggests that guesses are remembered quite well, which argues against a suppression account of the benefits of errorful generation

(Knight et al., 2012; Vaughn & Rawson, 2012; Yan et al., 2014). Having good guess memory, however, can be explained in several ways. When you see the cue, the guess is arguably your pre-potent response. It is whatever your idiosyncratic strongest association is for that word. When you are asked at the final test what your guess was, you could either generate it again (as your pre-potent response), or generate it from episodic memory of the study phase. Better guess memory, and having more targets remembered when guesses are also available, could also indicate that the entire encoding episode was more memorable, not necessarily that at the time of the final test the participant is relying on memory for the guess to trigger memory for the target. When guesses and alternative responses were presented, as in Experiment 2, performance did not improve, even though if the correct pairs and generated responses were bound together as triples, those additional cues should have been helpful links to the correct target.

Huelser and Metcalfe (2012) differentiate between generated errors potentially acting as mediators versus acting as elaborators, with the latter perspective minimizing the role of the specific response, and instead placing emphasis on the activation in memory that went in to producing the response. With the presence manipulation from the current experiments, we can consider the role of generating alternative responses in the presence condition from a response as an elaborator perspective. So in addition to thinking of the presence condition as a way for learners to generate better mediators, we can think of this manipulation as investigating whether the network must be activated prior to the encoding of the target, as in the typical, absence condition, or whether activating the network after initial encoding of the target would be just as beneficial. Our results suggest, then, that activating the network first is more beneficial than activating it after encoding.

However, we do not have a pure measure of activating first versus activating after, as we also changed the nature of the activation—it may be more constrained in scope, as the participant already knows what the correct pairing is. It is difficult to say whether more focal activation is helpful or harmful, but we do have evidence that constraining guess retrieval to a pre-specified related response hurts learning of the correct response (Grimaldi & Karpicke, 2012). In some cases, alternatively, participants in the presence condition might not produce as much relevant activation, if they were trying to think of something totally different related to the cue word, instead of something that might connect the correct cue and target.

In Chapter 4, we consider the role of errors as elaborators in more depth, and assess recent challenges to a semantic activation hypothesis.

## Chapter 4: Assessing the Semantic-activation Interpretation of the Benefits of Errorful Generation

Why does guessing an answer, even if you are always wrong, sometimes enhance learning of the correct answer later? Many researchers have posited that the act of guessing activates a semantic network. When conditions are right, that is, when the to-be-learned feedback is presented soon after the guess attempt (before activation dissipates), and the feedback is semantically relevant to the activation, target encoding will be better than without a guess (e.g. Huelser & Metcalfe, 2012). A semantic activation hypothesis can plausibly explain why, among other things, benefits are found with related but not unrelated word pairs and with immediate but not delayed feedback (e.g., Hays et al., 2013).

Recent findings in the literature, however, have cast some doubt on the plausibility of a semantic activation hypothesis, at least as a complete explanation of the benefits of errorful generation. Namely, studies have found benefits for errorful generation with delayed feedback and unrelated word pairs (Kornell, 2014; Potts & Shanks, 2014). In the current experiments, we test predictions of a more general explanation of errorful generation benefits—beyond a semantic activation account.

### Benefits of Delayed Feedback

In experiments where the materials were trivia questions, Kornell (2014) found benefits of errorful generation even when feedback was delayed. Participants were asked questions like "What is the world's tallest grass?" Participants either studied the question together with its answer: "What is the world's tallest grass? Bamboo," or saw the question and typed in their answer (typically an incorrect guess, given the nature of the materials) before receiving the

correct answer. On trials where participants had to type in their guesses, feedback was either given immediately or after approximately six minutes. Performance in both immediate and delayed feedback conditions was better than performance in the study-only condition. These findings differ from those of studies using semantically related word pairs, where errorful generation benefits over reading are typically only found with immediate feedback. Errorful generation effects were also found when feedback was delayed by 24 hours after participants had generated guesses three times for each trivia question.

The results from Kornell (2014) extend prior findings by Richland et al. (2009). Participants completed short-answer pretests before reading a text passage about vision, or studied the text for a longer time. Here, no explicit immediate feedback was given to each question; instead, the participants read the answers to the questions within the text. Participants in the pretested group outscored those in the extended study group.

Kornell (2014) offered an explanation that is perhaps consistent with a semantic activation hypothesis, though not as previously construed. Using word pairs, other researchers had concluded that any benefits of semantic activation must be short lived, as no benefits of guessing were found if feedback was delayed. Kornell argued that learners know that when guessing or encoding paired associates, there is no one right answer. There is an answer for experimental purposes, and while it may be related to the cue, it is a somewhat arbitrary association. With trivia questions, in contrast, learners know that there is a correct answer, and that they could have known it. Meaningful, more interesting questions could thus lead to broader, lasting activation of semantic networks.

**Benefits with Unrelated Materials**

While benefits of delayed feedback for interesting materials could lead us to broaden our view of what semantic activation might entail, results showing benefits for unrelated pairs are perhaps most troubling for a pure semantic activation account (Potts and Shanks, 2014). As these results have important theoretical implications, I now describe the rationale and results of Potts and Shanks' investigation in detail.

Potts and Shanks (2014) point out that in almost all of the recent studies on errorful generation, learners are presented with a cue word that is familiar to them. When asked to guess what goes with the cue word, then, learners generate something that is associated—almost always semantically—to the cue. When given the cue *whale,* for instance, learners might generate *dolphin* or *squid* or *Shamu,* but probably not *chocolate*. In the case where the to-be-learned pair is related, as in *whale-mammal,* the generated guess will then be at least marginally related to the target, and so the guess may act as a helpful link. In the case where the to-be-learned pair is unrelated, however, the guess will not be related to the target, and may interfere with access to the target at test. In all cases, the act of guessing may indirectly benefit encoding of the correct response by increasing attention to feedback in some way. Any benefits, however, could be masked by interference from the generated guess, as the cue-guess association should be strengthened by retrieval.

The potential implications of Potts and Shanks' work are interesting: it could be the case that guessing first before studying a correct answer always benefits learning—through more direct benefits of retrieval, indirect benefits, or both. Sometimes, however, interference dynamics could hide the benefits. Our question of when errorful generation is beneficial changes: when are benefits displayed or hidden?

To eliminate the semantic relationship between the cue and guess that has been present in other studies, Potts and Shanks used study materials with unfamiliar cues. Cues were obscure English words and responses were one-word definitions, such as *hispid-bristly,* and in other experiments, foreign language (Euskara) words with English translations such as *igel-frog*. Because participants were very unlikely to hold any pre-existing associations with the cues, it is unlikely that the generated guess would be related to either the cue or the target. For the participant, even though *hispid* does have a correct definition, pairing it with that definition creates, for our purposes, an unrelated pairing.

In the generate condition, participants saw the obscure word and had to guess its definition before studying the word with its correct definition. A 'choice' condition was also investigated, where participants had to choose from two potential definitions, one of which was the correct one (we do not focus on this condition for the current studies). As a control, in a read condition some words were studied together with their definitions for the entire time. Time on task for all trial types was held constant.

In contrast to prior results that did not yield errorful generation benefits for unrelated pairs, guessing an incorrect definition produced better performance on a final test than did studying the word and correct definition the whole time (Potts & Shanks, 2014).

**Uncovering Hidden Benefits of Generation**

Finding benefits of errorful generation with unrelated materials (Potts & Shanks, 2014) and delayed feedback (Kornell, 2014) might cast doubt on explanations requiring relevant semantic activation. Perhaps there is always a benefit of generation, but the degree to which we see it depends on competing dynamics.

When Potts and Shanks found benefits for errorful generation with unfamiliar word and definition pairs, the final test had a multiple-choice format. For some of their experiments, in addition to the correct definition, the lures on the test included the lure that had been present in the choice condition (the participant had only seen that lure for the choice trials), and two other lures. In one experiment, the responses that participants generated for errorful generation trials were also included as lures.

Why could it be important that a multiple choice test was used to show generation benefits? Interference could be avoided at test in two ways. First, as there are no pre-existing associations to the cue, as Potts and Shanks argue, no other cue-guess associations are strengthened. Second, the presentation of the correct answers at final test could allow participants to bypass any interference that a guess could have potentially created.

Hays et al. (2013) used a backwards test in an effort to get around effects of interference. With weakly related word pairs and a typical (forward) cued-recall test, they found that delaying feedback got rid of errorful generation benefits over study. Furthermore, for generate trials with delayed feedback, participants were somewhat likely to provide their incorrect guesses at final test. By using a test where participants were instead provided with the target (*mammal*) and instructed to produce the cue word (*whale*), the goal was to avoid competition from other associations to the cue. On the backwards test, participants recalled more items from the delayed feedback condition than the study condition (Hays et al., 2013).

Finding benefits for errorful generation with delayed feedback on a backwards test suggests that manipulating test type might be a good strategy for uncovering hidden benefits of errorful generation that are masked by interference on a typical cued-recall test. It is also possible that some of the benefit of generation with unrelated pairs found in Potts and Shanks'

experiments was due to their use of a multiple choice test, and no benefit would be found on a cued-recall test.

From the prior literature, it remains unclear whether under some circumstances, there are no effects of errorful generation on subsequent encoding, or alternatively, if effects in those cases are masked by the competing dynamic of interference. In six experiments, we explored whether by changing interference dynamics at test we could uncover benefits in cases where limits on the effect have been found before, along the dimensions of semantic relationship, retrieval dependence, and constraining retrieval.

## Experiment 1

The aim of Experiment 1 was to replicate the errorful-generation benefit found by Potts and Shanks (2014) with unrelated pairs and unfamiliar cues on a multiple choice test, and determine whether benefits are still present when probed with a cued-recall test. Benefits of errorful generation have in other cases only been found with semantically related pairs.

### Method

**Participants.** Participants were 73 UCLA students.

**Design.** Encoding activity—reading or generating definitions of Swahili words—was manipulated within participants. Test type—multiple choice or cued recall—was manipulated between participants.

**Materials and procedure.** Sixty Swahili words with their English translations were taken from Nelson and Dunlosky's norms (1994). During the learning phase, half of the pairs were in the generate condition, where the Swahili word was presented and participants had 8 s to guess the definition. The Swahili word was then presented along with its English translation for 5 s. The other half of the pairs were assigned to the study condition, where the Swahili word was

presented along with its English translation for 13 s. Assignment of Swahili words with their translations to the generate or study condition was random for each participant. Generate and study trials were randomized throughout the learning phase.

After a five-minute Tetris distractor phase, the participants took either a multiple-choice or cued-recall test. On the multiple-choice test, each Swahili word was presented with its correct definition and three lures. The lures were correct translations for other Swahili words studied in the experiment. On the cued-recall test, the Swahili word was presented alone and participants were instructed to type in the correct English translation. Both test types were self-paced.

**Results and Discussion**

If benefits of errorful generation with unfamiliar cues (such as those shown by Potts and Shanks, 2014) rely on multiple-choice testing, we would expect an interaction such that the advantage for generating over reading would be bigger on the final multiple-choice test than on the cued-recall test, and we might even see a benefit of studying over generating on the cued-recall test. If benefits of errorful generation with unfamiliar cues emerge because of the general lack of interfering guesses associated with the cue, and thus do not depend on test type, we would expect benefits of generating over reading on both test types, with no interaction between test type and encoding activity.

Results (plotted in Figure 4.1) were analyzed with a 2(study activity) by 2 (final test type) ANOVA. Overall, performance was much lower on the cued-recall test than on the multiple choice test, $F(1,71) = 160.06$, $p < .01$. Overall, there was no significant difference between the generate condition and the study condition, $F(1,71) = 2.67$, $p = .11$. The interaction did not reach significance, $F(1,71) = 1.33$, $p = .25$.

Follow-up t-tests were used to assess whether we replicated Potts and Shanks' (2014) findings. On the multiple-choice test, there was a marginal benefit for generating ($M$ = .63, $SD$ = .19) over studying ($M$ = .59, $SD$ = .21), $t(38) = 1.87$, $p = .07$. This result is in line with Potts and Shanks' findings. On the cued-recall test, however, there was no difference between generating ($M$ = .15, $SD$ = .13) and studying ($M$ = .15, $SD$ = .11), $t(33) = .38$, $p = .71$. Performance on the cued-recall test was very low, however, which might obscure any meaningful differences.

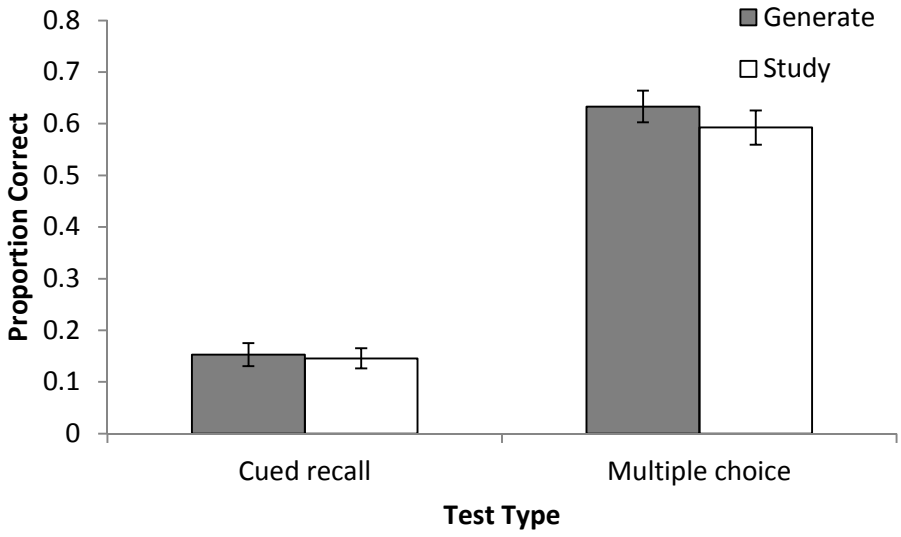


*Figure 4.1*. Final test performance from Experiment 1. Materials are Swahili-English pairs.

## Experiment 2

The aim of Experiment 2 was to determine if errorful generation benefits could be found with unrelated pairs that have familiar cues (*dog-spoon*), if potential interference was avoided by using a backwards or multiple-choice test.

### Method

Participants were 121 UCLA students.

The major difference between Experiment 1 and Experiment 2 is that the materials in Experiment 1 were Swahili-English pairs, and in Experiment 2 the materials were 60 unrelated English pairs. As in Experiment 1, generating versus studying was manipulated within participants, and test type was manipulated between participants. Here, in addition to cued-recall and multiple-choice tests, we included a backwards cued-recall test, where the target was presented and participants were instructed to retrieve the corresponding cue word.

**Results and Discussion**

If generation benefits do exist with unrelated pairs that have familiar cues (unrelated English pairs), and those benefits are being masked by interference, we would expect an interaction such that on a forward test, we would see a replication of typical (for these materials) null effects of study activity, but on a backwards test or multiple-choice test, there would be a benefit of generating over studying intact.

If, however, errorful generation does not potentiate encoding of feedback with unrelated pairs that have familiar cues, we would expect null effects of study activity, (or even benefits of studying intact), on multiple-choice, forward, and backward tests.

As shown in Figure 4.2, results revealed that overall, there was a substantial effect of test type, in that scores were much higher on the multiple-choice test than they were on the forward or backward cued-recall tests, $F(2,118) = 165.22$, $p < .01$. Overall, there was no main effect of generating versus studying, $F(1,118) = 2.26$, $p = .14$. There was however, a suggestion of an interaction of test type and encoding activity, $F(2, 118) = 2.30$, $p = .10$.

For the forward cued-recall test, there was a trend for generating ($M = .19$, $SD = .16$) to be worse than studying intact ($M = .23$, $SD = .19$), $t(35) = 1.67$, $p = .10$. For the backward cued-recall test, there was also a trend for generating ($M = .15$, $SD = .12$) to be worse than studying

intact ($M$ = .18 $SD$ = .15), $t(42) = 1.70$, $p = .097$. For the multiple-choice test, there was no significant difference between generating ($M$ = .73, $SD$ = .17) and studying ($M$ = .71, $SD$ = .19), $t(41) = .89$, $p = .38$.

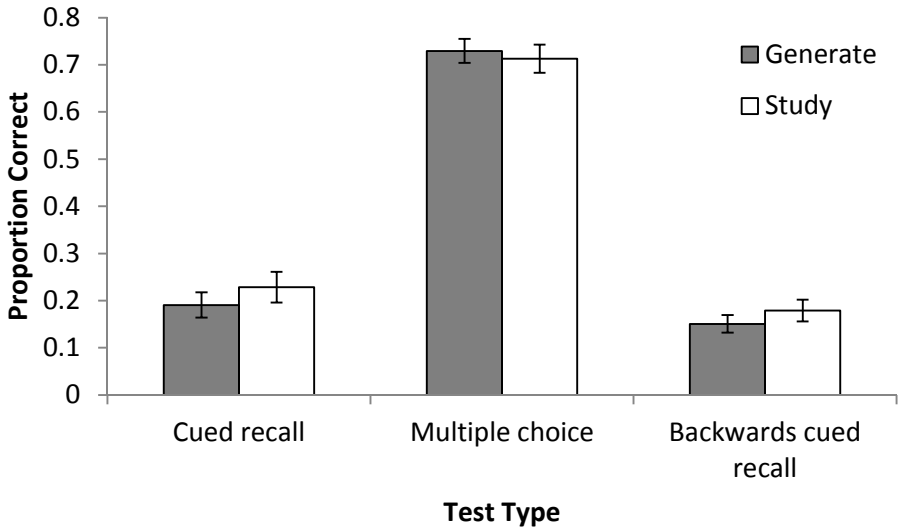


*Figure 4.2*. Final test performance for Experiment 2. Materials are unrelated English pairs.

## Experiment 3

As a point of comparison, a small sample of participants was collected using the same procedure as Experiment 2, but with related English pairs. The participants were 48 Mechanical Turk workers.

Results are shown in Figure 4.3. Due to much higher performance on the multiple-choice test than either of the cued-recall tests, there was a main effect of test type, $F(1,45) = 34.22$, $p < .01$. Overall, performance in the generate conditions was higher than in the study conditions, $F(1,45) = 12.64$, $p = .001$. Additionally, there was a trend toward an interaction between test type and encoding activity, $F(2,45) = 2.48$, $p = .095$.

Follow-up tests revealed that in the multiple-choice condition, performance was at ceiling, and generating ($M$ = .99, $SD$ = .022) did not produce significantly better performance than studying ($M$ = .98, $SD$ = .053), $t(14) = .73$, $p = .48$. On the forward cued-recall test, results replicated the typical finding that generating ($M$ = .56, $SD$ = .25) was better than studying ($M$ = .45, $SD$ = .27), $t(16) = 2.34$, $p = .033$. Performance on the backwards cued-recall test was almost identical to performance on the forward test, and generating ($M$ = .55, $SD$ = .21) was significantly better than studying intact ($M$ = .44, $SD$ = .23), $t(15) = 2.80$, $p = .013$.

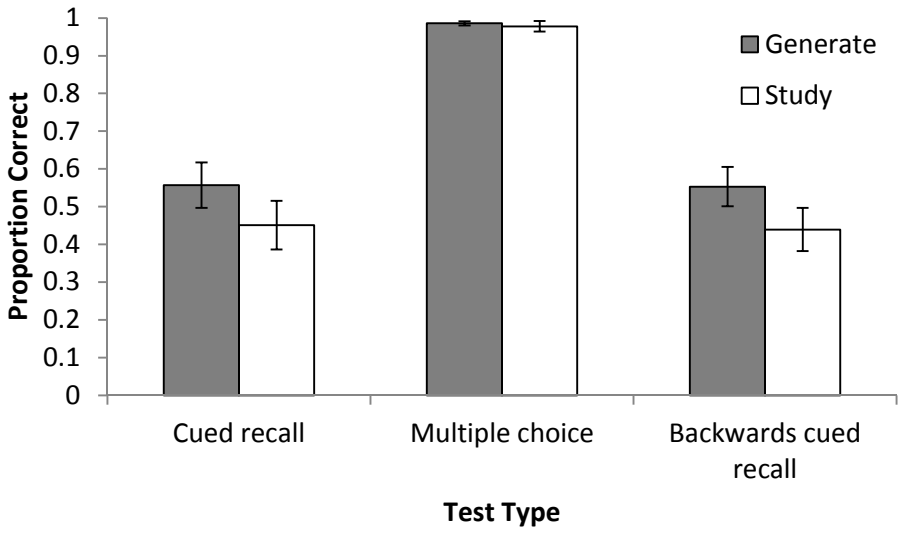


*Figure 4.3*. Performance by final test type on Experiment 3. Materials are related English word pairs.

## Experiment 4

In Experiments 4 and 5, we used free recall tests to determine whether elements of the encoding phase are strengthened by errorful generation in a way that is not amenable to the associative tests used thus far in Experiments 1-3. In Experiment 4, the to-be-learned materials were unrelated English pairs. The encoding phase was the same as in Experiments 1-3. For the

test phase, participants were told that they should recall all words they remembered studying in the experiment, not their original guesses. The test was self-paced. Participants were 50 students from UCLA.

**Results and Discussion**

When scoring the data it became apparent that participants occasionally offered guesses that they had seen as to-be-learned words earlier in the experiment, or—very rarely—the other way around, where a participant would give a guess in the encoding phase that would turn out to be a cue or a target for a later pair. In these cases, if the item was recalled on the free-recall test, it was classified however it had first appeared during the study phase. So if a study intact cue was later offered as a guess for a generate trial, it was classified as a study intact cue.

Performance (plotted in Figure 4.4) on the free recall test was very low. When interpreting means, it should be noted that there were 60 word pairs in the study phase, for a total of 120 possible words to recall.

Overall, memory for words from generated pairs was better than memory for words from studied pairs, $F(1,49) = 14.74$, $p < .001$. There was no main effect of cue versus target memory, $F(1,49) = 1.77$, $p = .19$. There was, however, an interaction between encoding type and element of the pairing $F(1,49) = 13.40$, $p = .001$. For cue words, generate words ($M = 4.38$, $SD = 3.00$) were remembered better than study words ($M = 2.18$, $SD = 1.78$), $t(49) = 4.93$, $p < .01$. For target words, however, the difference was not as substantial between generate words ($M = 3.32$, $SD = 2.3$) and study words ($M = 2.74$, $SD = 1.83$), $t(49) = 1.45$, $p = .15$.

Again, performance was very low, but this was our first demonstration of a generation benefit with unrelated pairs. The generation benefit was shown primarily for recall of cue words.

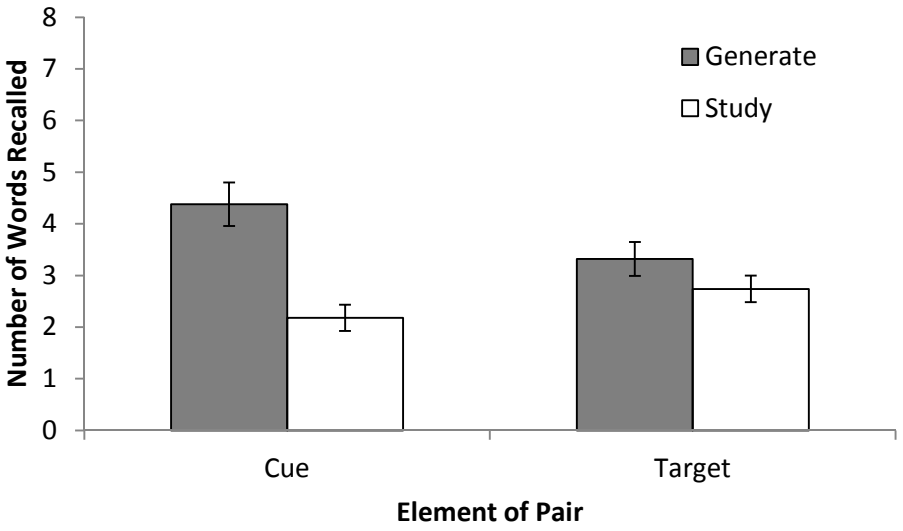


*Figure 4.4*. Free recall performance from Experiment 4. Materials are unrelated English pairs.

## Experiment 5

Experiment 5 was identical to Experiment 4, except the to-be-learned materials were weakly related English pairs. Participants were 53 students from UCLA.

### Results and Discussion

Overall, words from generated pairs were remembered better than words from studied pairs, $F(1,52) = 37.42$, $p < .01$. Cues, moreover, were remembered marginally better than were targets, $F(1,52) = 3.54$, $p = .066$. There was also a trend toward an interaction between encoding activity and element of the pair, $F(1, 52) = 3.21$, $p = .079$, which took the form that for cue words, there was a significant benefit of generation ($M = 6.53$, $SD = 3.85$) over studying ($M = 3.11$, $SD = 2.38$), $t(52) = 6.50$, $p < .01$, whereas for target words, there was also a benefit of generation ($M = 6.15$, $SD = 3.74$) over studying ($M = 3.13$, $SD = 2.45$), but the difference was slightly smaller, $t(52) = 5.50$, $p < .01$. Compared to Experiment 4, where a generation benefit

was apparent in the cue words but not target words, with related pairs, the benefit is apparent for both cues and targets.

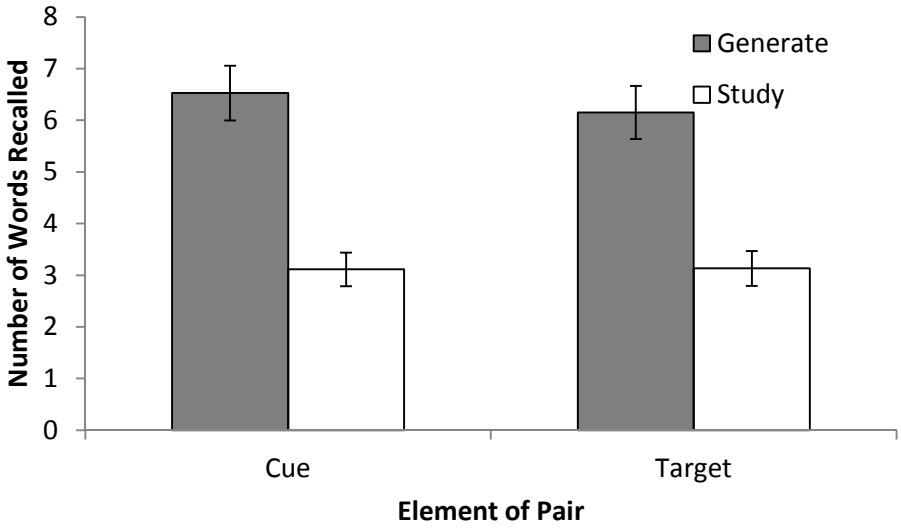


***Figure 4.5*.** Cued recall performance from Experiment 5. Materials are weakly related English pairs.

## Experiment 6

Grimaldi and Karpicke (2012), found that retrieval, and more specifically unconstrained retrieval of guesses, seemed to be required to see benefits of errorful learning. When learners studied the cue with the normatively given guess, or generated the normatively given guess from a word stem before studying the correct cue-target pairing, performance was worse than when learners only studied the correct pairing or were allowed to generate their own guesses.

In Experiment 6, we assessed whether studying or generating lures might have some benefit over only studying the correct pairing when probed by a backwards test.

## Method

In this experiment, study activity (generating lures vs. studying lures) was manipulated between participants, and each of those groups had their own study intact control conditions (so processing a lure vs. studying was manipulated within participants). Test type (a forward cued-recall or backwards cued-by-target recall test) was manipulated between participants. Participants were 293 UCLA students.

To-be-learned pairs were 60 weak semantic associates. These pairs were identical to the pairs used in Experiment 3 with one exception, as one pair did not have a higher associate than the one already used as a target. For the study and generate lures conditions, lures were constructed by taking the highest forward associate to the cue, unless that associate was present elsewhere in the experiment. In that case, the second highest associate was used. For the set of lures, the average forward association strength was .32.

In the study lure condition, participants saw a cue word paired with the lure for 8 s ("*golf-ball"*), before studying the to-be-learned pairing for 5 s ("*golf-sport"*). In the generate lure (constrained retrieval) condition, the cue word was presented for 8 s paired with the lure with its vowels removed ("*golf- b_ll"*). Participants were instructed to complete the lure. The to-be-learned pair was then presented intact for 5 s.

## Results and Discussion

If unconstrained retrieval is required for benefits of learning from feedback, then we would expect a main effect of study activity, such that both studying lures and generating lures is worse than reading intact pairs on both forwards and backwards tests. If, however, there are benefits of constrained retrieval and studying lures, but they are masked by interference, we

might expect an interaction such that studying lures and generating lures would be better than reading on a backwards test, but worse than reading on a forward test.

Results, presented in Figure 4.6 for the study lure conditions and Figure 4.7 for the generate lure conditions, were analyzed with two separate 2(lure processing vs. study intact) by 2(forward vs. backward cued-recall) ANOVAs.

**Study lure**. Overall, performance was higher on the forward test than on the backward test, $F(1,142) = 10.95$, $p = .001$. Additionally, there was a main effect of encoding activity such that the study lure condition outperformed the study intact condition, $F(1,142) = 14.03$, $p < .001$. There was no interaction between encoding activity and test direction, $F(1,142) = .018$, $p = .89$. The benefit of studying lures over studying intact was not expected, and we did not replicate effects demonstrated by Grimaldi and Karpicke, (2012) where studying lures numerically hurt learning on a forward cued-recall test.

**Generate lure**. As in the study lure case, performance on the forward cued-recall test was better than performance on the backward test, $F(1,147) = 9.78$, $p = .002$. There was a trend for performance in the generate lure condition to be better than in the study intact condition, $F(1,147) = 2.72$, $p = .10$. There was also a trend toward an interaction between encoding activity and test direction, $F(1,147) = 2.98$, $p = .087$. On the forward cued-recall test, there was no difference between the generate lure ($M = .50$, $SD = .23$) and study intact conditions ($M = .50$, $SD = .23$), $t(76) = .053$, $p = .96$. On the backward test, however, the generate lure condition ($M = .42$, $SD = .19$) outperformed the study intact condition ($M = .37$, $SD = .20$), $t(71) = 2.30$, $p = .025$.

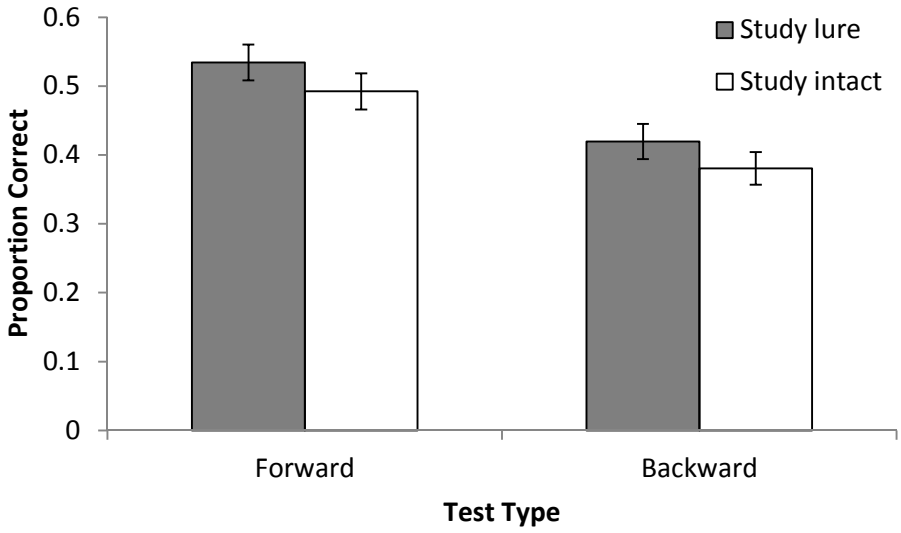


*Figure 4.6*. Final test performance for study lure vs. study intact conditions from Experiment 6.

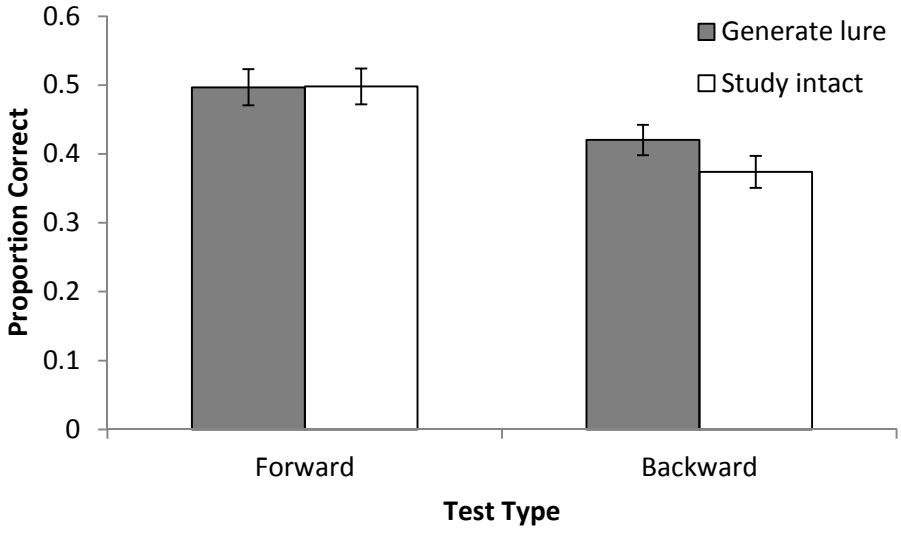


*Figure 4.7*. Final test performance for generate lure vs. study intact conditions from Experiment 6.

## General Discussion

The aim of the current investigation was to determine whether constraints of the errorful generation effect—semantic relatedness, delayed feedback, retrieval dependence, and constrained retrieval—are true constraints, or whether different test types could uncover hidden benefits. Our logic was that if hidden benefits were uncovered, benefits of errorful generation for encoding subsequent feedback might be quite general, based on general indirect benefits of testing, not only semantic ones, but those benefits can sometimes be hidden by the interference dynamics present at test.

Our results from the current experiments are mixed. In Experiment 1, our findings with Swahili-English pairs suggest that the benefits of errorful generation with unrelated pairs and unfamiliar cues, akin to those obtained by Potts and Shanks (2014), are replicable effects, even with different materials and test constructions. With a cued-recall test, however, we did not see any benefits of generation. Performance was very low, however, which could make it more difficult to see any benefits. Notably, in a similar experiment with fewer to-be-learned pairs and extended study time, a small generation benefit was obtained (Potts, 2014).

In our experiments with unrelated English pairs (2 and 4), we sought to determine whether we could uncover any hidden benefits of generation—using backward, multiple-choice, and free-recall tests. Perhaps any benefits of generation in prior studies were hidden by interference from guesses associated with the familiar cue. The results of our cued-recall test in the forward direction replicated other findings in showing no generation benefit and a trend toward worse performance in the generate condition than in the study condition. The backward cued recall test, which we thought might circumvent interference, did not uncover any benefits. In fact, performance on the backwards test was very similar to the forwards test and also showed

a trend for generating to be worse than studying. On the multiple choice test, performance was much better overall, and there was no difference in performance between generating and reading (with the slight numerical advantage going to the generate condition).

It was only on the free recall test (Experiment 4) where we saw a generation benefit for unrelated English pairs, and the benefit was larger for cues than for targets. This finding could suggest that what is strengthened in this case is the cue alone, but not any associative links between the cue and the target (otherwise, we should have seen generation benefits on the backward cued-recall test). However, performance on the final free-recall test was exceedingly low, which suggests that this finding should be re-examined using a shorter list of pairs.

In Experiment 6, we attempted to assess the role of backwards tests on potentially uncovering generation benefits after studying lures or generating lures from cues. Unlike in Grimaldi and Karpicke's (2012) investigation, where studying lures impaired participants' ability to encode the correct answer, we found that performance was better for the study-the-lure condition on a forward cued recall test. So in our case, we had no benefit to uncover on the backwards test (though it did remain on the backwards test). And for the generate-a-lure case, we found no effect on a forward test and we found a benefit of generating lures over studying intact on the backward test. One potential reason for different patterns of results between the current work and Grimaldi and Karpicke's studies is that their lures were constructed by looking at prior participants' guesses and choosing the most common one for each cue. On average, these words had been produced 22% of the time in earlier studies. In Experiment 6, lures were constructed by taking the highest associate available from norms, with a forward association strength of .32. Another difference between our studies is that relatedness (related or unrelated

English pairs) was manipulated within participants in Grimaldi and Karpicke's study, whereas only related English pairs were used in Experiment 6.

On the whole, our attempts to uncover benefits of errorful generation for unrelated pairs were unsuccessful, and our results were mixed when we attempted to replicate the boundary effects of generating and studying lures. To explain the former, one possibility is that our tests did not, for whatever reason, circumvent the interference generated by having a familiar cue.

Another possibility is that the tests that we chose did not tap in to aspects of memory that are strengthened by errorful generation for unrelated pairs. With related pairs, we know that cue-to-target memory, target-to-cue memory (Experiment 3) and the individual cue and target elements (Experiment 5, though interpreted with caution) are strengthened. Furthermore, for unrelated pairs with unfamiliar cues (Swahili-English pairs), at least the forward association is strengthened (Experiment 1; Potts and Shanks, 2014). With unrelated English pairs, perhaps neither the cue to target nor the target to cue associations are strengthened, but individual elements are strengthened, as suggested by Experiment 4. After generating an error to a familiar English word, perhaps there is more attention to feedback, but that attention is focused on the elements separately, instead of on their association.

Another reason we may not see the same benefits of errorful generation with unrelated English pairs as compared to obscure English or foreign language vocabulary-definition pairs, is that we might expect the indirect effects on feedback to be different in the two cases. Students, for instance, might be more motivated to learn vocabulary, as it is a familiar task to them and something that seems worthwhile. After an initial retrieval error, it might be natural to pay more attention to feedback, and to try to link the foreign word with its translation. A follow-up study

could examine study time choices after errorful generation in response to a foreign language term and compare them to study time after errorful generation in response to a familiar word.

# Chapter 5: Overview and Conclusions

Generating errors on tests designed for assessment can lead to bad grades and frustrated students. Generating errors on tests designed for formative assessment, however, can potentiate subsequent learning. The goal of the current work was to better understand the boundary conditions of such effects, and to clarify the mechanisms underlying errorful generation.

## Overview of Findings

In Chapter 2, we tested how changing the type of guess could affect learning. Generating guesses related to the cue word resulted in improved learning of semantically related targets, but not of rhyming targets. When generating words that rhyme with the cue, there was no improved learning for related or rhyming targets. These results suggest that not just any type of generation from memory is enough to improve learning, even if that type of generation is consistent with the type of to-be-learned pairing (as in the case of rhyming guesses and rhyming pairs).

In Chapter 3, we tested a mediation-based explanation of errorful generation by creating circumstances where generating good mediators was either more or less likely. We first presented the entire to-be-learned pairing, and then instructed participants to generate an alternative response (in effect, guessing after seeing the pair), instead of guessing before, as is typical in errorful generation and pretesting scenarios. A mediation account of errorful generation, on our view, would predict better learning when generating guesses in the presence of both cue and target because participants would be able to make better mediators than when guessing in presence of only the cue word. Counter to our expectations, guessing in the presence of the correct answer did not show benefits over guessing in the absence of the correct answer.

Finally, in Chapter 4, we assessed the adequacy of a semantic-activation explanation of errorful generation benefits. While many past studies are in line with such an explanation, recent

findings suggest that more general mechanisms may be responsible for errorful generation benefits. We tested an implication of these more general benefits of errorful generation by attempting to uncover benefits that might have been created, but otherwise hidden by interference. To do so, we manipulated the types of materials and final tests used in the paradigm. The results from these attempts were mixed, but largely unsuccessful, suggesting that depending on the materials, elements and associations between the words in a word pair might be differentially strengthened.

**Seeking a Cohesive Explanation for Errorful Generation Effects**

In searching for explanations of errorful generation effects, one puzzle to be solved is why there are somewhat different boundary conditions for errorful versus successful generation and testing effects. For instance, in successful generation tasks, many generation rules bring about memory benefits, such as coming up with synonyms, rhymes, and computations. But as explored in Chapter 2, making guesses following a rhyme generation rule for errorful generation does not enhance subsequent learning.

Why might there be different boundary conditions on successful and errorful generation? To understand the complexities of the current studies, and how they differ from typical testing or generation, it is useful to think of the elements and associations involved in errorful generation tasks. We can consider the nature of the generation processes, cues, to-be learned targets, and their associative strength. We also must consider the cue-to-guess and guess-to-target relationships, as well as the identity of the guess itself as an element of the context of encoding.

In cases where generation succeeds, the type of generation is largely relevant to the cue and target (though there may be differences in depth), and the generated response (the guess, so to speak) is the same as the target. When the guess is different from the target, so there is a

potentially competitive relationship between the guess and the target, matters become more complex. The type of generation may or may not be relevant to the target, the nature of the guess will be determined by the type of generation, and the guess will be different from the to-be-learned target.

The current studies, using established errorful generation paradigms, make it difficult to study these task dynamics independently. As discussed in Chapter 2, manipulating the nature of the generation also manipulates the products of generation, the guesses. As a necessary side effect, we also manipulated the cue-to-guess and guess-to-target relationships. In Chapter 3, where we primarily sought to make the guess a better link between the cue and target, we also, in all likelihood, manipulated the nature of the generation itself, perhaps limiting its scope. In Chapter 4, where we manipulated the type of materials, we also manipulated the potential relationship between the guess and the to-be-learned target. And by manipulating the test types (forward and backward cued-recall, multiple choice, and free recall), we changed the components of memory that we were evaluating. For each manipulation, then, we have to consider competing dynamics.

## When and Why does Learning Profit from the Introduction of Errors?

We now seek to offer an answer to the question posed by the title of this dissertation. From the evidence we have so far, we argue that benefits of errorful generation are found in cases where either direct or indirect effects of generation (or both) are present, and there is an absence of interference. In the following section, we discuss direct and indirect effects of generation and possible sources of interference in more detail.

**Direct and indirect benefits of generation**. The distinction between direct and indirect benefits of retrieval has been drawn by Roediger and Karpicke (2006a) and Arnold and

McDermott (2013), among others. Direct benefits are produced by process of retrieval itself, while indirect benefits are mediated by the act of testing. For instance, in an experiment where participants studied and then either took free-recall tests without feedback or restudied the passage, we can measure in a relatively pure way the direct benefits on a later test of retrieving over restudying (e.g., Roediger & Karpicke, 2006b). But in cases where learners are given feedback after their retrieval attempts, there are also indirect benefits of retrieval, such as better encoding of feedback. With errorful generation, therefore, it is necessary to consider both direct effects and indirect effects, as feedback about the correct answer must be given after an incorrect guess is generated.

What is the nature of direct benefits of retrieval? For successful retrieval, Soloway (1986) demonstrated that contact with semantic memory is important for generation benefits. On generate trials, participants generated targets from a cue word and letter stem. On *trivially generate* trials, the correct response was presented 3 s before participants were asked to generate it. On a final free-recall test, the generate group performed best, followed by a copy control group and then the trivial generate group. Soloway concluded that these results support a semantic-activation theory of generation effects, given that generating from a recent, shallowly encoded episode was not as effective as generating from semantic memory.

Elroy and Slamecka (1982), after failing to find a generation effect with nonwords, argued that their results were evidence that some type of contact with semantic memory is necessary to show benefits of generation. In contrast, Johns and Swanson (1988) found generation effects with nonwords, presumed not to have semantic content, though those effects were somewhat attenuated from those obtained with word stimuli, and were only found when

feedback was provided. In Bertsch and colleagues' (2007) meta-analysis, the generation effect with nonwords (Cohen's $d$ = .05) was substantially smaller than with words (Cohen's $d$ = .41).

While some contact with semantic memory might be needed to produce generation effects, the depth of such activation might not matter that much, as discussed in Chapter 2. With successful generation effects, a broad spectrum of generation rules, including generating synonyms, filling in letters, and rhyming, produces benefits. For testing effects, however, depth, or effort, involved in retrieval tends to be important (e.g. Carpenter & DeLosh, 2006)

For errorful generation, benefits have largely only been found when learners generate information that is semantically related to the cue word. So perhaps benefits of errorful generation are due, in part, to direct benefits of semantic processing. Generations based on word stems or on the basis of rhyming are unlikely to produce as much semantic activation, and do not show the beneficial effects of generation. Cyr and Anderson (2015) argue that even the errorful generation benefits Potts and Shanks (2014) found with foreign words might be, in part, explained by learners doing conceptual processing of the foreign words, even if the words themselves are unfamiliar. Learners might, for instance, consider the meaning of words that are imbedded in the unfamiliar word.

Any possible direct effects have to be considered within the context of the to-be-learned materials. Even with semantic generation, generation benefits are not found when the subsequent feedback is unrelated to the cue (and both cue and target are English words). The aim of Chapter 4 was to test a claim made by Potts and Shanks (2014) that perhaps independent of whether there is a direct benefit of generation, generating guesses always potentiates subsequent learning through indirect means, and sometimes that benefit is hidden by interference. Potts and Shanks speculate that guessing trials feel more difficult for the learner

(as demonstrated by lower JOLs for generated than studied pairs), and that experience of difficulty may lead to increased attention to subsequent feedback. Our attempts to uncover such benefits produced mixed results, and more study is needed before this issue is resolved.

Broadly speaking, claims by Kornell and Vaughn (in press) that there is no reason at this point to assume entirely different mechanisms underlying errorful generation and those underlying successful generation, seem reasonable. Even Slamecka and Fevrieski (1983) suggested that errorful generation effects, in some cases, at least, represent successful but partial recovery of elements—the semantic elements, in their case, but not the surface ones.

**Exploring potential sources of interference**. While there may be no reason to posit separate mechanisms of retrieval on successful as compared to errorful generation, we should consider how the products of an errorful retrieval might interact with the to-be-learned cue or target information, as any direct or indirect benefits of generation might be overshadowed by interference.

Bridger and Mecklinger (2014) made a distinction between high-constraint cues and low constraint cues. High-constraint cues were word stems that had two strong associates (for instance, *bro-* is highly associated with *brother* and *broom*), whereas the low-constraint stimuli were weak semantic associates. Low constraint stimuli, as in other work with similar materials, showed an errorful generation benefit. High constraint stimuli, however, did not. One reason for this lack of a benefit could have been from the absence of a direct effect—there was not sufficient semantic activation with the word-stem materials. Perhaps generating in response to *bro-* activates lexical, but not semantic properties of *brother* and *broom*. Additionally, at the time of test, it might be much more difficult to exclude the incorrect error, because both responses are highly active. In the case of semantically related words (here, the low-constraint condition), even

though the guess might also be active, it might be active in the context of a semantic network, making the guess and the target easier to distinguish from one another.

Cyr and Anderson (2015) showed similar effects with stimuli cued by a category ("The category is a pastry…"), which served as the conceptual condition, or a word stem ("ta____"), which served as the lexical condition. In both conditions, participants gave two guesses before receiving feedback (*tart*). Conceptual, but not lexical, guesses enhanced learning. Cyr and Anderson argued that it is harder to integrate lexical errors into a memory trace than conceptual errors, making them more likely to cause interference later. This explanation is in line with Bridger and Mecklinger's framing of contextualized errors.

Thinking of lexical errors as harmful is consistent with findings that reading or retyping the incorrect spelling of commonly misspelled words decreased accuracy on a later test (Jacoby & Hollingshead, 1990). Commonly misspelled words may also be a type of high constraint stimuli, as there may be two strong candidates for a given letter choice (e.g., does *necessary* have one *c* or two). Future work should explore errorful generation with highly confusable terms in vocabulary and concept learning.

Finally, as Potts and Shanks (2014) argue, interference could potentially take the form of other associations to the cue. When the cue is familiar—in the current studies that means it is in English—other pre-existing associations could get in the way of later remembering an unrelated association. It is unclear, though, how much this type of interference contributes to learning from errorful generation in realistic educational scenarios as compared to cases with arbitrarily defined, unrelated word pairs.

## Scaling up to Educationally Relevant Scenarios

Considering direct and indirect effects of errorful generation also informs how the effect may (or may not) scale up to complex, realistic educational scenarios. With pretesting in the classroom, as compared to in the lab, for instance, we might imagine many more possibilities for indirect effects. Pretest questions, for example, can tell the student what level of detail to expect on a later test, or what topics the professor thinks are important. In the following sections, we explore how dynamics of errorful generation may change with more complex learning scenarios.

**Confidence in errors.** In the current studies, participants were probably not very curious about what the correct answers were, nor were they responding with deeply held beliefs about, for instance, what word should be paired with *whale*. With more complex learning materials, however, curiosity and confidence in one's responses might matter quite a bit. Work on the hypercorrection effect, for instance, has demonstrated that high-confidence errors are more likely to be corrected for a later test than are low confidence errors (e.g., Butterfield & Metcalfe, 2006). Fazio and Marsh, (2009), argue that after committing a high confidence error, participants pay more attention to feedback. When feedback appeared in red or green font, for instance, participants were more likely to remember its color when they committed a high confidence error than a low confidence error.

**Retrieving from episodic versus semantic memory.** One limitation of the current work is that we have largely conflated dynamics involving generation and testing effects. There may, however, be important differences between the two effects, taking generation as retrieval from semantic memory, not attached to a particular episode, and testing as retrieval from an episode. Karpicke and Zaromb (2010) directly compared generation and testing effects. First, participants were exposed to a list of words. In a second phase, they saw cue words with letter stems and

were instructed to complete the targets with either the first word that came to mind (generation from semantic memory) or from the first phase (testing). Even though performance was equated in the second phase, the testing condition produced better memory than the generate condition on a final test.

On pretests in the classroom, or practice tests given early in learning, students likely generate some guesses from their prior (decontextualized) knowledge and some from a particular episode or lesson. Arnold and McDermott (2013) point out differences between some test-potentiated learning paradigms and errorful generation paradigms, and note that the same boundary conditions do not necessarily apply (e.g., delaying feedback).

**Productive failure.** Principles of errorful generation, loosely defined, have been applied in the classroom in work on *productive failure*. Kapur and Bielaczyc (2012) compared groups of students taught with productive failure (PF) techniques to groups taught with direct instruction (DI) techniques on performance on complex mathematical word problems. In the PF groups, students first worked in groups to try to solve the problems. Despite many attempts to represent and solve the problems, none of the students' efforts were successful. In a second phase, the PF groups were asked by their teachers to explain their reasoning and compare their different strategies. Throughout the discussion, the teachers led the students to the correct ways to solve the problems. In contrast, students in the DI groups learned how to solve the problems with the teacher always present and giving frequent scaffolding and feedback. At later test, PF students performed better than the DI students on the complex problems, suggesting that the opportunity to activate and explore prior knowledge and to generate multiple possible solutions for problem solving was beneficial for learning.

Additionally, the PF students outscored DI students on well-defined problems, even though the DI students had done more practice on well-defined problems. Kapur and Bielaczyc (2012) argued that the deeper understanding generated in the PF condition transferred to well-defined problems. Follow up work found that productive failure was also better than "vicarious failure" when students studied and evaluated the (modified) work of other students who had been in a productive failure lesson (Kapur, 2014). Understanding the differences between productive and vicarious failure, or comparing more generally how errors that are self-generated affect learning differently from those that are read or overheard, is an important direction for future work

The results from work on productive failure are interesting, though it is likely that in addition to the errors that students are retrieving throughout the productive failure exercises, there are many opportunities for successful retrieval of concepts that are important, but do not happen to apply to the current problem. Perhaps this work should be viewed as an overall success story of implementing desirable difficulties more generally, not just the principles of errorful generation.

A productive frame for future work on errorful generation—for both theoretical and applied reasons—could be to assess the different functions of pretesting. Roediger, Putnam and Smith (2011) delineate ten benefits of testing effects. For instance, testing produces better organization of information and improves transfer of information to new contexts. Some of the benefits, such as identifying gaps in knowledge and potentiated subsequent study, recognize the utility of generating errors during learning.

## Concluding Comments

Decades of research on interference dynamics suggests that generating an incorrect answer should impair subsequent encoding. Errorful generation studies, however, show that generating meaningful errors can help set the stage for effective encoding. The scope of when generating guesses will help versus hurt learning is still under investigation. Some known boundary conditions, such as the failure to find benefits of generation with unrelated English words, should not necessarily be feared by students, as most educationally relevant materials are more meaningfully related. Investigations of other—currently unknown—dimensions of the effect will be of more interest to students, such as whether it appears when a given set of materials is highly confusable, or whether it is more beneficial to novices or learners with some experience in a given domain.

# Appendix A

Aggregate metacognitive judgments of preferred learning conditions from Chapter 2.

| | Encoding activity | | | Pair type | | | Type of generation | | |
|---|---|---|---|---|---|---|---|---|---|
| | Guess first | Intact | Equal | Rhyming pairs | Related pairs | Equal | Rhyming | Related | Equal |
| Exp 1a: Rhyming | 0.28 | 0.53 | 0.19 | 0.47 | 0.22 | 0.31 | - | - | - |
| Exp 1a: Related | 0.21 | 0.45 | 0.34 | 0.24 | 0.63 | 0.13 | - | - | - |
| Exp 1b: Rhyming | 0.48 | 0.50 | 0.03 | 0.31 | 0.51 | 0.18 | - | - | - |
| Exp 1b: Related | 0.36 | 0.40 | 0.24 | 0.07 | 0.90 | 0.02 | - | - | - |
| Exp 1c: Rhyming | 0.26 | 0.68 | 0.06 | 0.32 | 0.56 | 0.12 | - | - | - |
| Exp 1c: Related | 0.37 | 0.51 | 0.11 | 0.12 | 0.79 | 0.09 | - | - | - |
| Exp 2 | 0.41 | 0.56 | 0.03 | 0.17 | 0.75 | 0.08 | 0.23 | 0.65 | 0.13 |
| Exp 3 | 0.35 | 0.47 | 0.18 | 0.27 | 0.60 | 0.13 | 0.26 | 0.62 | 0.13 |
| Exp 4 | 0.44 | 0.44 | 0.12 | - | - | - | 0.05 | 0.75 | 0.20 |

## References


Arnold, K. M., & McDermott, K. B. (2013). Test-potentiated learning: Distinguishing between direct and indirect effects of tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(3), 940.

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & Cognition*, 35(2), 201–210.

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

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

Bjork, R. A. (2001). Recency and recovery in human memory. In H. L. Roediger, J. S. Nairne, I Neath, & A. M. Surprenant (Eds.), *The nature of remembering: Essays in honor of Robert G. Crowder* (pp. 211-232). Washington, DC: American Psychological Association Press.

Bridger, E. K., & Mecklinger, A. (2014). Errorful and errorless learning: The impact of cue–target constraint in learning from errors. *Memory & Cognition*, 1-14.

Butterfield, B., & Metcalfe, J. (2006). The correction of errors committed with high confidence. *Metacognition and Learning*, 1(1), 69–84. doi:10.1007/s11409-006-6894-z

Carpenter, S. K. (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), 1547–52. doi:10.1037/a0024140

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

Cyr, A. A., & Anderson, N. D. (2015). Mistakes as stepping stones: Effects of errors on episodic memory among younger and older adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 841.

Elroy, L. A., & Slamecka, N. J. (1982). Memorial consequences of generating nonwords : implications for semantic-memory interpretations of the generation effect. *Journal of Verbal Learning and Verbal Behavior*, *259*, 249–259.

Fazio, L. K., & Marsh, E. J. (2009). Surprising feedback improves later memory. *Psychonomic Bulletin & Review*, 16(1), 88–92. doi:10.3758/PBR.16.1.88

Grimaldi, P. J., & Karpicke, J. D. (2012). When and why do retrieval attempts enhance subsequent encoding? *Memory & Cognition*, 40(4), 505–13. doi:10.3758/s13421-011-0174-0

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

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

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*(6), 649-667.

Jacoby, L. L., & Hollingshead, A. (1990). Reading student essays may be hazardous to your spelling : Effects of reading incorrectly and correctly spelled words. *Canadian Journal of Psychology*, *44*(3), 345–358.

Johns, E. E., & Swanson, L. G. (1988). The generation effect with nonwords. *Cognition*, 14(1), 180–190.

Kane, J. H., & Anderson, R. C. (1978). Depth of processing and interference effects in the learning and remembering of sentences. *Journal of Educational Psychology*, *70*(4), 626.

Kapur, M. (2014). Comparing learning from productive failure and vicarious failure. *Journal of the Learning Sciences*, *23*(4), 651-677.

Kapur, M., & Bielaczyc, K. (2012). Designing for productive failure. *Journal of the Learning Sciences*, 21(1), 45–83. doi:10.1080/10508406.2011.591717

Karpicke, J. D., & Zaromb, F. M. (2010). Retrieval mode distinguishes the testing effect from the generation effect. *Journal of Memory and Language*,*62*(3), 227-239.

Knight, J. B., Ball, B. H., Brewer, G. A., DeWitt, M. R., & Marsh, R. L. (2012). Testing unsuccessfully: A specification of the underlying mechanisms supporting its influence on

retention. *Journal of Memory and Language*, *66*(4), 731–746. doi:10.1016/j.jml.2011.12.008

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

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *35*(4), 989–98. doi:10.1037/a0015729

Kornell, N., & Vaughn, K. E. (in press). How retrieval attempts affect learning: A review and synthesis. *Psychology of Learning and Motivation*.

Libkuman, T. M. (1994). Norms for words that rhyme. *Behavior Research Methods, Instruments, & Computers*, *26*(3), 278-322.

McGillivray, S., & Castel, A. D. (2010). Memory for age-face associations in younger and older adults: the role of generation and schematic support. *Psychology and Aging*, 25(4), 822–32. doi:10.1037/a0021044

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

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

Osgood, C. E. (1949). The similarity paradox in human learning: A resolution. *Psychological Review*, 132–143.

Potts, R. (2014). *Memory interference and the benefits and costs of testing.* (Doctoral dissertation, UCL (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.

Pyc, M. A., & Rawson, K. A. (2010). Why Testing Improves Memory: Mediator Effectiveness Hypothesis. *Science*, 330(October), 2010.

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

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

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

Roediger III, H. L., Putnam, A. L., & Smith, M. A. (2011). 1 Ten Benefits of Testing and Their Applications to Educational Practice. *Psychology of Learning and Motivation-Advances in Research and Theory*, *55*, 1.

Slamecka, N. J., & Fevreiski, J. (1983). The generation effect when generation fails. *Journal of Verbal Learning and Verbal Behavior*, *22*(2), 153-163.

Slamecka, N. J., & Graf, P. (1978). The Generation Effect : Delineation of a Phenomenon. *Journal of Experimental Psychology: Human Learning and Memory*, 4(6), 592–604.

Soderstrom, N. C., Clark, C. T., Halamish, V., & Bjork, E. L. (2015). Judgments of learning as memory modifiers. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(2), 553.

Soloway, R. M. (1986). No generation effect without semantic activation.*Bulletin of the Psychonomic Society*, *24*(4), 261-262.

Soraci, S. A., Carlin, M. T., Chechile, R. A., Franks, J. J., Wills, T., & Watanabe, T. (1999). Encoding variability and cuing in generative processing. *Journal of Memory and Language*, *41*(4), 541-559.

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

Yan, V. X., Yu, Y., Garcia, M. A., & Bjork, R. A. (2014). Why does guessing incorrectly enhance, rather than impair, retention?. *Memory & Cognition*, *42*(8), 1373-1383.