RETRIEVAL CUE VARIABILITY:
WHEN AND WHY ARE TWO MEANINGS BETTER THAN ONE?

BY

JASON R. FINLEY

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Psychology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2012

Urbana, Illinois

Doctoral Committee:

Associate Professor Aaron S. Benjamin, Chair
Professor Emeritus William F. Brewer
Professor Gary S. Dell
Assistant Professor Brian D. Gonsalves
Professor Brian H. Ross

Abstract

Much is known about the memory benefits of encoding variability, but the effects of retrieval variability (or diversity) remain largely unexplored. The current project investigates the possible benefits and detriments of retrieval cue variability in episodic memory tasks, the processes underlying such effects, and how those effects may interact with encoding conditions. Six experiments tested participants' recall of balanced homographs when cued with a single meaning or with two meanings. Based on the principle of congruity between encoding and retrieval (e.g., transfer-appropriate processing), I predicted that double-meaning cues would be superior by virtue of providing two routes to retrieval, at least one of which would likely overlap with an encoded single meaning. However, single-meaning cues were in fact superior when target homographs had been studied alone (Experiment 1) or not studied at all (Experiment 2). However, when the cue words were disambiguated by being presented with the targets during study, double-meaning retrieval cues indeed yielded higher recall (Experiments 3 and 6). Experiment 4 showed that, when the procedure allowed it, participants often used two double-meaning retrieval cues together in a synergistic way to better home in on the target. Experiments 5 and 6 showed that retrieval cue variability can yield benefits or costs depending on encoding conditions. Double-meaning cues yielded higher performance than single-meaning cues that were incongruent with the encoded meaning, but lower performance than single-meaning cues that were congruent with the encoded meaning. Experiment 6 also showed that participants recalled more when tested with the specific cues that they selected from a set of choices at study, but that they did not have good insights on the benefits of retrieval cue diversity. Overall,

results suggested that retrieval cue variability is beneficial to the extent that cues are unambiguous and that it is not redundant with variability induced at encoding. Furthermore, retrieval cue variability can be useful as a hedge against uncertainty about the past and changing interpretations of ambiguous stimuli.

Keywords: retrieval variability, encoding variability, retrieval strategy, metacognition, recall

Acknowledgments

At the time of this writing, I comprise a small self-aware subset of totality. Other such subsets I would like to thank include: my parents (Elizabeth Anne Robertson/Finley/Kahrs and Richard Terry Finley) for enabling and encouraging my perpetual education, my life partner and fellow traveler Laurel Marie Methot for love and laughter, and cockatiels Beaky and Mr. Wiggles for companionship and sundry hijinks.

I am furthermore grateful to Professor Aaron S. Benjamin for judicious meta-guidance and intrepid rigor, and to Professor William F. Brewer for a cornucopia of arcane knowledge and imaginative blue sky thinking. I thank both for my hybrid intellectual heritage and for fostering my skills with firmness, kindness, and/or good humor. I am proud to become a fellow scientist.

The research herein was supported by funding from the National Institute of Health to ASB (R01 AG026263).

For indexing purposes, my full name is Jason Richard Finley, date of birth July 3rd, 01980, location: Santa Ana, California, United States of America, Earth, Sol System, Milky Way Galaxy, Local Cluster, Virgo Supercluster, Known Universe.

If you are reading this in the distant future, greetings from the past!
The day you stop learning is the day you die.

Table of Contents

Introduction

Memory provides a way of transferring information between points in time. Thinking beings, such as the author and the reader, experience time linearly. We inhabit a single moment of perpetual change. Only memory enables subjective continuity across moments, and the ability to adapt to challenges within a single lifetime. With memory we can relate current experience to prior experiences, access previously encountered information, and store the current contents of awareness for potential future use. In short, we can exist across moments.

Encoding processes send information into the future. Retrieval processes recover information from the past. For over 100 years we have applied the methods of science to understanding these processes (Ebbinghaus, 1885/1913). Of the peculiarities of human memory that we have determined so far, which might a judicious memory user be able to exploit, and how? The primary characteristic of human memory that I consider here is the following: the more alike two points in time, the easier the transferal of information between them. That is, congruity between encoding and retrieval enhances memory performance. This indicates that access to information in human memory is context- and process-dependent, and that successful transfer can be increased by altering characteristics of the two time points.

Congruity Between Encoding and Retrieval

Two points in time can be similar or different in a variety of ways. They can never be identical, but to the extent that some of their features overlap, memory is enhanced. Overlap, or congruity, can occur for: (a) aspects of the external environment surrounding a memory user, (b) the user's own internal affective and cognitive states, and

(c) the cognitive processes the user is executing. I will first review research that has documented the facilitating effects of congruity between encoding and retrieval for each of these three aspects of momentary existence. I will then confront the practical limitations of ensuring congruity across the conditions of encoding and of retrieval, and will proceed to explore how the same overarching principle of congruity indicates that memory use can be improved by introducing variability at encoding or at retrieval. I will review research on the benefits of encoding variability, and the much sparser research on the benefits of *retrieval variability*. Finally, I will report on the current research project, which concerns the processes underlying the possible benefits and detriments of retrieval variability in simple episodic memory tasks. Those experiments will reveal that the consequences of introducing variability at retrieval vary meaningfully but not always obviously with the conditions of encoding.

Environmental context. One way in which similarity can exist between points in time is for two moments to occur in the same location. When the location in which encoding occurred is identical or similar to the location for retrieval, recall performance is generally enhanced (though not always). Numerous laboratory studies have documented this phenomenon (see Bjork & Richardson-Klavehn, 1989; Smith, 1988; Smith & Vela, 2001). For example, in an experiment by Smith, Glenberg, and Bjork (1978, Experiment 3), participants studied a word list in one of two quite different rooms, then were tested one day later in either the same room or the alternate room. Free recall performance was higher for participants who were tested in the same room where they had previously studied. A more dramatic demonstration of the effects of environmental context on memory was provided by Godden and Baddeley (1975, Experiment 1), whose

participants were scuba divers who studied a word list either under water or on land and were later tested either under water or on land. Recall performance was again higher when study and test context were congruent. It has even been shown that mentally reinstating the study environment can enhance test performance (e.g., Smith, 1984). It is worth noting, however, that such environmental context effects are generally small in magnitude—a point to which I will soon return.

Mood- and state-dependency. The success of retrieval is also sometimes enhanced to the extent that an individual's mood and/or mental state are similar at encoding and retrieval. For example, in an experiment by Eich and Metcalfe (1989, Experiment 1), music was used to induce a happy or sad mood at encoding, and again to induce a happy or sad mood at retrieval two days later. Recognition of target words was higher when the two moods matched than when they mismatched. With regard to state dependency, it has been found that memory performance is higher when participants are under the influence of alcohol at both time points versus just one (Goodwin, Powell, Bremer, Hoine, & Stern, 1969), and under the influence of marijuana at both time points versus just one (Eich, Weingartner, Stillman, & Gillin, 1975). As with environmental context effects, mood- and state-dependency effects tend to be small in magnitude, when they are found at all.

Transfer appropriate processing. The preceding examples highlight the memory benefits of congruent external and internal circumstances. As I noted, although these effects are fairly consistent, they are small in magnitude. This probably reflects the fact that external and internal contexts are rarely central to our understanding of information at encoding, unless those contexts are highly distinctive or impinge greatly

on attention (Dallett & Wilcox, 1968). Furthermore, we often cannot control the nature of external and internal contexts at both encoding and retrieval moments. Thus, for the practical memory user, exploiting context effects may not be an efficient use of time and effort.

Let us then consider a different form of congruity between time points, one which yields larger effects and over which the average memory user is more likely to have control: the nature of the cognitive operations carried out by the user. The idea of *transfer-appropriate processing* is that the way in which information is processed effects the representation of that information in memory, and that memory performance is enhanced to the extent that cognitive processes are similar at encoding and at retrieval. I will next review two key studies that illustrate this phenomenon.

Morris, Bransford, and Franks (1977, Experiment 1) presented participants with single words, half of which were preceded by a sentence that induced semantic processing (e.g., "The — had a silver engine." ... "TRAIN") and the other half of which were preceded by a sentence that induced phonetic processing (e.g., "– rhymes with legal." ... "EAGLE"). Participants gave a yes/no judgment on whether the word made sense in the sentence or indeed rhymed, respectively. Participants then received either a standard recognition test on the studied words or a recognition test on new words that rhymed with the studied words. For the standard recognition test (which Morris et al. presumed would emphasize the meaning of words; cf. Craik & Lockhart, 1972), performance was highest for items that had undergone semantic processing at encoding. In contrast, for the rhyming recognition test, performance was highest for items that had

undergone phonetic processing at encoding. Thus, for a given type of test, similar processing at study yielded better performance.

Fisher and Craik (1977, Experiment 3) presented participants with cue-target word pairs, half of which featured an associative relation (e.g., sleet - hail) and the other half of which featured a rhyming relation (e.g., pail - hail). Participants then received a cued recall test in which some trials featured new associative cues (e.g., "associated with snow") and some trials featured new rhyming cues (e.g., "rhymes with bail"). For target words that were studied with associative cues, performance was highest for the associative test cues. In contrast, for target words that were studied with rhyming cues, performance was highest for the rhyming test cues. Thus, for a given type of study, similar processing at test yielded better performance.

The review so far shows that there are several forms of congruity between time points (i.e., encoding and retrieval) that can increase the probability of successful transfer of information across time within an individual: external context, internal context, and cognitive processing. These can be exploited to some extent by a judicious memory user. A user may study in a location where s/he is likely to be tested, or s/he may physically return to a prior location to assist retrieval of information encoded there. A user may postpone encoding until his/her mood and state of mind are congruent with foreseen retrieval scenarios, or s/he may attempt to improve retrieval by reproducing or emulating a prior internal context (e.g., using emotionally arousing stimuli, exercise, food, or drugs). Perhaps most effectively, a user may think about information at encoding in a way likely to be similar to the processes required at retrieval, or s/he may attempt to perform retrieval processing that imitates encoding processing. Any such strategies could be used

to increase congruity between encoding and retrieval, and thus enhance successful memory use.

The Relationship Between Congruity and Variability

We have seen that increasing congruity between the moment of encoding and the moment of retrieval can enhance memory. A savvy memory user can exploit this characteristic of memory to his/her advantage. But such a strategy suffers from two major practical limitations: uncertainty about the future, and the very fallibility of memory itself.

First, except in special cases such as academic tests or artistic performances, we can rarely foresee the precise circumstances under which the information we are currently encoding may be needed in the future. In fact, to some extent the future retriever will be a different person from who we are at the moment of encoding. So simulation of retrieval contexts and processes may not be possible. Second, when the time comes to retrieve, we may not be able to remember the precise conditions under which needed information was encoded. In fact, for generalized knowledge, all specific prior moments may be lost into an amalgamation. So recapitulation of encoding contexts and processes may not be possible. Thus, it is often not feasible to seek congruity between encoding and retrieval.

What then *is* a good strategy for improving the chances of successfully transmitting information between two uncertain situations? That is, how can memory be made more generalizable? The answer is variability. Introducing variability at either encoding or retrieval should enhance transfer on average, because it increases the probability of some congruity between time points. Thus, for example, a judicious

memory user may study in multiple locations if the test location is unknown, or s/he may attempt retrieval of elusive information in a variety of locations. A user may encode the same information under various affective and cognitive states, or s/he may alter internal context upon retrieval failure. A user may consider information from a variety of perspectives upon encoding, or do so at retrieval.

It may seem that the average memory user should be more easily able to implement variability at encoding than at retrieval. Encoding can be an intentional task performed under one's own terms, whereas retrieval is often instigated by demands from the world external to the user. But the increasing permeation of information technology into our lives brings more opportunities to share the burden of memory tasks between internal and external (offloaded) representations and processes. For such shared tasks, users often have the ability to pre-arrange future retrieval cues at the time of encoding. For example, the notes a user takes or the scheduled reminders s/he adds to a calendar might employ shorthand that is easily understood at the time of encoding, but may appear obtuse to the same user at a later date. The same challenge applies to many tasks of prospective memory (McDaniel & Einstein, 2007) and personal information management (Jones & Teevan, 2007). When sending information to our potential future selves, using greater diversity should increase the probability of successful transmission. Furthermore, there are in fact a number of ways in which memory users can exercise control at the time of retrieval, such as terminating search (Harbison, Dougherty, Davelaar, & Fayyad, 2009), switching retrieval cues (Young, 2004), controlling output specificity (Goldsmith & Koriat, 2008; Koriat & Goldsmith, 1996), referring to the environment (Ballard, Hayhoe,

Pook, & Rao, 1997; Sparrow, Liu, & Wegner, 2011), and using multiple retrieval strategies (Williams & Hollan, 1981).

The general claim about the memory benefits of variability is supported by evidence from diverse fields of memory research. I will first review the ample research on encoding variability, and then review the much sparser research on retrieval variability.

Encoding Variability

The principle of encoding variability (a.k.a. contextual variability, varied context repetition) originated with the fluctuation model of Estes (1955; cf. Bower, 1972; Martin, 1968), and has been promoted as a theoretical explanation for many aspects of learning, including acquisition and extinction (Chelonis, Calton, Hart, & Schachtman, 1999; Estes, 1955), interference effects (Mensink & Raaijmakers, 1988), and the spacing effect (Glenberg, 1979; Melton, 1970). The spacing effect is the finding that multiple instances of encoding are more effective when spaced apart (i.e., distributed practice) versus massed together (Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Glenberg, 1979; Hintzman, 1974; Melton, 1970). The idea behind encoding variability is that the context of encoding is always fluctuating, even if only minutely. Context may be physical, emotional, and/or cognitive. Furthermore, some subset of these contextual cues is encoded associatively along with the to-be-remembered information. The further apart in time two instances of encoding, the more diverse the contextual cues will be. The greater the diversity of encoded contexts, the more likely that one or more will provide a good match for an eventual retrieval context, thus increasing the rate of successful recovery of the to-be-remembered information (encoding specificity). More generally, the more

different ways something is encoded, the more routes to retrieval exist, and thus the greater the chances of retrieval success.

The merits of encoding variability as an explanation for the spacing effect are not at issue here (cf. Benjamin & Tullis, 2010; Ross & Landauer, 1978). Instead, I consider evidence consistent with the general claim that variability in the circumstances and processes of encoding renders memory traces that are more generally retrievable in a variety of conditions.

Context variability at encoding. Returning to the context-dependency literature, benefits have been found for studying material in multiple contexts (see Bjork & Richardson-Klavehn, 1989; Smith. 1988). For example, in an experiment by Smith, Glenberg, and Bjork (1978, Experiment 1), participants studied a list of words either twice in the same room or once each in two different rooms, and then were tested in a new "neutral" room. Free recall performance was higher for participants who had studied in two different rooms, demonstrating what Smith has called a multiple-learning-context effect (Smith, 2007). It has also been found that studying different sets of material in different contexts reduces proactive interference (Dallett & Wilcox, 1968) and retroactive interference (Bilodeau & Schlosberg, 1951).

The benefits of contextual variability at encoding have even been applied to advertising research. In an experiment by Burnkrant and Unnava (1987), participants who had seen three different versions of an ad for Dewar's Scotch liquor (each ad featuring a picture of a different person) were more likely to free recall the brand name than participants who had seen a single version three times. This recall difference was not accompanied by a reliable difference in self-reported attention paid to the ads.

Variability in practice for motor learning. The scheduling of practice trials when multiple motor sequences are to be learned presents another opportunity to introduce variability at encoding. Practice trials for a single movement pattern can either be blocked all together or interleaved with practice trials for other movement patterns (a.k.a. random practice, variable practice). The latter schedule introduces more variability in the learner's activities across an entire practice session.

In an experiment by Shea and Morgan (1979), participants grasped a tennis ball and used it to knock over a series of barriers in one of three particular sequences, as cued by three different lights. Trial scheduling was either blocked or interleaved. Ten days later, participants returned and performed the three motor tasks again, as quickly as they could. Speed was highest for participants who had learned under the interleaved schedule, even though their performance had been slower during acquisition. Thus, the interleaving schedule constituted what Bjork has termed a desirable difficulty (Bjork, 1994). Similar findings across a variety of motor tasks are reviewed by Brady (2004, 2008) and by Magill & Hall (1990).

Variability during practice can also be introduced to a single dimension of a particular motor task. For example, practicing tossing beanbags from variable distances yielded better final performance than practicing from a single distance (Kerr & Booth, 1978). Similar findings on variable practice have been reported by: Catalano and Kleiner (1984); Lee, Magill, and Weeks (1985); Shea and Kohl (1990); and Wulf and Schmidt (1997).

Semantic variability at encoding. Encoding variability has also been implemented on the semantic level. In an experiment by Greenberg & Verfaellie (2010),

amnesic patients and healthy control participants (mean age 57.5, SD = 13.1) studied pairs of nouns separated by a single verb or multiple verbs across three repetitions (e.g., ARMY invades CITY, ARMY flees CITY, ARMY patrols CITY). For the control participants, associative recognition performance (i.e., recognizing intact vs. rearranged pairs) was higher for the varied than for the fixed encoding context.

In an experiment by Hintzman and Stern (1978, Experiment 2) participants read famous names (e.g., Shakespeare) embedded in single repeated sentences or multiple different sentences (e.g., peeled an apple, bought a car, kicked a dog), and rated the plausibility of each sentence. On a surprise free recall test, performance was higher for names that had appeared in multiple different sentences versus a single repeated sentence. Similarly, Glanzer and Duarte (1971), whose participants were bilingual, found that studying a word in two languages led to higher free recall performance than studying a word twice in a single language, particularly at the smallest lags between repetitions.

A number of studies have taken advantage of homographs (a.k.a. polysemous words), which are written words with more than one meaning, such as foot. For example, in two experiments by Gartman and Johnson (1972, Experiments 2 and 3), participants studied a series of words in which some words were homographs that appeared twice, each time being immediately preceded by two context words. The context words either induced the same meaning of the homographs on both occasions (e.g., leg, neck, foot; arm, hand, foot) or induced a different meaning on each occasion (e.g., leg, neck, foot; inch, meter, foot). Free recall performance was considerably higher for items in the double-meaning condition versus the single-meaning condition (M_{diff} = 42% across experiments and lags).

Retrieval Variability

Just as encoding information in multiple ways can increase the probability of successful transmission between time points, attempts to retrieve information in multiple ways should also improve performance. However, much less research has been done on varying context and/or processes at retrieval, especially for basic episodic memory tasks. The principles behind the benefits of encoding variability may or may not generalize to retrieval variability. To my knowledge, no one has considered this before. I will next review the few disparate studies that have borne on retrieval variability and have suggested the benefits thereof. Note that these examples have not, until now, been interpreted in the context of the general value of retrieval variability.

Autobiographical and eyewitness memory. There is some evidence that retrieval variability can benefit autobiographical memory. In one study (Williams, 1977; Williams & Hollan, 1981; Williams & Santos-Williams, 1980), four participants attempted to recall names of high school classmates across 4 to 10 one-hour sessions, all while thinking aloud. Although most of the participants stated early in the first session that they could not recall any more names, they were all still recalling new names even in their final sessions. Remarkably, each had recalled at least 90 correct names by the end of the study. Analysis of the verbal protocols revealed five common retrieval strategies: activity-cued search, location scanning, image scanning, general association, and name generation. Williams (1977) stated that "frequently a subject would shift from one strategy to another when the second seemed to hold out hope for improved success, and then shift back a few moments later when the second strategy ceased to be productive." (pp. 26-27). Although not clear from the reported data, it was in fact the case that

strategy shifts were accompanied by upticks in recall (D. M. Williams, personal communication, December 11, 2011). While cumulative recall continually climbed across sessions, strategy use was not equally distributed across sessions. For example, the activity-cued search (e.g., searching for names of people who had been on the baseball team) was used less and less across sessions as it became harder to generate new activities. In contrast, the name generation strategy (e.g., traversing the alphabet to generate common names to use as retrieval cues) was used more in later sessions. This pattern suggests that the different strategies enabled access to different names.

In an experiment by Whitten & Leonard (1981, Experiment 2), participants spent up to 30 minutes recalling names of their teachers from grade 1 through 12, all while thinking aloud. Analysis of the verbal protocols revealed at least four different retrieval strategies: subject enumeration, location search, reference to disruptions in life, and directional search. Although not clear from the results reported, it is implied that individual participants used more than one strategy, and that they invoked different strategies when retrieval became difficult. Unfortunately, performance was not reported as a function of self-reported retrieval strategy use.

In the realm of eyewitness memory research, varied retrieval cues have been found to increase the recall of previously unrecalled information (i.e., reminiscence). In a study by Gilbert and Fisher (2006), participants watched a 3 minute police training video of a simulated bank robbery attempt. After a 10 minute distractor task, they took a test which instructed them to use one of four retrieval strategies: chronological order, reverse order, police's perspective, or robber's perspective. Participants returned 48 hours later and took a second test, which either had the same retrieval strategy instructions as the

first test or one of the other three strategies. A separate baseline group of participants took a free recall test on both occasions (no particular retrieval instructions). Participants who changed retrieval strategies between tests produced a reliably larger number of reminiscences than those who had the same retrieval strategy on both tests, including the free recall group. The use of multiple retrieval strategies has been included as part of the cognitive interview, a method for improving recall by crime eyewitnesses and victims (Fisher, Geiselman, Raymond, Jurkevich, & Warhaftig, 1987; Memon, Meissner, & Fraser, 2010).

Semantic memory. In a study by Walker and Kintsch (1985), participants spent 12 minutes attempting to recall members of categories such as automobiles, all while thinking aloud. Analysis of verbal protocols revealed that individual participants made use of multiple retrieval strategies (in this case, different types of retrieval cues used to search memory), many of which were episodic in nature (e.g., friends' cars). Furthermore, most participants initially reported a handful of easily recalled category members, then began to think of likely situations to use as retrieval cues. The switch to a new retrieval strategy or cue when the current one becomes unfruitful is one way of explaining the scalloped shape of individual participants' cumulative recall curves. Although the data from this study clearly illustrate that participants indeed used multiple retrieval strategies, the data merely hint that such variability directly improved recall.

When humans make quantitative estimates under uncertainty (e.g., when forecasting probabilities of future events), it has been found that averaging across the estimates of multiple individuals tends to yield higher accuracy than the individual estimates themselves (see Wallsten, Budescu, Erev, & Diederich, 1997). This

phenomenon has been termed the *wisdom of crowds* (Surowiecki, 2004), and the idea has also been extended to the case of multiple estimates made by a single human, the *crowd within* (Vul & Pashler, 2008). The idea is that estimates are made at least in part based on retrieval of some relevant information from semantic memory. To the extent that two estimates sample different subsets of an individual's memory, the average of the estimates will tend to be closer to the true value. Simply separating two estimates across time can yield variability in the information retrieved, and thus also in the estimates (Vul & Pashler, 2008). Hourihan and Benjamin (2010) demonstrated that this effect was even more pronounced for participants with lower working memory spans, whose smaller samplings of memory were presumably less likely to overlap on the two occasions.

Herzog and Hertwig (2009) proposed a strategy to further induce variability in estimates, which they termed *dialectical bootstrapping*. The strategy encouraged participants making their second estimate to retrieve different information from memory, for example by considering ways in which their first estimate might have been off. In their experiment, participants made two estimates of the dates of historical events (e.g., the discovery of electricity), and were either told to use the dialectical bootstrapping strategy for their second estimate, or were given no strategy for their second estimate. The dialectical bootstrapping strategy yielded higher accuracy (averaged across both estimates) than no strategy, demonstrating the benefit of introducing variability to retrieval in the service of estimation.

Episodic memory. In a study by Anderson & Pichert (1978), participants were instructed to adopt the perspective of either a burglar or a homebuyer before reading a story about two boys exploring a house. The story included some details important to a

burglar but not a homebuyer (e.g., a coin collection) and some details important to a homebuyer but not a burglar (e.g., a leaky roof). After a 12 minute distractor task, participants took a first free recall test (with no special instructions about perspective), completed a 5 minute distractor task, then finally took a second free recall test that either repeated the same perspective instructions given before encoding or gave the alternate perspective instructions. It is not clear from the reported results whether participants who switched perspectives recalled more overall (across both tests), but it was the case that switching perspectives led to recall of additional information that had been unimportant to the first perspective, and did so to a greater degree than keeping the same perspective. Thus, a shift in retrieval perspective appeared to enable recovery of information that was unrecoverable using the first perspective.

In an experiment by McLeod, Williams, and Broadbent (1971), participants studied target words (15 nouns and adjectives) presented without any cues, completed a 30 second math distractor task, then took a free recall test followed by two cued recall tests. The first cued recall test included only targets that the participant failed to retrieve in free recall. For half of the participants the cues for each target were one cue word that was highly associated to the target; for the other half of the participants, the cues for each target were *two* cue words highly associated to the target. The second free recall test included only targets that the participant failed to retrieve in the first cued recall test, and used the alternate number of cues. The order of cue condition was counterbalanced between subjects. Each target's two cue words were never associated with each other, indicating that target words likely had multiple meanings (though the authors did not state that they used homographs). Cued recall performance was higher with two cues (M

= .84) than with one cue (M = .73). This suggests that variability in retrieval cues (or cue diversity) may improve recall, although in this case cue diversity was confounded with the number of cues.

Current Project

The aim of the current project was to investigate the possible benefits and detriments of retrieval cue variability in episodic memory tasks, to elucidate the processes underlying such effects, and to examine how they may interact with encoding conditions. An additional goal was to investigate memory users' metacognition with respect to retrieval variability. For example, to what extent would users appreciate the value of variability, and to what extent would their study and testing choices exploit it?

As demonstrated in the review above, much is known about encoding variability, and comparatively little about retrieval variability. Searching memory in multiple ways or with multiple cues may or may not be equivalent to encoding information in multiple ways or contexts. Any benefits of retrieval variability may depend on: (a) the retrieval strategy or strategies used; (b) any retrieval cues used in those strategies; and (c) the way in which the information was encoded (e.g., it could be that variability at encoding obviates the need for variability at retrieval). My overall approach to these factors will be to: (a) allow participants to use whatever retrieval strategies they wish, but also record self reports and questionnaire responses about strategies; (b) systematically manipulate retrieval cues within-subjects; and (c) allow encoding strategies to vary naturally but also manipulate encoding cues, across experiments or conditions (Experiments 1-4) or within-subjects (Experiment 5).

The stimuli used in this project's experiments were homographs. Retrieval cue variability (or diversity) was manipulated within-subjects such that some recall trials featured cues pointing to just one meaning of a target (single-meaning or convergent) and other recall trials featured cues pointing to two different meanings of a target (double-meaning or divergent).[1] Unlike in most previous studies, multiple encoding and/or retrieval cues were presented simultaneously in most cases. The first three experiments served both to test the prediction that retrieval cue variability would improve recall, and to begin to elucidate the cognitive processes underlying the effects of retrieval cue variability. They accomplished this by each using different encoding conditions (study of targets only, no study, and study of targets with four cues, respectively). The fourth experiment served to provide insight into the retrieval processes participants used when confronted with double-meaning retrieval cues, by manipulating study condition (cues absent vs. present) and test condition (simultaneous vs. sequential presentation of retrieval cues). The fifth experiment served to investigate the effect of retrieval cue diversity as a function of number of meanings encoded, by manipulating the types of cues presented at encoding, within-subjects. The sixth experiment explored what metacognitive choices and judgments memory users would make in a situation where cue diversity may be beneficial.

[1] The exception is Experiment 4, in which only double-meaning cues were used.

Experiment 1

The purpose of Experiment 1 was to investigate the effect of retrieval cue variability on performance when participants had studied homograph target words alone, with no cues presented at study. Winograd and Conn (1971) found that participants studying such words typically encoded only one of the possible meanings (usually the more dominant one), as evidenced by recognition performance as a function of meaning context.[2] The homographs in the current experiment had two meanings of roughly equal dominance, so let us assume that each participant will encode one of a target's two meanings with 50% probability. Given this encoding situation, we can make predictions about relative performance for single-meaning versus double-meaning retrieval cues, based on the principle of encoding-retrieval congruity reviewed earlier. When participants receive a pair of single-meaning cues at test, there should be a 50% chance that the test meaning is congruent with the (likely) single meaning they encoded at study. In such cases, the probability of retrieval should be relatively high. However, in the cases where the test meaning is incongruent with the encoded meaning, the probability of retrieval should be relatively low. This issue will be addressed directly in Experiment 5. Overall single-meaning performance will of course be an average across both scenarios. When participants receive a pair of *double*-meaning cues at test (i.e., two words that each point to one of the two meanings), one of the cues should always be congruent with the encoded meaning. To the extent that retrieval success is equally probable in the double-meaning and congruent single-meaning cases, performance should be superior for the

[2] Unlike the current project, the experiments by Winograd and Conn (1971) used homographs with one dominant meaning (i.e., polarized or unbalanced). However, the extension of their finding to balanced homographs will be borne out by the results of the current project (e.g., questionnaire data in Experiments 4 and 5).

double-meaning condition, because it should not be dragged down by incongruent cases. Thus, I predicted that retrieval cue diversity would enhance cued recall performance.

Method

Participants. Participants were 32 undergraduates who received partial course credit. Eleven were female, six reported that English was not their first language (although they were all fluent in English), and their mean age was 19.1 years (SD = 1.3). One additional participant did not follow instructions and the data from this participant were excluded from analysis.

Materials. Materials were 60 English homograph target words along with four associated cue words for each target (two cues for each of two meanings). For example, one target word was *bat* and its four cue words were *swing, hit, fangs,* and *cave*. Target words were 3-8 letters long (M = 4.6, SD = 1.2) and their HAL frequency ranged from 716 to 552,532 (M = 63,771, SD = 103,417; Balota et al., 2007; Lund & Burgess, 1996). Target words were balanced homographs, having two meanings of roughly equal dominance (Twilley, Dixon, Taylor, & Clark, 1994). The mean forward associative strength (cue-target) across all meanings and targets was .05 (SD = .05, *range* = .01 - .53) as obtained from the University of South Florida Word Association, Rhyme and Word Fragment Norms (Nelson, McEvoy, & Schreiber, 1998).[3] Across target items, there was no reliable difference in associative strength between meanings ($t(48) = 0.25$, $p = .800$). Stimuli are presented in Appendix A.

Design. There was one within-subjects independent variable with two levels: single-meaning retrieval cues at test versus double-meaning retrieval cues at test. The

[3] Normed associative strength data were not available for all cue-target pairs.

dependent measure was cued recall performance. Self reports on study and test strategies were also collected.

Procedure. For reference, Table 1 shows an overview of the procedures for all experiments in the current project. For all experiments, participants were run individually on computers programmed with REALBasic. Unless otherwise noted, all instructions and stimuli were presented visually on the computer screen, and all participant responses were made using the mouse and/or keyboard. For Experiment 1, initial instructions informed participants that they would be studying target words on which they would later be tested. Participants were then presented with the target words in a randomized order for 8 seconds each with a 0.5 second blank screen inter-stimulus interval. All words were shown in black type on a white background and the label "Target Word" appeared above the target word. Instructions that remained at the top of the screen read: "Study the below words for the upcoming test.".[4] After this initial study phase, participants engaged in a 5 minute distractor task similar to the game Bejeweled (http://en.wikipedia.org/wiki/Bejeweled), in which they attempted to match 3 or more tiles with the same symbol in an 8 x 8 grid of tiles where 7 different symbols were possible.

Participants then completed a self-paced cued recall test, in which they were shown two cue words for each target word, and were instructed to type in the corresponding target word or to type a question mark if they could not remember the

[4] I do not follow the custom of placing trailing punctuation within quotation marks, on the grounds that it is silly at best and misleading at worst. Therefore, punctuation will only appear within the quotation marks if that punctuation is in fact part of what is being quoted. Any trailing punctuation will always appear outside of the quotation marks (just as it does for parentheses). This may look unusual but it improves clarity.

target word. Test order was random, and the cue words were positioned in a vertical column labeled "Cue Words", in a random order, to the left of the response field labeled "Target Word". Target words were randomly assigned such that half were tested with a pair of cues that pointed to a single meaning of the target (single-meaning condition; e.g., *swing, hit*) and half were tested with a pair of cues that pointed to two meanings of the target (double-meaning condition; e.g., *swing, fangs*). Which particular meaning was used for the single-meaning condition for a given target was counterbalanced between-subjects. For the double-meaning condition, one cue was randomly selected from each of a target's two meanings.

After completing the test, participants answered two free response questions which asked them to describe any strategies they had used during the study phase, and any strategies they had used during the test phase.

Results

For all experiments in the current project, an alpha level of .05 is used for all tests of statistical significance. Effect sizes for comparisons of means are reported as Cohen's *d* calculated using the pooled standard deviation of the groups being compared (Olejnik & Algina, 2000, Box 1 Option B). Standard deviations reported are uncorrected for bias (i.e., calculated using *N*, not *N*-1). Finally, response times reported are measured from stimulus onset to the first letter typed.

Of key interest, cued recall performance was in fact reliably higher for the *single-meaning* condition ($M = .25$, $SD = .16$) versus the double-meaning condition ($M = .19$, $SD = .12$), $t(31) = 2.58$, $p = .015$, $d = 0.39$. Participants' median response times were reliably longer for the double-meaning condition ($M = 5.88$ s, $SD = 3.03$) versus the

single-meaning condition (M = 4.66 s, SD = 1.79), $t(31) = 3.20$, $p = .003$, $d = 0.49$. Thus, participants spent more time on double-meaning test trials, but performed less well on these trials.

The modal self reported study strategies were making associations across targets (n = 15) and rote rehearsal (n = 14). Some participants reported both strategies. Other reported strategies included imagery and relating targets to something personal (cf. Finley & Benjamin, in press). Participants self reported a variety of test strategies. The most commonly mentioned strategy component was to figure out the relationship between the two cue words (n = 12). That relation was sometimes used to search for target words, sometimes compared to target words that had already been generated by free recall, sometimes used to generate candidate words from semantic memory, and sometimes elicited an immediate and apparently effortless response. One participant eloquently described how s/he confronted the challenge of incongruity between studied and tested meanings: "Also the strategie was to think of all aspects and meanings of the word because it was clear that they purposefully gave cue words that were not necessarily related to what you would associate the word with initially, but rather another meaning.". Finally, some participants reported that they had used no strategy at all. The range of responses informed the creation of standardized strategy questionnaires used in Experiments 4 and 5. Analyses relating strategy questionnaire responses to test performance will be reported in the General Discussion.

Discussion

Recall results were the opposite of those predicted based on the congruity principle. That is, retrieval cue diversity was in fact detrimental to performance. How

can we explain this? Let us consider the kinds of retrieval processes that could have been at work, and how those might function in the face of double-meaning retrieval cues.

A relevant theory of recall is that of generate-recognize (for a review, see Watkins & Gardiner, 1979). Although it was devised to explain free recall, it can be readily adapted for cued recall. The original version of the theory first posits that items (e.g., words) are stored discretely in a permanent knowledge base (i.e., semantic memory). When an item is encountered during a study phase, its representation in the knowledge base is marked with an occurrence tag. Later, on the free recall test, the participant generates (i.e., searches for) a set of candidate words from semantic memory, then checks the candidates for occurrence tags (i.e., recognizes), and reports words that have that tag. Note that the same general two-process framework can be applied even if we suppose that encoding and recognition do not employ an occurrence tag process. For the present purposes I am not committed to a particular theory of how recognition works. As for the generation/search process, it can be guided by cues, and/or organization (e.g., categories), so that the entire contents of semantic memory do not need to be searched every time. Also note that although generate-recognize was first proposed as a theory to account for all free recall, we can also think of it as just one possible retrieval strategy at participants' disposal. To the extent that participants used such a strategy (and results from Experiments 4 and 5 suggest that they did), there are then two processes for which single- versus double-meaning retrieval cues might have differential effects: generate/search, and recognition. If one process can be ruled out, we can focus on the other. I first consider the recognition process.

Suppose that the target word is in fact generated as one of the candidate responses; it must still be recognized as having been previously studied. Are there reasons to suspect that the success of the recognition process may vary for single- versus double-meaning retrieval cues? First let us consider the case in which the single-meaning cues are incongruent with the encoded meaning yet the target word has nevertheless been generated. If an occurrence tag was attached to the target word at encoding, as proposed by the original versions of the generate-recognize theory, it may be that the tag itself was associated with features of the word relevant to the encoded meaning. Thus the tag should be less easily recovered when the target word is activated with features relevant to the unencoded meaning. Alternatively, it may simply be that the target word is more difficult to recognize in the presence of other candidates generated from the semantic area of the unencoded meaning. This would be in line with the phenomenon of recognition failure of recallable words (Tulving, 1974; Tulving & Thomson, 1973; Watkins & Tulving, 1975). This same problem would occur for the double-meaning retrieval cue cases, perhaps to a lesser extent since some encoded-meaning candidates could also be generated.

For single-meaning cues, half of the cases should pose some contextual challenge to recognition and half should not. For double-meaning cues, every case should pose some degree of contextual challenge to recognition. If these latter challenges outweigh those in the single-meaning cases, then double-meaning cue performance may suffer relative to single-meaning performance. Experiment 2 investigated this possibility by eliminating the recognition process to see if the performance disparity still remained.

Experiment 2

The purpose of Experiment 2 was to test whether retrieval cue variability affects the generation of cues from semantic memory in the absence of any episodic effects (i.e., recognition processes). Toward this end, the study phase was completely eliminated. If the advantage of single-meaning retrieval cues in Experiment 1 was solely due to the recognition process of a generate-recognize strategy, then that advantage should disappear here, because there can be no recognition process in the absence of a study phase.

Method

Participants, materials, and design. Participants were 33 undergraduates who received partial course credit. Eighteen were female, fifteen reported that English was not their first language, and their mean age was 19.1 years (SD = 1.1). Materials and design were the same as in Experiment 1, with the exception that self reports were only collected for test strategies.

Procedure. The procedure was identical to that of Experiment 1, except that there was no initial study phase. That is, participants began with the distractor task. On the test, participants were instructed to type in the word that both cue words were associated with. After the test, participants free-responded to just the question asking them to describe any strategies they had used during the test phase.

Results

Performance was once again reliably higher for the single-meaning condition (M = .12, SD = .08) versus the double-meaning condition (M = .06, SD = .05), $t(32) = 4.43$, $p < .001$, $d = 0.95$. Participants' median response times were reliably slower for the

double-meaning condition (M = 6.33 s, SD = 4.28) versus the single-meaning condition (M = 4.59 s, SD = 4.66), $t(32) = 4.92$, $p < .001$, $d = 0.39$. Thus, participants again spent more time on double-meaning test trials, but performed less well on these trials.

As in Experiment 1, the most commonly mentioned retrieval strategy component was to figure out a relationship between the two cue words (n = 22). In some cases participants stated that if they could not figure out that relationship, they would respond with a word that was associated with only one of the cue words. Other strategies described by multiple participants were using one cue word at a time and relying on the first candidate word that came to mind (i.e., free association).

Discussion

Single-meaning retrieval cues were again more likely to elicit the target homographs than double-meaning retrieval cues, still in contrast to the prediction based on the principle of congruity. First, this indicates that the results of Experiment 1 were likely not due solely to differences in a recognition process of the generate-recognize strategy, because no such process could have been carried out in Experiment 2. Thus, the generate/search process is implicated. Furthermore, because the inferiority of double-meaning cues was replicated in a situation with no study phase whatsoever, then it must be that the generate/search process was affected by something about the retrieval cues themselves. The best explanation is *under-specificity of retrieval cues* (or cue ambiguity). In the double-meaning cases, each of the two meanings is represented by only one cue word. It may be that a single word is often inadequate to delimit the intended meaning, so the search ends up sampling a broader semantic area. Thus, participants may end up searching irrelevant areas of memory, generating candidate responses that do not even

include the target. Some of the cue words were themselves homographs, which by themselves could point to both relevant and irrelevant meanings. Recall the example target word, *bat*, and its four cue words: *swing, hit, fangs,* and *cave*. Now consider the case of a double-meaning cue pair, in which only one word is given for each meaning (e.g., *swing* and *fangs*). The word *swing* on its own may instigate search across the semantic area concerning the action of swinging a limb or implement through space, *and/or* the semantic area concerning parks and playgrounds, which in this case happens to be irrelevant. In contrast, in the case of a single-meaning cue pair, the cue word *swing* would be accompanied by *hit* and the combination of the two would delimit search to just the relevant semantic area.

Thus, double-meaning retrieval cues may only be beneficial when they point to both of the two relevant meanings just as reliably as the single-meaning cues point to a single relevant meaning. Experiment 3 served to test this idea.

Experiment 3

The purpose of Experiment 3 was to investigate the effect of retrieval cue variability when the problem of retrieval cue under-specificity was ameliorated. This was accomplished through an initial study phase in which participants studied all four cue words alongside each target word. I reasoned that this initial exposure should help disambiguate the intended meaning of individual cue words. As such, when cue words were encountered on the test, they would be more likely to delimit search to only the relevant semantic area, whether in the context of a single-meaning or double-meaning pair. To the extent that this was accomplished, the principle of congruity leads to the prediction that retrieval cue diversity should enhance cued recall performance, unlike what was seen in Experiments 1 and 2.

Method

Participants, materials, and design. Participants were 50 undergraduates who received partial course credit. Twenty-eight were female, nine reported that English was not their first language, and their mean age was 19.3 years (SD = 1.5). Materials and design were the same as in Experiment 1.

Procedure. The procedure was similar to that of Experiment 1, except that the four cue words were presented next to the target words during the study phase. For each target, the four cue words were positioned in a vertical column, in a random order, to the left of the target word. The label "Cue Words" appeared above the cue words, and the label "Target Word" appeared above the target word. Instructions that remained at the top of the screen read: "Study the below cue and target words for the upcoming test.". Participants had also been told in the initial instructions that they would be studying

target words with four cues each, and that they should study the cues to help them remember the target words on the test. Presentation duration was again 8 s. The distractor task, test, and final strategy questions were exactly the same as in Experiment 1. Note however that in all cases, the two cue words given at test had previously been studied alongside the target word. That is, cue words were not new words.

Results

Cued recall performance was higher for the double-meaning condition (M = .48, SD = .19) versus the single-meaning condition (M = .45, SD = .19), although this difference did not reach statistical significance ($t(49) = 1.71$, $p = .094$, $d = 0.14$).[5] Participants' median response times were slightly slower for the double-meaning condition (M = 4.13 s, SD = 2.58) versus the single-meaning condition (M = 3.71 s, SD = 2.39), although this difference did not reach statistical significance, $t(49) = 1.93$, $p = .059$, $d = 0.17$.

The modal self-reported study strategy was making associations between cue and target words. Test strategy responses were lost due to a programming error. Strategy self reports from the first three experiments informed the creation of a standardized questionnaire used in Experiments 4 and 5.

[5]Higher overall performance levels for Experiment 3 versus Experiment 1 are consistent with the results from Tulving and Osler (1968), who found that associative retrieval cues were only beneficial if they had been presented with the targets at study. The fact that Experiment 1 performance was above floor is curiously inconsistent with that previous finding. But this discrepancy may be attributed to Tulving and Osler's shorter presentation time (2 s vs. 8 s) and lower mean cue-target associative strength (.013 vs. .051).

Discussion

Results from Experiments 1 through 3 are plotted together in Figure 1. By analogy to encoding variability, I originally predicted that retrieval cue variability would benefit recall of previously studied balanced homographs. Experiment 1 employed a study situation (targets only) which has been found to induce encoding of a single meaning per homograph (Winograd & Conn, 1971). Based on the principle of congruity, recall should be facilitated when the meaning cued at test is the same as the one encoded. Because that congruity should only happen about 50% of the time for a pair of single-meaning retrieval cues, and 100% of the time for a pair of double-meaning retrieval cues, I predicted that double-meaning retrieval cues should yield superior performance. In Experiment 1, the opposite occurred.

The generate-recognize theory of recall provided a framework for thinking about what processes may have led to the unexpected superiority of single-meaning retrieval cues. This superiority was again obtained in Experiment 2, suggesting that the shortcomings of retrieval cue variability could not be restricted to a recognition process, since that experiment had no episodic memory component. Thus, the effect likely stems from a generate/search process.

With Experiment 3 I sought to determine whether the detriment to the generate/search process for double-meaning cues in Experiments 1 and 2 was caused by cue under-specificity. By presenting all four cues at the time of study, their meaning was presumably disambiguated, and indeed the pattern of recall performance reversed, although not to the point of statistical significance. One reason for the lack of significance could be that the presentation of all four cues at study led participants to

encode both meanings of targets, thus evoking encoding variability and curtailing the multiple-route benefit of double-meaning cues predicted from the principle of congruity. Experiment 5 will address this directly by manipulating the number of meanings presented at study. Experiment 4 served to investigate the processes underlying the use of double-meaning cues, given the evidence from Experiments 1-3 that double-meaning cues lead to at least equivalent performance compared to single-meaning cues only once cue under-specificity has been resolved. More specifically, Experiment 4 investigated the extent to which double-meaning cues were routinely used in a synergistic way rather than used independently.

Experiment 4

Experiment 4 served to elucidate the processes behind the use of double-meaning cues. Originally, based on the principle of congruity, I reasoned that double-meaning cues would provide two retrieval routes to the target, thus greatly improving the chances of matching meanings at study and test. However, participants' self reported test strategies from Experiments 1 and 2 suggested that double-meaning cues may assist retrieval in ways beyond simply providing two independent routes. Participants often reported trying to use the two cue words together. So, participants may also be able to use double-meaning cues to constrain their generate/search processes to home in on the target better than could be done with single-meaning cues. Thus double-meaning cues may be able to augment performance not only by providing multiple retrieval routes, but also by affording *cue interactivity*.

We can conceive of several different retrieval strategies in which double-meaning cues could be used in an interactive way that would yield the higher performance hinted at in Experiment 3. For example, the two cues could be combined into a single composite query to memory which would effectively search the intersection of the two meanings, with both cues constraining each others' interpreted meanings. Such a class of strategies would encompass numerous self-reports from Experiments 1 and 2, such as: "…I would look at the cue words seperatly and then together and see how they were connected (and if that connection was something I remember seeing).". Using the relationship between two cues as a query to memory is likely more effective for single-meaning than double-meaning cue pairs. But the attempt to figure out a relationship between the two cue words might in and of itself elicit activation of the target so quickly

as to defy conscious awareness of the process. For example, a participant in Experiment 1 wrote: “I first pictured the cue words in my head and tried to make an association between the two words and the words that I had studied earlier. The majority of words that I knew popped into my head immediately after reading the cue words.”.

Another type of retrieval strategy that exploits cue interactivity would be to use one cue to generate candidate responses, then check to see if any of those candidates could be related to the second cue word. For example, a participant in Experiment 2 wrote: “I thought of a word that relates to the first word and tried to match that word with the second one. If they did not match, I tried the same strategy with another word.”. Alternatively, the two cues could be used to filter candidate responses generated from free recall. For example, a participant in Experiment 1 wrote: “I tried to compare them to any of the target words I actually still remembered.”.

The purpose of Experiment 4 was to investigate the extent to which participants’ generate/search process in these tasks employed any such joint use of double-meaning cues. Toward this end, only double-meaning cues were used, and a new within-subjects manipulation was introduced: on the test, a pair of cues was either presented simultaneously in one trial, or sequentially across two trials (i.e., one cue per trial). I reasoned that the simultaneous condition should afford the synergistic use of cues together, while the sequential condition should preclude it or at least make it very impractical.

To the extent that participants use cues interactively, conditions that discourage such strategies should decrease performance when interactive cues are effective and increase performance when they are ineffective. I reasoned that the effectiveness of cue

interactivity would depend on the level of cue ambiguity (i.e., under-specificity), which in this experiment was manipulated between-subjects by presenting no cues at study (as in Experiment 1) versus all four cues at study (as in Experiment 3).

When only targets are presented at study, thus making cue words ambiguous, the interactive use of cues should put the simultaneous test condition at a disadvantage compared to the sequential test condition. This is because any attempt to, say, combine or superimpose two under-specified cues would often yield only confusion and would be less effective and yield more omissions than using one cue at a time. For example, without knowing the intended meaning of cue words, participants may end up trying to find the intersection of semantic areas that don't in fact intersect on a target word. Thus, the effectiveness of the two cues presented together should be *sub-additive* with respect to the effectiveness of both cues presented apart.

When targets are presented with all four cues at study, thus alleviating cue ambiguity, the interactive use of cues should put the simultaneous condition at an advantage compared to the sequential condition. This is because, for example, attempts at determining an intersection of the two cue meanings would now be more likely to succeed, and thus to yield a small set of candidate responses very likely to contain the target. The sequential condition precludes such an effective search. Thus, the effectiveness of the two cues presented together should be *super-additive* with respect to the effectiveness of both cues presented apart.

To summarize, the interactive use of cues predicts an interaction between study condition (0 vs. 4 cues) and test condition (simultaneous vs. sequential) such that simultaneous test cuing should be inferior to sequential cuing when no cues were studied,

and superior to it when four cues were studied. If participants are *not* using cues interactively (e.g., if they are only attempting multiple retrieval routes independently), then there should be no interaction because separating the cues across two trials would neither hinder nor help non-interactive strategies, whether cues are under-specified or not.

Method

Participants and materials. Participants were 44 undergraduates who received partial course credit. Twenty-eight were female, 29 reported that English was not their first language, and their mean age was 19.5 years (SD = 1.3). Materials were the same as in Experiment 1 (i.e., 60 balanced English homographs, each having 4 cue words, 2 per meaning).

Design. The experiment used a 2 x 2 mixed design, with study condition manipulated between-subjects (study targets only vs. study targets with four cues), and test trial condition manipulated within-subjects (cues presented simultaneously in one trial vs. sequentially across two trials). Only double-meaning cue pairs were used in this experiment. The dependent measure was cued recall performance. Responses to a study and test strategy questionnaire (described in the Procedure) were also collected.

Procedure. Participants were randomly assigned to study target words only (study procedure as in Experiment 1) or study target words accompanied by 4 cue words (study procedure as in Experiment 3). Study was followed by the 5 minute Bejeweled distractor task. On the self-paced cued recall test, participants were given only double-meaning cue pairs, and for each target the two randomly selected cues were presented either simultaneously in one test trial or sequentially across two test trials (i.e., one cue per trial). For each participant, half of the targets were randomly assigned to the

simultaneous condition and the other half assigned to the sequential condition. Trial order was randomized for each participant, with the constraint that pairs of sequential trials were separated by 0-4 intervening trials (mean lag = 2). The instructions preceding the test informed participants that the same target word might be the answer on more than one trial.

After completing the test, participants completed a paper questionnaire in which they rated how much they had used a list of study strategies and a list of test strategies. The strategy lists are presented in Appendices C and D; they were constructed based on open-ended self reports collected in the previous experiments, and on prior research (Finley & Benjamin, in press). Participants rated each strategy on a scale of 1 to 7, where 1 was labeled *Didn't use* and 7 was labeled *Used extensively*. Participants could write in and rate any additional strategies they used that were not listed. Participants also answered two yes/no questions asking whether they had ever noticed that target words had multiple meanings during the study phase, and during the test phase.

Results

Recall. In the case of the sequentially tested targets, recall was counted as accurate if the participant typed the correct answer on either or both of the two test trials. Figure 2 shows recall performance as a function of study and test conditions. For participants who studied the target words only, recall performance was reliably higher for the sequential condition (M = .15, SD = .12) versus the simultaneous condition (M = .08, SD = .06), $t(21) = 3.56$, p = .002, d = 0.77. For participants who studied the target words with four cues each, recall performance was numerically higher for the simultaneous condition (M = .43, SD = .19) versus the sequential condition (M = .38, SD = .17),

although this difference did not reach statistical significance, $t(21) = 1.75$, $p = .095$, $d = 0.22$. However, the predicted interaction was in fact reliable, $t(42) = 3.64$, $p < .001$, $d = 1.12$.

Additivity. Next, I analyzed the additivity of the cues in the simultaneous conditions by comparing actual simultaneous performance to that predicted by a simple additive model assuming independent effectiveness of the sequential cues:

$$p(\text{T|X or T|Y}) = p(\text{T|X}) + p(\text{T|Y}) - p(\text{T|X}) * p(\text{T|Y}) \quad (1)$$

where $p(\text{T|X})$ is the probability of cue X eliciting the target, and $p(\text{T|Y})$ is the probability of cue Y eliciting the target. In the simultaneous presence of both cues, recall is counted as successful if either cue elicits the target (i.e., both cues do not have to elicit the target), hence the use of "or" on the left side of the equation. The additive model assumes that two cues are used independently on a simultaneous test trial.[6] To the extent that participants are in fact using the two cues together interactively, this model should be false.

For each participant, mean actual performance in the simultaneous condition was compared to performance predicted by Equation 1, where $p(\text{T|X})$ was mean performance across first sequential trials and $p(\text{T|Y})$ was mean performance across second sequential trials. When participants had studied targets only, their actual simultaneous performance ($M = .08$, $SD = .06$) was reliably lower than their predicted simultaneous performance ($M = .16$, $SD = .13$), $t(21) = 3.70$, $p = .001$, $d = 0.82$. That is, simultaneous cuing was sub-

[6] Recall that only double-meaning cue pairs were used in this experiment, lending plausibility to the independence assumption as a starting point.

additive. When participants had studied targets with four cues, their actual simultaneous performance (M = .43, SD = .19) did not reliably differ from their predicted simultaneous performance (M = .44, SD = .20), $t(21) = 0.81$, p = .428, d = 0.10. That is, simultaneous cuing was approximately additive. The interaction between additivity and study condition was not quite statistically significant, $t(42) = 1.93$, p = .060, d = 0.60. The additivity of cues in both study conditions is illustrated in Figure 3.

Omissions. If participants were indeed using cues interactively, then the sub-additivity in the study-targets-only condition was likely due to confusion arising from attempts to reconcile two ambiguous cue words. Such confusion would be indicated by the relative proportion of omissions for simultaneous versus sequential cuing. Omissions were trials for which participants typed a question mark in order to proceed. Proportions are used instead of counts because there were twice as many individual trials for the sequential versus the simultaneous conditions. When participants had studied targets only, the proportion of omissions for simultaneous test trials (M = .59, SD = .27) was reliably higher than that for sequential test trials (M = .50, SD = .27), $t(21) = 3.44$, p = .002, d = 0.31. In contrast, when participants had studied targets with four cues the proportion of omissions for simultaneous test trials (M = .42, SD = .22) was reliably *lower* than that for sequential test trials (M = .52, SD = .23), $t(21) = 3.95$, p = .001, d = 0.44. The interaction was statistically significant, $t(42) = 5.24$, $p < .001$, d = 1.62. These results indicate that simultaneous cuing indeed led to more confusion than sequential cuing when cues were under-specified, but not when cues were disambiguated.

Response times. Just as the omission analyses suggested participant confusion stemming from attempted use of cue interactivity strategies with under-specified cues, the

same analyses performed on response times may yield similar evidence. For each participant, median RT for simultaneous trials was compared to the mean of the median RTs for the first and second sequential trials. When participants had studied targets only, RT was reliably longer for simultaneous trials (M = 6.19 s, SD = 3.83) versus sequential trials (M = 3.88 s, SD = 2.10), $t(21) = 3.99$, $p = .001$, $d = 0.75$. When participants had studied targets with four cues, RT was again reliably longer for simultaneous trials (M = 3.16 s, SD = 1.18) versus sequential trials (M = 2.10 s, SD = 0.69), $t(21) = 6.29$, $p < .001$, $d = 1.09$. The difference in RTs was reliably larger in the study-targets-only case, $t(42) = 2.08$, $p = .044$, $d = 0.64$, but this interaction should be interpreted with caution because of potential scaling effects (i.e., there is less room left for difference as RTs become smaller).

Strategy questionnaire. Table 2 shows the mean usage frequency ratings for encoding and retrieval strategies as a function of study condition. The table also shows the number of participants who rated each strategy higher than 1, indicating some degree of usage. Of particular interest are the high ratings across both study conditions for the "Cue Relationship: Simultaneous" retrieval strategy and comparatively low ratings for the "Cue Relationship: Sequential" retrieval strategy. These provide further evidence that participants were indeed attempting to use cues together interactively, and were much better able to do so in the simultaneous versus sequential condition. Of further note are the relatively high ratings for the "Generate-Recognize" retrieval strategy, supporting the use of that strategy as a framework for analysis across experiments. The "Free Recall + Match" retrieval strategy, also rated highly, may itself have included a generate-recognize process, though without search explicitly guided by the cue words. Finally, it

is worth noting the low ratings for the "Multiple Target Meanings" encoding strategy in the study-targets-only condition, which is consistent with the previous work suggesting that participants mostly encode a single meaning for target homographs studied without cues (Winograd & Conn, 1971).

For the participants who studied targets without cues, 16 of 22 reported that they had noticed that target words had multiple meanings during the study phase, and 16 of 22 reported so for the test phase. For the participants who studied targets with four cues, the frequencies were 20 of 22, and 20 of 22. Whether for noticing multiple meanings at study or at test, there was no statistically significant difference in proportion as a function of study condition, $z = 1.56$, $p = .118$.

Discussion

The results of Experiment 4 provide several lines of converging evidence that when faced with multiple retrieval cues participants mostly attempt to use cues interactively. For recall performance there was an interaction between study condition and test condition such that: for participants who had studied targets only, and for whom cue words were thus underspecified, cues presented simultaneously on the test yielded lower performance than did cues presented sequentially; and for participants who had studied targets with four cues, and for whom cue words were thus less ambiguous, simultaneous cues yielded numerically higher performance than did sequential cues. The effect of simultaneously presented cues was sub-additive when participants had studied targets only, and additive when participants had studied targets with four cues. The failure to find the predicted super-additivity in the latter case may have been because the method by which cue under-specificity was resolved (i.e., showing four cues at study)

also potentially introduced encoding variability, which may have enhanced the effectiveness of single-route retrieval for sequential cuing. The interplay between encoding and retrieval strategy will be directly addressed in Experiment 5. Simultaneous cues yielded a higher proportion of omissions than sequential cues when participants had studied targets only, and the opposite was true when participants had studied targets with four cues. Simultaneous cues yielded longer RTs than sequential cues to a greater degree for participants who had studied targets only versus those who had studied targets with four cues. Finally, participants' strategy questionnaire responses further suggested that they were attempting to use the two cues together to assist retrieval when they could.

Prior research on cue additivity. The issue of additivity for multiple retrieval cues has been addressed by numerous prior studies. In psycholinguistics research, the possibility of inhibitory processes between competing meanings of an ambiguous word (Simpson & Kang, 1994) has been investigated using additivity of priming effects from two convergent (i.e., single-meaning) versus divergent (i.e., double-meaning) cues (Balota & Paul, 1996). But there is an important difference between psycholinguistic tasks and memory tasks. When ambiguous words such as homographs occur in the use of language, a single meaning is usually intended (except in the case of puns, which is better left closed). Thus, the listener/reader must rapidly disambiguate the meaning in order to comprehend the message. Once a meaning is chosen, comprehension would be facilitated by *inhibition* of the alternative meanings. In contrast, in order to succeed on the memory tests in the current project, a participant needs only to type in the characters comprising the correct target word. S/he does not have to remember any particular meaning of the word in order to be counted correct, and in fact having encoded multiple

meanings can assist in recovering the word.[7] Therefore, disambiguation with inhibition would not be helpful and in fact may be a detriment, especially in the case of incongruent single-meaning test cues, as we will see in the final two experiments. Thus, it is overall not clear to what extent the psycholinguistic research using multiple meanings can be synthesized with the current research, though this would make a good direction for further study.

One other vein of research concerning additivity for multiple retrieval cues has been aimed at determining the structure of human long-term memory traces (i.e., knowledge representation; cf. Tulving & Watkins, 1975). These studies have broadly contrasted associative or non-configural models (e.g., Anderson & Bower, 1973; Jones, 1976) with Gestalt or configural models based on the idea of holistic processing of perceptual stimuli (e.g., Koffka, 1935; Kohler, 1947; Lockhead, 1972). The latter predict super-additivity of multiple cues while the former predict additivity at most.

In an experiment by Anderson and Bower (1972), participants studied simple sentences of the form subject-verb-object (e.g., "The hippie touched the debutante."), where each sentence shared its object with one other sentence. On a test of object recall, *crossover cues* (consisting of the subject and verb from two sentences) were found to be additive with respect to cues consisting of only a single subject or single verb, whereas cues consisting of the subject and verb from a single sentence were found to be sub-additive. However, Foss and Harwood (1975) found the exact opposite result even though they used very similar methods. The inconsistency of results could be due to the

[7] One may argue that such a task is artificial. But as a counter-argument I submit the naturalistic tasks of recalling passwords and common nouns.

focus on the format of stored information and an under-appreciation of the versatility of human retrieval strategies.

There are some hints that the greater the difference between the way two cues point to the target, the greater their synergy in enhancing retrieval. For example, Rubin and Wallace (1989) found that simultaneous presentation of a rhyme cue and an associative cue yielded remarkable super-additivity in the recall of target words. They interpreted this finding in terms of "multiple constraints limiting the number of responses" (p. 698), which is one way that cue interactivity could be implemented in the generation of candidate responses. In a demonstration that facilitated retrieval doesn't necessarily increase performance, Watson, Balota, and Roediger (2003) found that presenting both semantic and phonological cues (e.g., hound, puppy, log, dot) produced super-additive *false* recall and recognition of critical words that had not in fact been presented (e.g., dog). However, super-additivity has also been demonstrated using more similar cuing methods. For example, McLeod, Williams, and Broadbent (1971) found super-additivity of cue words that were associated to the target but not to each other. Although super-additivity was not found in the study-targets-with-cues condition of Experiment 4 in the current project, the overall picture from prior research is that multiple constraints can enhance retrieval to the extent that they can be made to overlap (Figure 4).

Whereas the previous work has largely concerned inferences about the structure of information stored in memory, my focus in the current study is on the effectiveness of retrieval cue variability in improving memory performance, and in how precisely multiple-meaning cues are being used together. I have used additivity as one way to infer the nature of retrieval strategies and processes. In the context of the current project,

Experiment 4 has provided evidence that participants, where permitted by the methods, are largely using cues interactively, which is effective when the cues are unambiguous and ineffective when they are not.

Addressing a problem with the materials. There is one problem with Experiment 4. When two cues were presented simultaneously, each test trial uniquely identified a single target, as was the case in Experiments 1-3. By separating the cues into sequential trials, this was no longer the case. That is, some single cue words by themselves were associated with multiple target words. Thus, sequential performance may have been artificially driven down, particularly in the study-targets-only condition. But if anything this should have worked *against* the pattern observed in the results, that performance was higher for sequential versus simultaneous when participants had studied targets only.

Nevertheless, I re-conducted all of the analyses using the subset of data in which no problematic cues appeared in any sequential test trials. That is, for each participant, data were excluded on a per-target basis whenever a problematic cue occurred in either sequential test trial. This left a mean of 15 targets in the sequential condition across participants (SD = 2.6).

The primary interaction for recall performance remained intact, $t(42) = 2.08$, $p = .043$, $d = 0.64$, as did the simple main effect in the study-targets-only condition, $t(21) = 2.47$, $p = .022$, $d = 0.19$. Simultaneous cues were again reliably sub-additive in the study-targets-only condition, $t(21) = 3.18$, $p = .004$, $d = 0.77$, and this was no longer close to being reliably different from the additivity in the study-targets-with-cues condition, $t(42) = 0.70$, $p = .489$, $d = 0.22$. The pattern of omissions held, with more omissions for

simultaneous versus sequential trials in the study-targets-only condition, $t(21) = 3.67$, $p = .001$, $d = 0.33$, and the opposite in the study-targets-with-cues condition, $t(21) = 2.47$, $p = .022$, $d = 0.37$. The interaction for omissions was again reliable, $t(42) = 4.11$, $p < .001$, $d = 1.27$. Finally, the pattern of RTs was the same as before, with slower responses on simultaneous versus sequential test trials for the study-targets-only condition, $t(21) = 3.51$, $p = .002$, $d = 0.72$, and for the study-targets-with-cues condition, $t(21) = 4.75$, $p < .001$, $d = 1.01$. The interaction was no longer reliable, $t(42) = 1.37$, $p = .177$, $d = 0.42$.

Considering encoding variability. Experiment 4 provided insight into the nature of the retrieval processes participants used when confronted with double-meaning retrieval cues. Participants often (but not exclusively) appear to use the two cues together interactively to generate candidate responses. But such strategies do not work well when cues are under-specified, as when no cues were present during study. This implies that double-meaning cues go beyond simply providing two routes to retrieval, also providing an opportunity to use those cues in an interactive way. Does this mean that the principle of congruity is in fact not relevant to the current tasks? Not necessarily. Even if a single meaning is combined with or used in conjunction with another meaning for retrieval, that retrieval may well still be facilitated by an overlap with the encoded meaning. However, the congruity principle implies that the effects of retrieval cue variability may depend on which meaning(s) were encoded at study. In fact, it could be the case that variability at encoding obviates any benefits of variability at retrieval. If a single match between encoded target meaning and retrieval cue meaning provides most of the congruity benefit (i.e., additional matches yield diminishing benefits), then that single match could be provided by multiple encoded meanings and a single retrieval cue meaning, or by a single

encoded meaning and multiple retrieval cue meanings. Thus, when a match has already been enabled by encoding variability, the additional matches provided by retrieval cue variability would be unnecessary and redundant. The experiments so far have not provided any way to be sure of how particular items were encoded at study, and thus how the encoded meaning(s) may or may not have overlapped with the retrieval cues. Experiment 5 addressed just this issue.

Experiment 5

In order to better understand any benefits of retrieval cue variability and how those may interact with encoding variability, in Experiment 5 I manipulated which cues were presented during study, so I could be certain about how many (and which) meanings were encoded for a given item. This was accomplished by presenting only two cue words at study, either single-meaning or double-meaning. But that presented a new problem if the number of retrieval cue meanings at test was to again be manipulated as in Experiments 1-3: some test conditions would re-present the exact cue words previously studied, while other test conditions would have to present unstudied cue words. To avoid this confound, I decided to use only new, unstudied cue words at test, for all conditions. The use of different cue words at study and test would also provide a good test of generalizability. That is, do any benefits of retrieval cue diversity depend on an exact match with encoded cues?

Thus, the purpose of Experiment 5 was to investigate the interplay between encoding and retrieval variability by fully crossing single- versus double-meaning cues at encoding and at retrieval, within-subjects, without repeating any previously studied cue words on the test. This way, I could control the number of meanings encoded by participants at study, and analyze the effect of retrieval cue diversity as a function of number of meanings encoded. My predictions, again based on consideration of generate/search and recognize processes as well as the principle of congruity, were as follows. For a single-meaning encoding, performance should be highest for congruent single-meaning retrieval cues, followed by double-meaning retrieval cues, then incongruent single-meaning retrieval cues. For a double-meaning encoding, performance

should be highest for double-meaning retrieval cues, followed by single-meaning retrieval cues.

Method

Participants. Participants were 81 undergraduates who received partial course credit. Forty-seven were female, 23 reported that English was not their first language, and their mean age was 19.3 years (SD = 1.7).

Materials. Materials were adapted from those of Experiments 1-4. They consisted of 45 English homograph target words along with eight associated cue words for each target (four cues for each of two meanings). For example, one target word was *foot* and its eight cue words were *mile, yard, meter, measurement, kick, boot, pedal,* and *sole*. All 45 target words were ones used in Experiments 1-4. Four additional cue words were added for each target word, and one or two of the original cue words had to be changed for five target words. Target words were 3-8 letters long (M = 4.5, SD = 1.2) and their HAL frequency ranged from 1,542 to 552,532. The mean forward associative strength (cue-target) across all meanings and targets was .05 (SD = .05, *range* = .01 - .53). Across target items, there was no reliable difference in associative strength between meanings ($t(40) = 0.73$, $p = .471$). Materials are presented in Appendix B.

Design. The experiment used a 3 x 3 fully factorial within-subjects design with the two independent variables being encoding cues (single-meaning-A vs. single-meaning-B vs. double-meaning), and retrieval cues (single-meaning-A vs. single-meaning-B vs. double-meaning). A and B indicate the two different meanings for a target word. Because the target meanings were roughly balanced, and because the effect of particular meanings was not of interest, the nine within-subjects conditions reduce to a

total of five conditions of interest (illustrated in Table 3): single-single-congruent, single-single-incongruent, single-double, double-single, and double-double. The dependent measure was cued recall performance. Responses to a study and test strategy questionnaire (described in the Procedure) were also collected.

Procedure. Initial instructions informed participants that they would be studying target words on which they would later be tested, and that each target word would also have two cue words that they should study to help them remember the target words on the test. Participants were then presented with, in a randomized order, the target words alongside two of their cue words, for 5 seconds each with a 0.5 second blank screen inter-stimulus interval. For each target, the two cue words were positioned in a vertical column, in a random order, to the left of the target word. All words were shown in black type on a white background. The label "Cue Words" appeared above the cue words, and the label "Target Word" appeared above the target word. Target words were randomly assigned such that two thirds (30) were accompanied by a pair of cues that pointed to a single meaning of the target, and one third (15) were accompanied by a pair of cues that pointed to two meanings of the target. After this initial study phase, participants engaged in the same 5 minute distractor task described in Experiment 1.

Participants were then instructed that they would take a test in which they would be shown two *new* cue words that would be related to one of the target words they had studied. They then completed a self-paced cued recall test, in which they were shown two cue words for each target word, and were instructed to type in the corresponding target word, or to type a question mark if they could not remember the target word. Test order was random, and the cue words were again positioned in a vertical column labeled

"Cue Words", in a random order, to the left of the response field labeled "Target Word". Target words were randomly assigned to test conditions as follows: for targets that had been encoded with a pair of single-meaning cues, one third were tested with a pair of cues that pointed to the same single meaning as used at encoding (10 items, single-single-congruent), one third were tested with a pair of cues that pointed to the different single meaning from that used at encoding (10 items, single-single-incongruent), and one third were tested with a pair of cues that pointed to two meanings (10 items, single-double). For targets that had been encoded with a pair of double-meaning cues, two thirds were tested with a pair of cues that pointed to a single meaning (10 items, double-single), and one third were tested with a pair of cues that pointed to two meanings (5 items, double-double). The particular cues used were randomly selected from those available for a given condition. In all cases, the two cue words given at test had *not* been previously studied.

After completing the test, participants completed a strategy questionnaire similar to the one used in Experiment 4. Complete strategy lists are presented in Appendices E and F.

Results

Recall. Cued recall performance is shown in Figure 5. For items that were studied with single meaning cues, performance for congruent single-meaning retrieval cues (M = .33, SD = .20) was reliably higher than performance for incongruent single-meaning retrieval cues (M = .14, SD = .13), $t(80) = 8.85$, $p < .001$, $d = 1.12$, and was also reliably higher than performance for double-meaning retrieval cues (M = .22, SD = .16), $t(80) = 5.32$, $p < .001$, $d = 0.61$. Furthermore, performance for double-meaning retrieval

cues was reliably higher than performance for incongruent single-meaning retrieval cues, $t(80) = 4.50$, $p < .001$, $d = 0.52$. For items that were studied with double-meaning cues, performance did not reliably differ for single-meaning retrieval cues ($M = .26$, $SD = .17$) versus double-meaning retrieval cues ($M = .25$, $SD = .24$), $t(80) = 0.30$, $p = .763$, $d = 0.04$.

For single-meaning encoded items, participants' median test response times did not reliably differ for double-meaning ($M = 5.30$, $SD = 3.00$) versus single-meaning-incongruent retrieval cues ($M = 5.40$, $SD = 3.22$), $t(80) = 0.41$, $p = .686$, $d = 0.03$. Double-meaning RTs were reliably slower than those for congruent single-meaning ($M = 3.92$, $SD = 1.70$), $t(80) = 4.62$, $p < .001$, $d = 0.57$, as were incongruent single-meaning RTs, $t(80) = 4.66$, $p < .001$, $d = 0.58$. For double-meaning encoded items, RTs were reliably slower for double-meaning ($M = 5.97$, $SD = 3.89$) versus single-meaning retrieval cues ($M = 4.79$, $SD = 2.56$), $t(80) = 2.98$, $p = .004$, $d = 0.36$.

Strategy questionnaire. Table 4 shows the mean usage frequency ratings for encoding and retrieval strategies, and the number of participants who rated each strategy higher than 1, indicating some degree of usage. The highest reported encoding strategy use was for "Cue-target Association", which is consistent with previous findings (Finley & Benjamin, in press; Hall, Grossman, & Elwood, 1976). Similar to the results of Experiment 4, predominant retrieval strategies were "Generate-Recognize", "Inter-cue Association", "Direct Search: Both Cues", and "Free Recall + Match".

Out of all 81 participants, 71 reported that they had noticed that target words had multiple meanings during the study phase (8 reported that they had not, and 2 did not respond). For the test phase, 60 reported yes, 20 reported no, and 1 did not respond.

Discussion

When participants had studied a single meaning of a target, a pair of test cues pointing toward that same meaning (congruent) was the best, a pair of test cues pointing toward the different meaning (incongruent) was the worst, and a pair of test cues pointing toward both meanings was somewhere in between. It appears then that when a single meaning has been encoded, retrieval cue variability can serve as a hedge against the worst-case scenario of retrieval cues based on the unencoded meaning. However, the pattern is more puzzling when participants had studied both meanings of a target. In that case, there appeared to be no benefit of receiving double-meaning test cues. It could indeed be the case that encoding variability obviates any benefits of retrieval variability. This would also be consistent with the results of Experiment 3, in which participants had studied both meanings and there was no reliable difference for single- versus double-meaning cues at test. But in Experiment 3 there was also a trend hinting that double-meaning cues may have been beneficial. Given this ambiguity, I sought with Experiment 6 to perform a conceptual replication of Experiment 3, while at the same time extending the investigation to participants' metacognitive monitoring and control (Nelson & Narens, 1990, 1994) with regard to retrieval cue diversity.

As mentioned in the introduction, choices about future retrieval cues will become more numerous and important as personal information management becomes more central to our lives. Investigating metacognition in the simple context of the current project represents a first step in understanding typical memory users' wisdom about sending the most helpful memory cues into the future, and whether they can exploit the potentially beneficial effects of retrieval cue diversity.

Experiment 6

The purpose of Experiment 6 was to investigate the extent to which participants' metacognitive choices and judgments would reflect any understanding of the effects of retrieval cue variability. Furthermore, a subset of conditions in Experiment 6 served to replicate Experiment 3. The basic method was that of Experiment 3 with the addition that, during the study phase, participants selected two of the four cues to receive on the test, and then made a judgment of learning (JOL). At test, half of participants' requests were honored and half were ignored.

Research on metacognition (more specifically, metamemory) has investigated several types of choices that memory users can make in guiding their encoding activities (for a review, see Finley, Tullis, & Benjamin, 2010). These include item selection (Kornell & Metcalfe, 2006), study-time allocation (Son & Metcalfe, 2000), scheduling (Benjamin & Bird, 2006; Son, 2004), self-testing (Kornell & Son, 2009), and selection of encoding strategies (Finley & Benjamin, in press). However, there has been very little research on memory users' choices in guiding retrieval activities. Exceptions include work on the control of output specificity (Goldsmith & Koriat, 2008; Koriat & Goldsmith, 1996) and work on termination of search (Harbison, Dougherty, Davelaar, & Fayyad, 2009; Young, 2004). Experiment 6 addressed several questions. Would participants request more double-meaning cue pairs than single-meaning cue pairs? Would their requests accord with the conditions leading to highest performance? Would their predictions reflect an appreciation for the potential value of retrieval cue diversity? How would retrieval cue diversity affect test performance when participants' requests were honored versus ignored? One basic prediction can be made with regard to honoring

versus ignoring participants' retrieval cue requests. Work by Mäntylä (Mäntylä, 1986; Mäntylä & Nilsson, 1983, 1988) has shown that when participants generate their own cues for target words, later cued recall performance is superior when they receive those cues versus other cues. Thus, I predicted that performance would be enhanced in the current experiment to the extent that participants received the cues they had chosen for the test.

Method

Participants and materials. Participants were 33 undergraduates who received partial course credit. Eleven were female, six reported that English was not their first language, and their mean age was 19.4 years (SD = 1.0). Materials were the same as those used in Experiments 1-4.

Design. There were two within-subjects independent variables, each with two levels. The first variable was whether a participant's test cue request for a given target was honored versus ignored. The second variable was nested within the ignored level of the first variable. In cases where a participants' test cue request was ignored, the test cues presented were either single-meaning or double-meaning. The dependent measures were participants' test cue requests, judgments of learning (JOLs), and cued recall performance. Self reports on cue request strategies and test strategies were also collected.

Procedure. Initial instructions informed participants that they would be studying target words on which they would later be tested, and that each target word would also have four cue words that they should study to help them remember the target words on the test. They were also informed that they would choose which two of the four cues they would most like to receive on the test. Participants were then presented with the

target words alongside their four cue words, in a randomized order. For each target, the four cue words were positioned in a vertical column, in a random order, to the left of the target word. All words were initially shown in black type on a white background. The label "Cue Words" appeared above the cue words, and the label "Target Word" appeared above the target word. Instructions that remained at the top of the screen read: "Study the below cue and target words, and choose the 2 cue words that you would most like to receive on the upcoming test to help you remember the target word.". Participants were given unlimited time to select two cues by clicking on them with the mouse cursor. Once a cue was selected, its text color was changed to blue and a rectangle was drawn around it. Participants could unselect cues by clicking on them again, and could not select more than two cues at a time. Once participants had made their selection, they clicked a "Done Choosing" button, at which point a JOL prompt appeared at the bottom of the screen. The cues and target remained visible. The JOL prompt read: "How sure are you that you will remember this target word when presented with your two chosen cue words on the upcoming test?" Participants clicked on a number from 1 to 6, where 1 was labeled *I am sure I WILL NOT remember* and 6 was labeled *I am sure I WILL remember*. Participants then clicked a "Continue" button to proceed to the next item, which appeared after a 0.5 second blank screen inter-stimulus interval. At the end of the initial study phase, participants engaged in the same 5 minute distractor task used in Experiment 1.

Participants then completed a self-paced cued recall test, in which they were shown two of the four cue words for each target word, and were instructed to type in the corresponding target word, or to type a question mark if they could not remember the target word. Test order was random, and the cue words were again positioned in a

vertical column labeled “Cue Words”, in a random order, to the left of the response field labeled “Target Word”. Target words were randomly assigned to test conditions such that the participant’s test cue requests were honored for half of the targets (request-honored condition), and ignored for the other half of the targets. For the request-ignored targets, half were randomly assigned to be tested with a pair of cues that pointed to a single meaning of the target (request-ignored: single-meaning condition), and the other half were assigned to be tested with a pair of cues that pointed to two meanings of the target (request-ignored: double-meaning condition). The particular single meaning used for the former condition was counterbalanced between-subjects. For the latter condition, one cue was randomly selected from each of the target’s two meanings. Note that for the two request-ignored conditions, participants’ requests were truly ignored, so the random assignment and selection of cues could have resulted in 0, 1, or 2 of the participant’s requested cues actually being given on the test. Thus, this was not an honor-dishonor paradigm (cf. Kornell & Metcalfe, 2006). Note also that the two request-ignored conditions served to replicate the test conditions of Experiment 3. In all cases, the two cue words given at test had previously been studied alongside the target word.

After completing the test, participants answered two free response questions which asked them to describe any strategies they had used in selecting cues during the study phase, and any strategies they had used during the test phase.

Results

For time spent on the study phase trials, the mean of participant medians was 7.02 s (*SD* = 1.74) for completing cue selection[8], 2.29 s (*SD* = 2.10) for completing the JOL,

[8] Note that experimenter-controlled study time in Experiment 3 was 8 s per trial.

and 9.81 s (*SD* = 4.77) overall. Participants' median test response times were reliably slower for double-meaning retrieval cues (*M* = 2.32 s, *SD* = 0.55) versus single-meaning retrieval cues (*M* = 2.16 s, *SD* = 0.51), $t(32) = 2.15$, $p = .039$, $d = 0.31$. Test response times were also reliably slower when participants' requests were ignored (*M* = 2.51 s, *SD* = 0.64) versus honored (*M* = 2.12 s, *SD* = 0.58), $t(32) = 3.27$, $p = .003$, $d = 0.65$.

Test cue requests. For each target in the study phase, participants requested two of the four cue words to be given on the test. Requests were classified as single-meaning versus double-meaning. Across participants, the mean proportion of double-meaning requests was .52 (*SD* = .23, *Mdn* = .50, *range* = .07 - .92). Figure 6 shows a histogram of percent double-meaning test cue requests. Interestingly, there was a reliable positive correlation between the number of double-meaning cues participants requested and their overall performance on items for which requests were ignored, $r = .65$, $t(31) = 4.76$, $p < .001$. This illustrates either the benefits of encoding variability (i.e., requesting and presumably rehearsing two meanings enhanced performance), or that participants who requested more double-meaning cues also happened to be those with better overall memory skills (which is why they performed better on the test).

Judgments of learning. Participants made JOLs based on the two cues they requested; thus, for analysis of JOL accuracy I only considered data for items in the request-honored condition. Across participants the mean gamma correlation between JOLs and cued recall accuracy was .05 (*SD* = .54), $t(27) = 0.48$, $p = .632$. That is, participants did no better than chance at predicting which targets they were more or less likely to remember. Across all items, JOLs were reliably higher for single-meaning requests (*M* = 4.27, *SD* = 0.71) versus double-meaning requests (*M* = 4.12, *SD* = 0.10),

$t(32) = 2.45$, $p = .020$, $d = 0.21$. This is the exact opposite of the pattern shown in actual recall performance, reported below. Thus, it appears that participants' metacognitive monitoring was not accurate for this task.

Recall. First I analyzed the recall data as in Experiments 1-3. Figure 7 shows mean cued recall performance as a function of whether participants' test cue requests were ignored versus honored, and as a function of single- versus double-meaning. When requests were ignored, performance was higher for double-meaning cues ($M = .82$, $SD = .16$) versus single-meaning cues ($M = .72$, $SD = .18$), $t(32) = 5.48$, $p < .001$, $d = 0.64$. This replicated the pattern found in Experiment 3, but with much higher overall performance, $t(81) = 7.98$, $p < .001$, $d = 1.77$. The higher overall performance was likely due to more elaborative processing (Craik & Tulving, 1975) invited by the cue selection and JOL tasks. Furthermore, the effect of double- versus single-meaning retrieval cues was larger in Experiment 6 versus Experiment 3, $t(81) = 3.25$, $p = .001$, $d = 0.74$. The greater amount of active processing of cues in Experiment 6 may have served to better disambiguate the cue meanings.

When requests were honored, performance was again higher for double-meaning cues ($M = .93$, $SD = .14$) versus single-meaning cues ($M = .83$, $SD = .15$), $t(32) = 4.51$, $p < .001$, $d = 0.65$. Collapsing across single- versus double-meaning, performance was higher when requests were honored ($M = .87$, $SD = .14$) versus ignored ($M = 77$, $SD = .16$), $t(32) = 6.09$, $p < .001$, $d = 0.68$. Note also that receiving double-meaning retrieval cues appeared to mitigate the harmful effect of having one's request ignored.

Figure 8 shows an additional way of considering these data: mean cued recall performance as a function of requested test cues (single- vs. double-meaning), received

test cues (single- vs. double-meaning), and request treatment (honored vs. ignored). Pairwise comparisons were made within request type. For single-meaning test cue requests, performance was reliably higher for single-meaning honored (M = .83, SD = .15) versus single-meaning ignored (M = .70, SD = .25), $t(31) = 3.22$, $p = .003$, $d = 0.65$. Double-meaning performance (M = .78, SD = .26) did not reliably differ from single-meaning honored, $t(32) = 1.21$, $p = .234$, $d = 0.26$, nor from single-meaning ignored, $t(31) = 1.14$, $p = .262$, $d = 0.28$. For double-meaning test cue requests, performance for double-meaning honored (M = .93, SD = .14) was reliably higher than performance for both single-meaning (M = .79, SD = .18), $t(32) = 4.50$, $p < .001$, $d = 0.84$, and double-meaning ignored (M = .80, SD = .22), $t(32) = 4.30$, $p < .001$, $d = 0.65$. Performance did not reliably differ for single-meaning versus double-meaning ignored, $t(32) = 0.38$, $p = .706$, $d = 0.07$. Thus, when participants requested a single-meaning, the most effective retrieval cues were those that pointed to that same meaning (congruent), followed by cues that pointed to both meanings, then cues that pointed to the different meaning (incongruent). When participants requested a double-meaning, the most effective retrieval cues were again the exact ones they chose, while the alternative double-meaning cues did not provide the same benefit and were in fact no more effective than single-meaning cues.

When targets were randomly assigned to the request-ignored conditions, test cues were randomly selected under the constraint that they pointed to one meaning or two, depending on the particular condition. This means that, for the request-ignored targets, participants could have received anywhere from zero to two of their requested cues. For the request-honored items, participants always received both of their requested cues.

Figure 9 shows mean cued recall performance as a function of the number of requested test cues actually received, and as a function of request type (single- vs. double-meaning), only including data from participants who happened to receive items in all six cells (n = 25). Separate simple linear regressions for each participant showed that, collapsing across request type, performance reliably increased with the number of requested test cues actually received, M_b = .10, SD_b = .12, $t(24) = 4.13$, $p < .001$. Furthermore, the slope was reliably steeper for single-meaning requests (M_b = .13, SD_b = .18) versus double-meaning requests (M_b = .03, SD_b = .11), $t(24) = 2.58$, $p = .017$, $d = 0.66$. These results suggest that: (a) receiving one's own chosen test cues enhanced memory, consistent with work by Mäntylä (1986; Mäntylä & Nilsson, 1983, 1988); and (b) violation of test cue choices was less detrimental for double-meaning requests.

Self reports. Participants' self reports on test strategies showed a range of responses similar to that found in Experiment 1 (including "no strategy"), with the addition of strategies that exploited previously requested cues, and ones that reinstated strategies used at study. Participants' self-reports about how they made their retrieval cue requests suggested that sophisticated thinking often went into their choices. Twelve participants reported using multiple strategies in making their choices, sometimes depending on the particular stimuli. For example: "I tried to pick two unrelated words so that I would be able to think of a common word between them. If I thought that it would be too difficult for me to recall the word using that strategy, I would choose two closely related words. For example; Tennis and Basketball for the word 'court'.". The modal reported strategy was to choose double-meaning cues (i.e., two cues indicating different meanings of the target; n = 18). Furthermore, there were two types of reasoning behind

this choice: two paths to retrieval (e.g., "I chose two cues that had two different meanings in hope that by looking at the two cues I would have multiple ways of thinking of the one word."), and constrained overlap of meaning (e.g., "…I tried to pick cues that could only be tied together by one word"). Note that these thoughtful pre-arranged retrieval strategies nicely illustrate non-interactive and interactive use of double-meaning cues, respectively, as discussed in Experiment 4. Overall, it appears that over half of participants thought that double-meaning cue pairs would be valuable when making their requests. This is in contrast to the JOL results in which they predicted greater performance for single-meaning cue pairs.[9] These results then apparently demonstrate an interesting dissociation between metacognitive monitoring and control.

Discussion

Confirming the pattern hinted at by Experiment 3, cue diversity at retrieval was beneficial for recall overall, whether participants' test cue requests were honored or ignored (Figure 7). At first this may seem to be discrepant with the results of Experiment 5, in which double-meaning retrieval cues did not provide any benefit when a double-meaning had been encoded. But recall that Experiment 5 used new cues at test. If we assume that participants in Experiment 6 focused their encoding efforts (e.g., elaboration, rehearsal) on the two cues that they requested, then the condition most analogous to double-double in Experiment 5 is the one in which participants requested a pair of double-meaning cues, but received the alternative pair of double-meaning cues (the leftmost white square in Figure 9). In that case, receiving double-meaning cues was not reliably different from receiving single-meaning cues (the middle white square in Figure

[9] The JOL pattern was the same regardless of whether participants did or did not self report double-meaning cue choices.

9), which is in accord with the results of Experiment 5. When participants requested a pair of double-meaning cues and received those exact cues (the rightmost white square in Figure 9), only then were the double-meaning cues more helpful than single-meaning cues. This was the condition most analogous to Experiment 3. Thus, considering Experiments 3, 5, and 6 together, when participants requested or studied both meanings of a target, there appeared to be no benefit for receiving double-meaning test cues, except in the case where those cues were identical to the ones requested/studied. This again suggests that retrieval cue variability may primarily be beneficial in the absence of encoding variability, because one type of variability is redundant with the other. That is, variability at either time point appears to be sufficient to invoke the benefits of congruity.

For cases in which participants requested single-meaning cues, Experiment 6 replicated the pattern of results found in Experiment 5. The benefits of double-meaning cues fell somewhere between the worst-case scenario of incongruent single-meaning, and the best-case scenario of congruent single-meaning.

When it came to selecting a pair of single- versus double-meaning cues, participants showed a sizable range of behavior (Figure 6). That is, they took advantage of retrieval cue variability to varying extents. Regardless, there may have been some merit to their choices, as performance was higher when participants received the test cues they had requested, though requesting double-meaning cues mitigated the cost of one's requests being ignored. However, participants incorrectly gave higher JOLs to items for which they made single-meaning requests, suggesting poor metacognitive monitoring even while metacognitive control was effective.

General Discussion

The current project has begun investigation of the largely unexplored memory effects of retrieval variability. Six experiments tested participants' recall of balanced homographs when cued with a single meaning or with two meanings. Based on the principle of congruity between encoding and retrieval (e.g., transfer-appropriate processing), I predicted that double-meaning cues would be superior by virtue of providing two routes to retrieval, at least one of which would likely overlap with an encoded single meaning. However, single-meaning cues were in fact superior when target homographs had been studied alone (Experiment 1). The generate-recognize theory of recall provided a framework for determining the reason for this unexpected disadvantage for double-meaning cues. The results of Experiment 2 ruled out differences in a recognition process as a sole cause, because performance was still superior for single-meaning cues even when no recognition took place (because no targets were studied).

Experiments 3 and 6 (the request-ignored conditions) served to test the idea that cue under-specificity differentially impairs the generate/search process for the double-meaning cues. Indeed, when the cue words were disambiguated by being presented with the targets during study, double-meaning retrieval cues yielded higher recall.

In Experiment 4 I manipulated study condition (cues absent vs. present) and test condition (simultaneous vs. sequential presentation of retrieval cues) in order to seek insight into the retrieval processes participants used when confronted with double-meaning retrieval cues. Participants appeared to often use the two cues together in a synergistic way to better home in on the target, but such strategies did not work well

when cues were under-specified, as when no cues were presented during study. These results suggest that the effects of retrieval cue variability are not simply a matter of multiple routes, but of combining those routes.

In Experiment 5 I manipulated the number of meanings presented at study and at test, and the results demonstrated that retrieval cue variability can yield benefits or costs depending on encoding conditions. Experiment 6 yielded a similar pattern of results. When a single meaning of a target had been encoded, double-meaning retrieval cues were better than incongruent single-meaning cues (i.e., those that cues the alternative meaning), but worse than congruent single-meaning cues. When both meanings of a target had been encoded, double-meaning retrieval cues were generally neither better nor worse than single-meaning cues, only providing a benefit when they were the exact same cues that had been studied/requested. Experiment 6 also showed that participants recalled more when tested with cues they selected from a set at study, but that they did not have good insights on the benefits of retrieval cue diversity.

The results of Experiments 5 and 6 suggest that variability at one time point obviates any benefits of variability at the other time point, because either is sufficient to ensure congruity between encoded and cued meanings. Alternatively, it is also possible that the double-meaning retrievals in these two experiments were vexed by the problem of cue under-specificity. In Experiment 5, completely new cue words were used at test, so it may have been difficult to identify the relevant meaning of a single cue word without the added context of an additional cue word pointing at the same meaning. In Experiment 6, the unrequested double-meaning cues may have been poorer at specifying the relevant semantic areas for search, for example because they had undergone less

rehearsal during study than the requested cues. However, given the consistency of the results from Experiments 5 and 6, it seems unlikely that I have severely underestimated the benefits of retrieval cue variability in these situations.

Overall, the results of these experiments suggest three major conclusions. First, retrieval cue variability is beneficial to the extent that cues are easily related to the target (i.e., cue under-specificity is alleviated) and that there was little variability at encoding. Note, however, that in these experiments encoding variability and retrieval variability occurred on the same dimension (i.e., number of meanings). It is quite plausible that encoding variability and retrieval variability could independently facilitate performance for situations in which they vary on different dimensions (e.g., cues vs. strategies). In fact I will report one analysis testing this idea later in the General Discussion.

Second, when the number of retrieval cues is held constant, there is a tradeoff between cue specificity and cue variability, both of which can facilitate retrieval. Cue specificity (e.g., single-meaning cues) helps to delimit search using only one particular route, while cue variability (e.g., double-meaning cues) provides multiple routes, each of which is less well delimited. The advantages of cue variability are evident when the cue under-specificity that can accompany double-meaning cues is alleviated. Furthermore, these advantages can stem from synergistic use of the two meanings together.

Finally, retrieval cue variability can be useful as a hedge against uncertainty about the past. That is, just as encoding variability helps us guard against incongruity with future retrieval conditions, retrieval variability can help us guard against incongruity with past encoding conditions.

Retrieval Strategy Variability

The current experiments have all focused on retrieval *cue* variability. However, as I reviewed in the introduction, this is but one kind of retrieval variability. Retrieval *strategies* can also be varied in the service of memory performance (e.g., in autobiographical memory, Williams, 1977, and in eyewitness memory, Fisher et al., 1987). The strategy questionnaires used in Experiments 4 and 5 allow us to analyze the relationship between strategy variability and overall performance.

My overall approach to this analysis was to correlate the number of strategies participants reported using (separately at study and at test) with their overall test performance. I combined the data from the three groups of participants (Experiment 5 and both between-subjects conditions of Experiment 4) in order to increase power and generalizability, $N = 125$. This entailed several standardizing procedures, which I will now describe.

For each strategy on the questionnaires, participants gave a rating from 1 to 7, where 1 indicated that they did not use that strategy at all. Thus, I considered any rating over 1 as indicating some usage of that strategy. For each participant, and separately for study strategies and test strategies, I calculated the number of strategies given ratings over 1, and added to that the number of additional strategies written in by the participant, if any.

Because the questionnaires listed slightly different numbers of strategies across the three groups, I converted participants' raw strategy counts into proportions. The denominator used was the appropriate total number of questionnaire strategies plus the maximum number of additional strategies written in by participants in that group

(separately for study and test). For example, on the questionnaire for participants in the study-targets-only condition in Experiment 4, there were nine test strategies listed, and the greatest number of additional strategies written in was one. So if a participant gave a rating over 1 for seven test strategies, and did not write in any additional strategies, that participant's proportion 7/10 for test strategies.

Each participant's overall performance (i.e., proportion correct cued recall) was calculated collapsing across all within-subjects variables. For participants in Experiment 4, performance was calculated using only the subset of data in which no problematic cues appeared in any sequential test trials (i.e., no cues related to more than one target). Finally, because overall performance levels varied across the groups, each participant's overall test performance was standardized with respect to the mean and standard deviation of the appropriate group.

The results of this analysis were as follows. The number of study strategies did not reliably correlate with test performance, $r = .06$, $t(123) = 0.71$, $p = .479$, but it did reliably positively correlate with the number of test strategies, $r = .35$, $t(123) = 4.20$, $p < .001$. Most interestingly, the number of test strategies was reliably positively correlated with test performance, $r = .196$, $t(123) = 2.22$, $p = .028$.

In order to confirm that the relationship between the number of test strategies and test performance was not mediated by number of study strategies, I conducted a three-variable mediation analysis as per MacKinnon, Fairchild, & Fritz (2007), illustrated in Figure 10. Coefficients were estimated using the below three regression equations:

$$Y = \beta_{01} + \tau X + \varepsilon_1 \quad (2)$$

$$Y = \beta_{02} + \tau' X + \beta Z + \varepsilon_2 \quad (3)$$

$$Z = \beta_{03} + \alpha X + \varepsilon_3 \quad (4)$$

where Y is test performance, X is number of test strategies, and Z is number of study strategies. The goal of the analysis was to determine whether the effect of X on Y via Z (i.e., the indirect effect, or mediation) was reliably above zero. The results are shown in Table 5. The indirect effect was estimated using the method of Sobel (1982; the product of α and β), and the standard error of the indirect effect was estimated using the method of Aroian (1944). This effect was not statistically significant, indicating that the number of study strategies did not mediate the effect of number of test strategies on test performance. So the number of test strategies used indeed had a reliable positive direct effect on test performance.

We must refrain from drawing any causal conclusions from these analyses, because strategy use was not randomly assigned. Thus, it could be that retrieval strategy variability improved test performance, and/or that those participants with better overall memory skills happened to also employ more variable retrieval strategies. Nevertheless, these results suggest that when it came to *strategies*, retrieval variability was more important than encoding variability.

Cue Variability and Strategy Variability

I mentioned earlier that encoding variability and retrieval variability along the same dimension (e.g., number of cues/meanings) may be redundant and thus not offer independent benefits to performance. Because encoding always precedes retrieval, this

means that same-dimension retrieval variability is unlikely to improve performance in the face of prior encoding variability. However, variability across different dimensions at encoding versus retrieval may in fact yield independent benefits from both. The strategy questionnaire data from Experiments 4 and 5 allow us to test this idea.

My approach was to correlate overall test performance with, on the one hand, the extent to which multiple meanings were encoded at study (i.e., variability in meanings), and on the other hand, the extent to which multiple retrieval *strategies* were used at test (i.e., variability in strategies). Test performance and variability in retrieval strategies were computed as described several paragraphs ago. Variability in encoded meanings consisted of each participants' usage rating for the appropriate study strategy on the questionnaire ("Separate Cue-target Associations" for Experiment 5, "Multiple Cue-target Associations" for Experiment 4 study-targets-with-cues condition, and "Multiple Target Meanings" for Experiment 4 study-targets-only condition; see Appendices C and E). Correlations and partial correlations are shown in Figure 11. Variability of encoded meanings (*E*) was reliably and positively correlated with test performance (*T*), $r_{TE} = .221$, $t(123) = 2.51$, $p = .013$. This correlation persisted when the effect of variability of retrieval strategies (*R*) was partialed out, $r_{TE.R} = .211$, $t(122) = 2.39$, $p = .018$. As already reported above, variability of retrieval strategies was also reliably and positively correlated with test performance ($r_{TR} = .196$). This correlation persisted when the effect of variability of encoded meanings was partialed out, $r_{TR.E} = .186$, $t(122) = 2.09$, $p = .039$. Importantly, the two forms of variability were not reliably correlated with each other, $r_{ER} = .070$, $t(123) = 0.78$, $p = .435$, making a mediation analysis unnecessary. Thus, in this

case where encoding and retrieval variability occurred along different dimensions, the two were independently associated with improved test performance.

Conclusion

Across six experiments the current project investigated the possible benefits and detriments of retrieval cue variability in episodic memory tasks, the processes underlying such effects, and how those effects interact with encoding conditions. The results tell a story more nuanced than the simple prediction based on the principle of congruity. It appears that retrieval cue variability is beneficial to the extent that cues are unambiguous and that there was little encoding variability (on that same dimension). Retrieval cues that offer multiple routes toward a target can enable strategies that harness cue interactivity to triangulate on the target. But the potential for variability at encoding and/or retrieval may be governed by the nature of the information to be transferred between time points. Multifaceted information, such as knowledge about another person, may afford a wider variety of retrieval routes than the simple stimuli used in the current project. That said, when exact recall of simple information will be required, such as for passwords, we may be wise to craft that information such that retrieval variability will be possible (e.g., multiple meanings, rhymes, personal relatedness). Finally, results from the current project suggest that memory users may be able to exploit retrieval cue variability in recovering information from an uncertain past, in order to guard against the incongruity between encoding and retrieval that may occur with changing interpretations of ambiguous stimuli. This may particularly be useful for cases in which information was encoded incidentally or cases in which a memory user's prior self lacked the resources or foresight to diversify encoding.

Although this project has concerned human memory, the concepts of encoding and retrieval variability have parallels in the use of external memory systems (e.g., computers), which are increasingly integral to human existence. For example, when saving a digital photograph, a user can implement encoding variability by tagging the file's metadata with multiple descriptive terms, thereby increasing the chances that those terms will overlap with whichever terms s/he may happen to use when searching for the photo in the future. Conversely, when trying to find the photo, a user can implement retrieval variability by using multiple search terms (analogous to retrieval cues) to increase the chances of overlap with whatever terms s/he may have used at the time of saving. Furthermore, retrieval variability could be implemented by the external system itself. For example, the system could generate synonyms to user-entered search terms by using the co-occurrence of descriptor tags in a large corpus such as the online photo-sharing service, Flickr (cf. *folksonomy*, Mathes, 2004; Vander Wal, 2007). The magnitude of personal data offloaded onto external memory systems will continue to increase as new technology enables the automated chronicling, or *life-logging*, of many aspects of daily human experience (see Bell & Gemmell, 2009; Finley, Brewer, & Benjamin, 2011). But with this increase, it will also become less practical for humans to manage encoding variability in the external systems. Thus, external retrieval variability may become ever more important. This is surely only one facet of the complex interplay between human memory and external memory. Further research into this interplay will continue to enhance our subjective continuity of existence across time.

Tables

Table 1

Overview of Procedures for All Experiments

		Experiment					
		1	2	3	4	5	6
Encoding Phase	Targets	60 homographs	--	60 homographs	60 homographs	45 homographs	60 homographs
	Cues	--	--	4 (2 per meaning)	0 vs. 4 (between-Ss)	2 (1x- vs. 2x-meaning, within-Ss)	4 (2 per meaning)
	Duration	8 s	--	8 s	8 s	5 s	self-paced
	Additional Tasks	--	--	--	--	--	test cue request, JOL
Retention Phase (Distractor)		5 min. Bejeweled	5 min. Bejeweled	5 min. Bejeweled	5 min. Bejeweled	5 min. Bejeweled	5 min. Bejeweled
Retrieval Phase	Task	cued recall	cued recall	cued recall	cued recall	cued recall	cued recall
	Cues	2 (1x- vs. 2x-meaning, within-Ss)	2 (1x- vs. 2x-meaning, within-Ss)	2, old (1x- vs. 2x-meaning, within-Ss)	2 (simultaneous vs. sequential, all 2x, within-Ss)	2, new (1x-congruent vs. 1x-incongruent vs. 2x, within-Ss)	2, old (1x- vs. 2x-meaning, request honored vs. ignored, within-Ss)
Additional Phase		strategy free-response	strategy free-response	strategy free-response	strategy questionnaire	strategy questionnaire	strategy free-response

Note. 1x-meaning = single-meaning; 2x-meaning = double-meaning; between-Ss = between-subjects; within-Ss = within-subjects; JOL = judgment of learning.

Table 2

Usage Frequency Ratings for Encoding and Retrieval Strategies as a Function of Study Condition in Experiment 4

Strategy	0 Cues at Study		4 Cues at Study	
	n rated >1	*M (SD)*	*n* rated >1	*M (SD)*
Encoding Strategies				
Cue-target Association	--	--	20	3.7 (1.6)
Multiple Cue-target Associations	--	--	22	4.8 (1.5)
Inter-item Association	19	3.7 (1.8)	21	3.5 (1.7)
Target Focus	--	--	21	4.6 (2.0)
Mental Imagery	19	3.6 (1.7)	17	3.3 (1.8)
Rote Rehearsal	20	4.5 (1.7)	20	4.4 (1.9)
Verbalization	15	3.2 (1.7)	13	3.2 (2.3)
Narrative	13	2.8 (1.8)	11	2.4 (2.0)
Personal Significance	16	2.8 (1.7)	13	3.2 (2.3)
Observation	22	5.3 (1.3)	20	4.5 (1.9)
Multiple Target Meanings	15	3.1 (2.0)	--	--
Retrieval Strategies				
Free Recall + Match	22	4.8 (1.3)	21	4.2 (1.8)
Previous Answer + Match	19	3.9 (1.7)	19	3.0 (1.3)
Generate-Recognize	21	4.6 (1.6)	21	4.2 (1.7)
Generate	21	3.6 (1.8)	15	2.3 (1.3)
Serial Cue Use	20	4.0 (1.8)	14	2.3 (1.5)
Cue Relationship: Simultaneous	21	4.9 (1.5)	20	5.6 (1.8)
Cue Relationship: Sequential	13	2.4 (1.6)	17	3.0 (1.7)
Direct Search: One Cue	21	4.4 (1.6)	21	4.2 (1.6)
Direct Search: Both Cues	20	3.8 (1.2)	19	4.6 (2.0)
Recall of Old Cues	--	--	18	4.4 (1.9)

Note. Ratings were on a scale of 1 (*Didn't use*) to 7 (*Used extensively*). $N = 44$ (22 per study condition). Complete strategy descriptions are provided in Appendices C and D.

Table 3

Illustration of Conditions in Experiment 5

		Test Condition (2 cues)		
		Single-Meaning: Congruent	Single-Meaning: Incongruent	Double-Meaning
Study Condition (2 cues)	Single-Meaning			
	Double-Meaning			

Table 4

Usage Frequency Ratings for Encoding and Retrieval Strategies in Experiment 5

Strategy	*n* rated > 1	*M (SD)*
Encoding Strategies		
Cue-target Association	79	5.2 (1.7)
Separate Cue-target Associations	71	3.6 (1.7)
Single Cue Focus	54	3.1 (2.0)
Inter-item Association	50	2.7 (1.8)
Target Focus	77	4.3 (1.7)
Mental Imagery	64	3.6 (2.1)
Rote Rehearsal	74	4.9 (2.0)
Verbalization	55	3.8 (2.4)
Narrative	39	2.4 (1.8)
Personal Significance	51	2.6 (1.8)
Observation	69	3.9 (1.9)
Retrieval Strategies		
Free Recall + Match	78	4.6 (1.7)
Generate-Recognize	79	5.1 (1.5)
Generate	60	2.8 (1.7)
Recall of Old Cues	70	4.0 (1.9)
Inter-cue Association	75	5.1 (1.7)
Serial Cue Use	62	2.8 (1.5)
Direct Search: One Cue	63	3.2 (1.7)
Direct Search: Both Cues	80	4.9 (1.5)

Note. Ratings were on a scale of 1 (*Didn't use*) to 7 (*Used extensively*). $N = 81$. Complete strategy descriptions are provided in Appendices E and F.

Table 5

Estimated Parameters of Three-variable Mediation Model

Effect	Parameter	Estimate	*SE*	*Z*	*p*
	α	0.41	0.10	4.20	< .001
	β	-0.04	0.62	-0.07	0.946
Total Effect	τ	1.47	0.66	2.22	0.027
Direct Effect	τ'	1.49	0.71	2.09	.037
Indirect Effect	αβ	-0.02	0.26	-0.07	0.948

Note. Parameters are defined in Figure 10 and Equations 2-4.

Figures

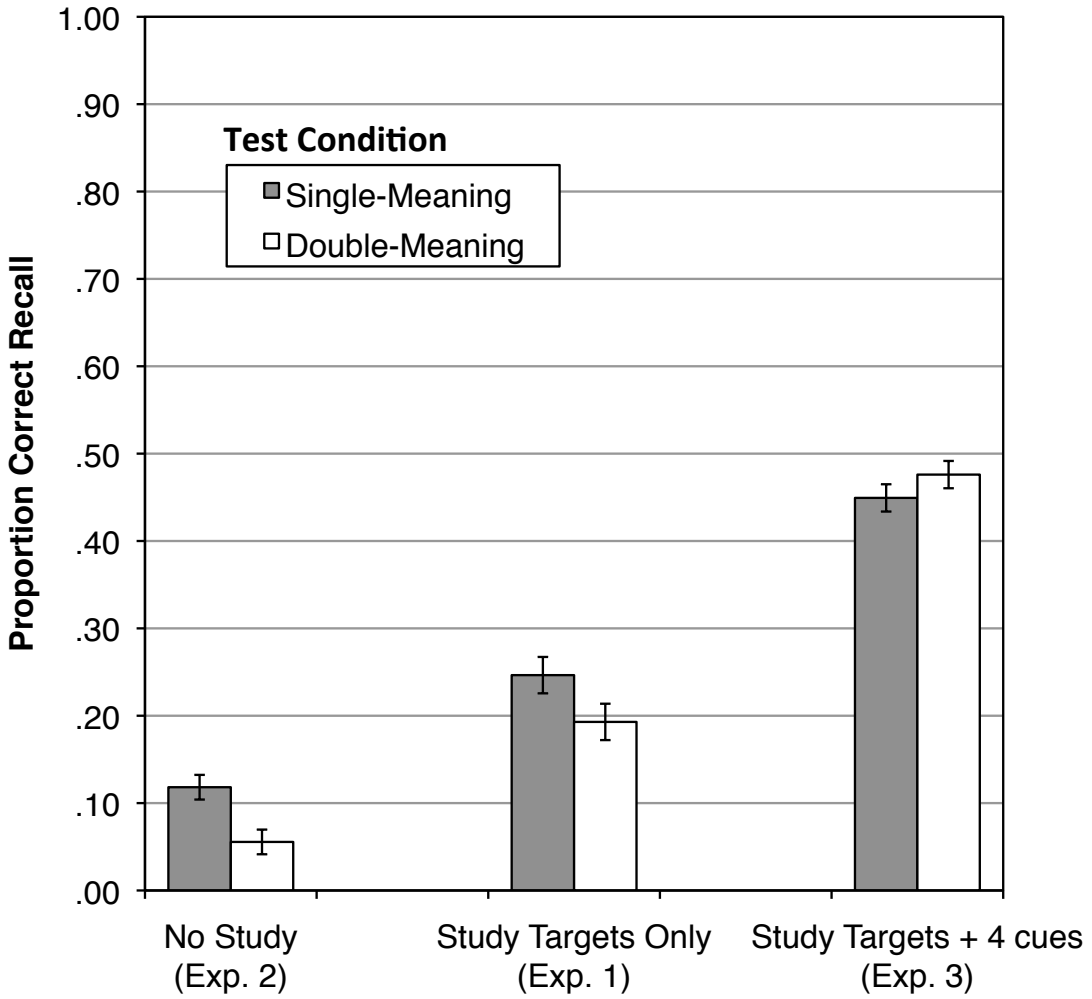

Figure 1. Mean cued recall performance as a function of test condition (single-meaning vs. double-meaning cues) in Experiments 1-3. Error bars represent the standard error of the difference scores.

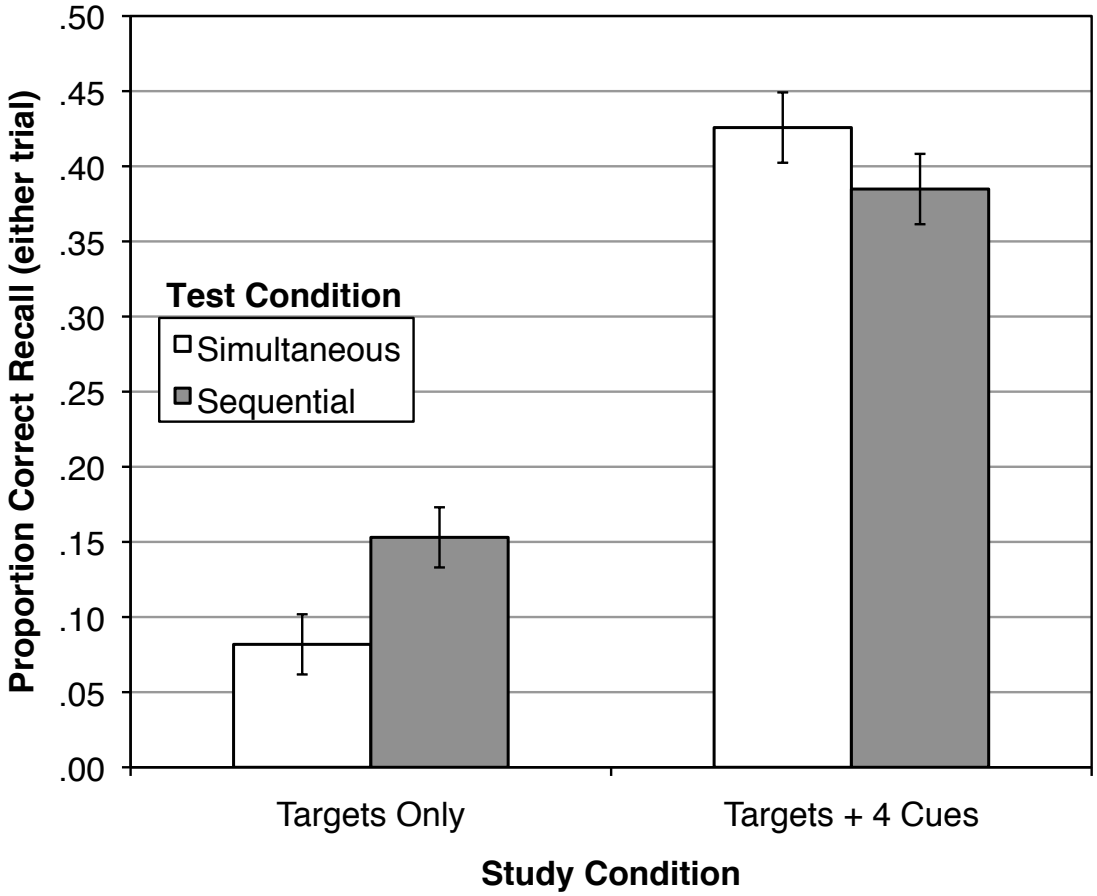

Figure 2. Mean cued recall performance as a function of study condition (0 vs. 4 cues presented with targets) and test condition (simultaneous vs. sequential presentation of retrieval cues) in Experiment 4. Sequentially tested targets were scored as recalled if the correct answer was given on either test trial. Error bars represent the standard error of the difference scores within study condition.

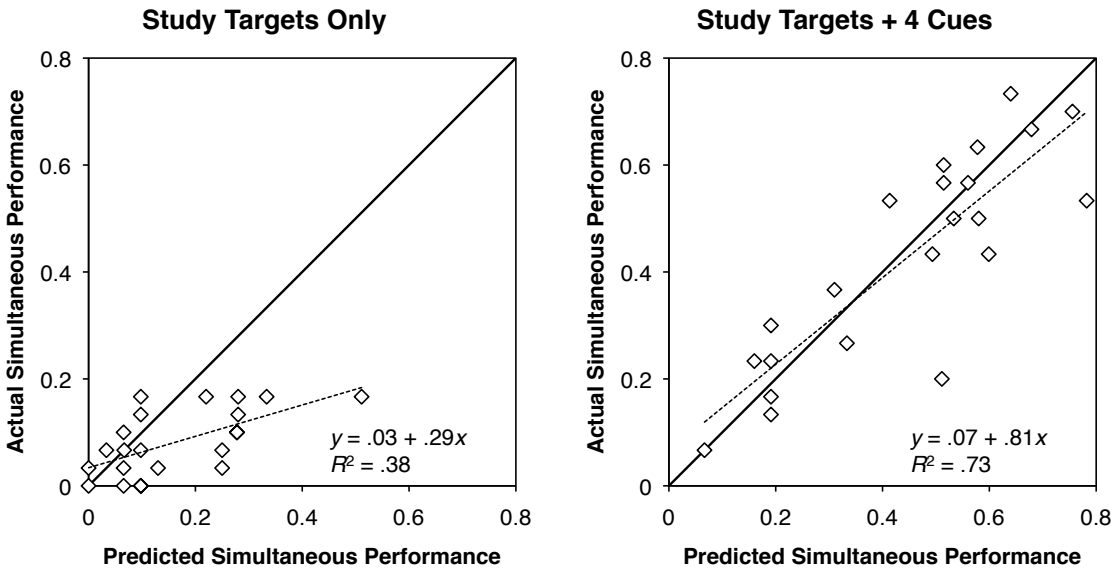

Figure 3. Actual performance on simultaneous test trials versus performance predicted by a simple additive model, separately for both study conditions in Experiment 4. Each point represents one participant, the solid diagonal line represents perfect correspondence, and the dotted line represents the simple linear trend across participants.

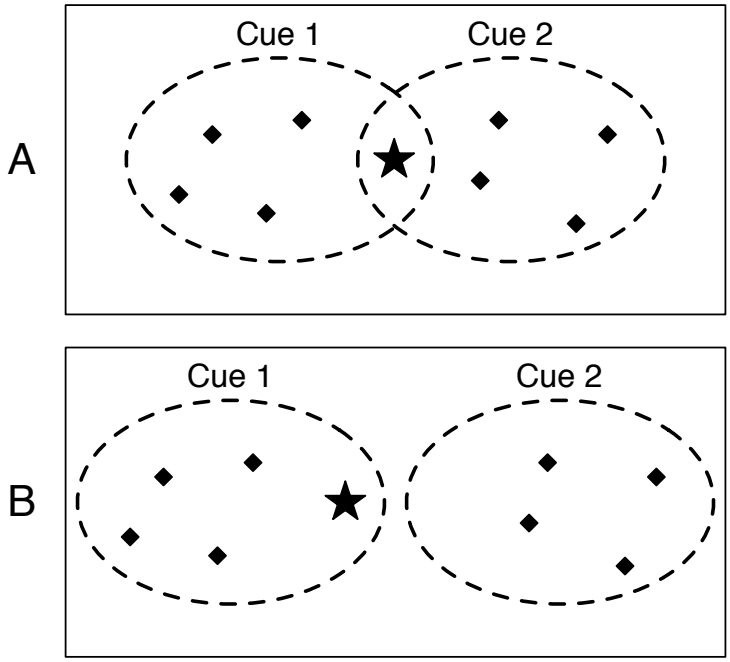

Figure 4. Diagram illustrating conception of a target item (star) stored in long-term memory. Diamonds represent other items. Two cues provide search constraints that overlap (A) or do not overlap (B).

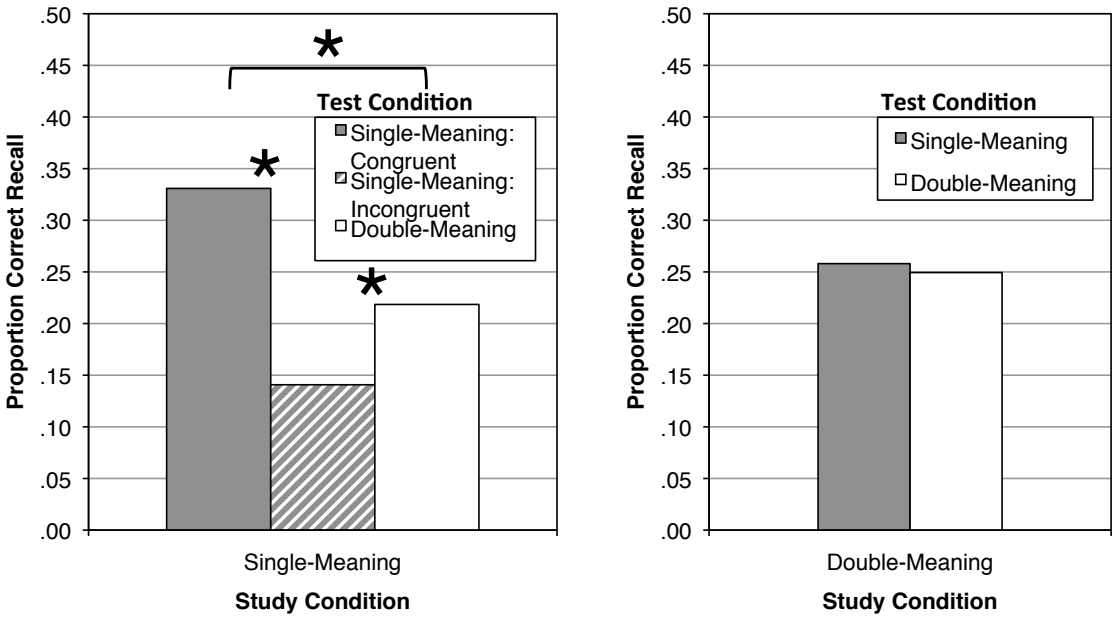

Figure 5. Mean cued recall performance as a function of study condition (single- vs. double-meaning) and test condition (single- vs. double-meaning) in Experiment 5. Asterisks indicate reliable differences ($p < .05$) given study condition.

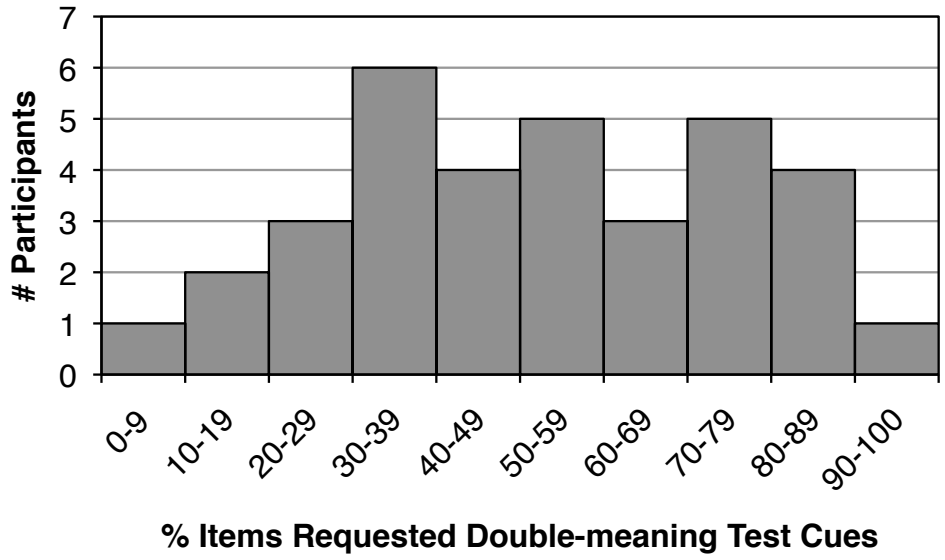

Figure 6. Histogram of percent double-meaning test cue requests in Experiment 6.

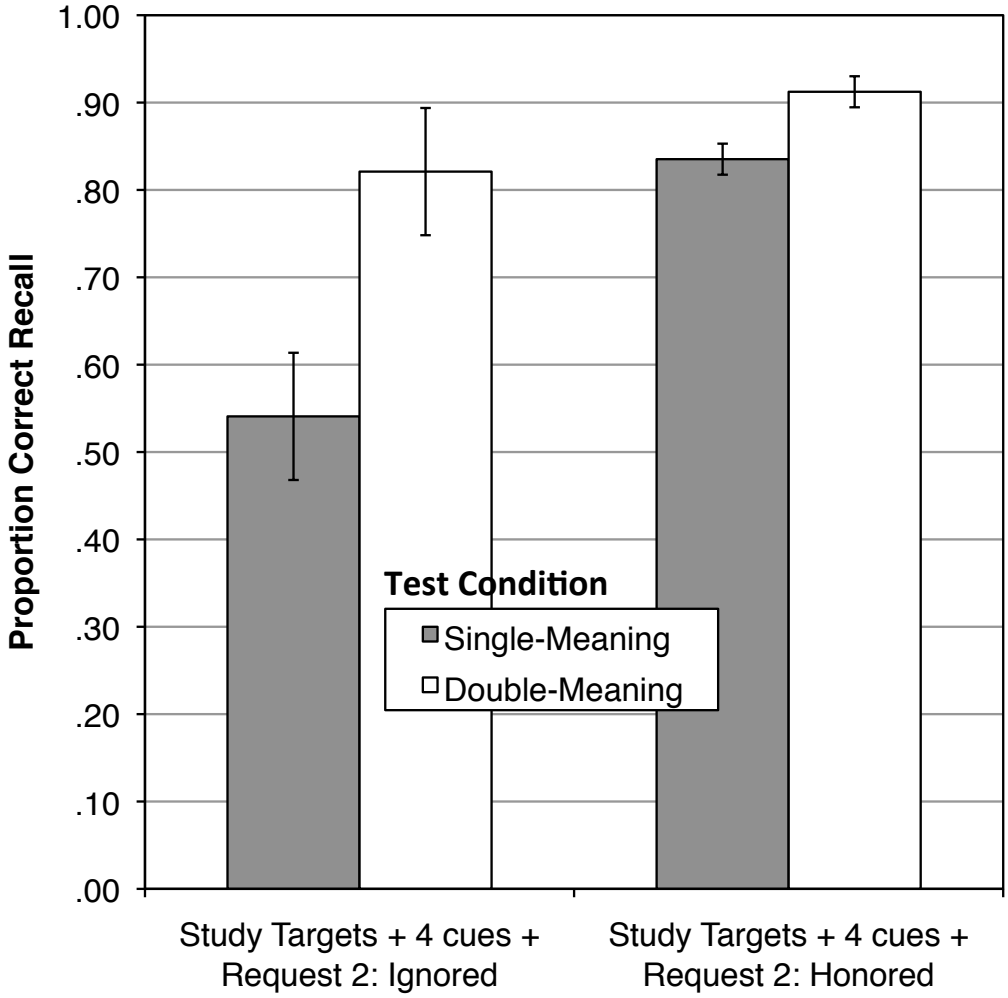

Figure 7. Mean cued recall performance as a function of test condition (single-meaning vs. double-meaning cues) and request treatment (ignored vs. honored) in Experiment 6. Error bars represent the standard error of the difference scores.

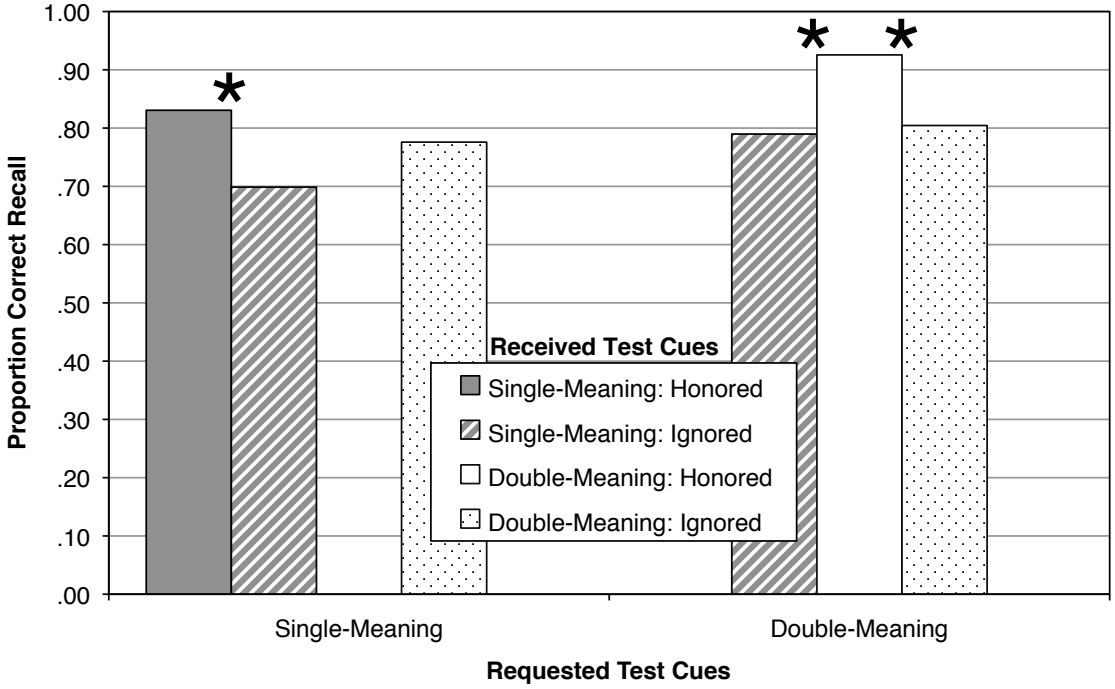

Figure 8. Mean cued recall performance as a function of requested test cues (single- vs. double-meaning), received test cues (single- vs. double-meaning), and request treatment (honored vs. ignored), for Experiment 6. Asterisks indicate reliable differences ($p < .05$) given request type.

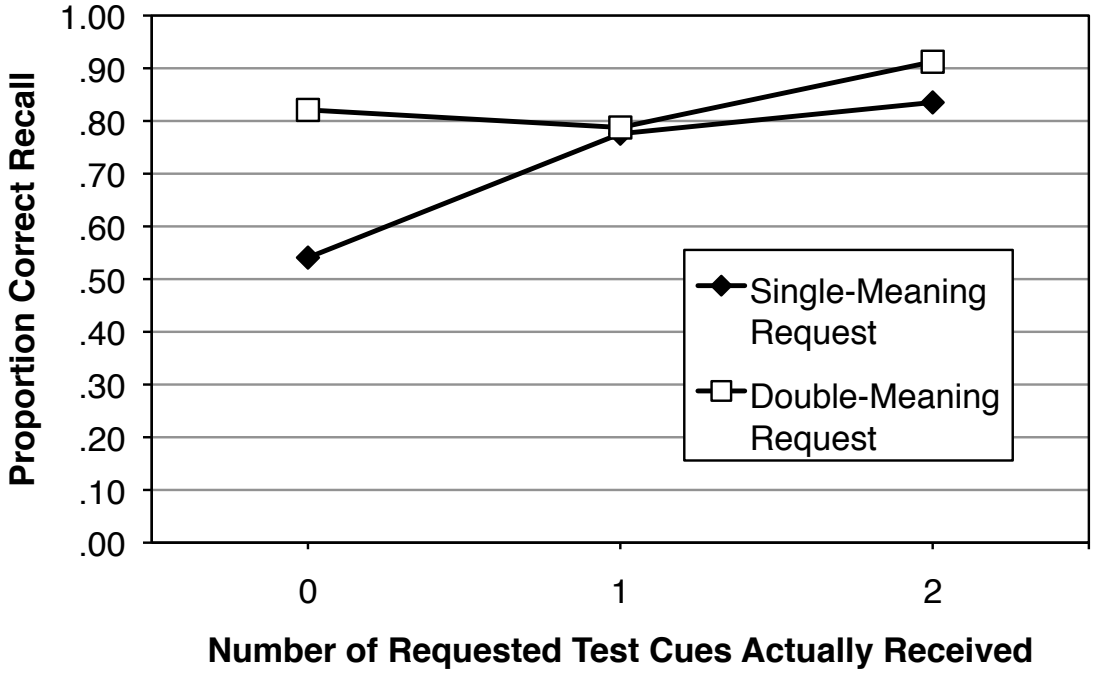

Figure 9. Mean cued recall performance as a function of the number of requested test cues actually received and request type (single- vs. double-meaning), only including data from participants who happened to receive items in all six cells (n = 25), in Experiment 6.

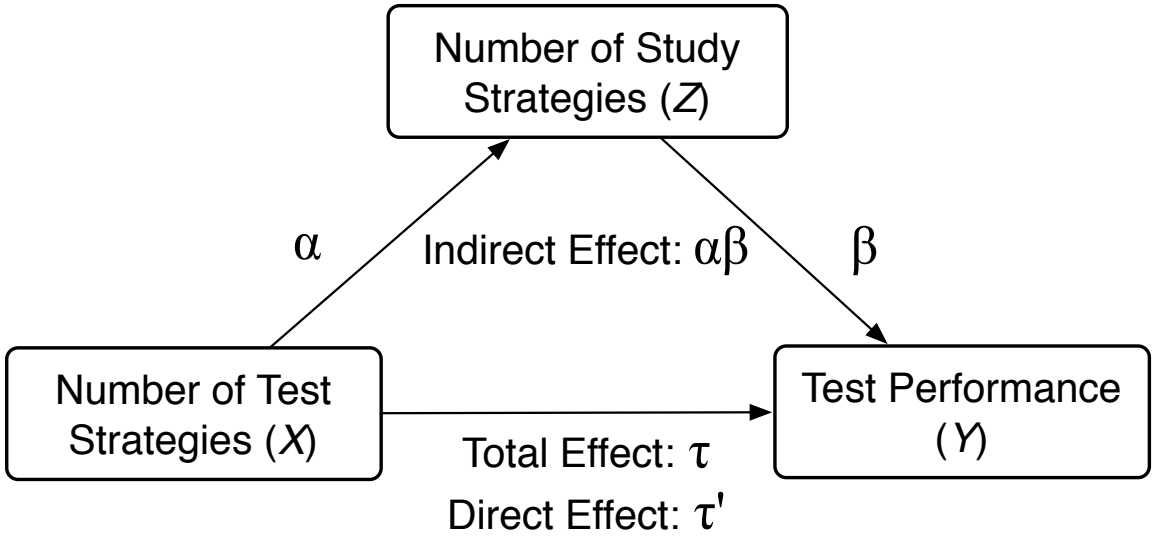

Figure 10. Three variable mediation model.

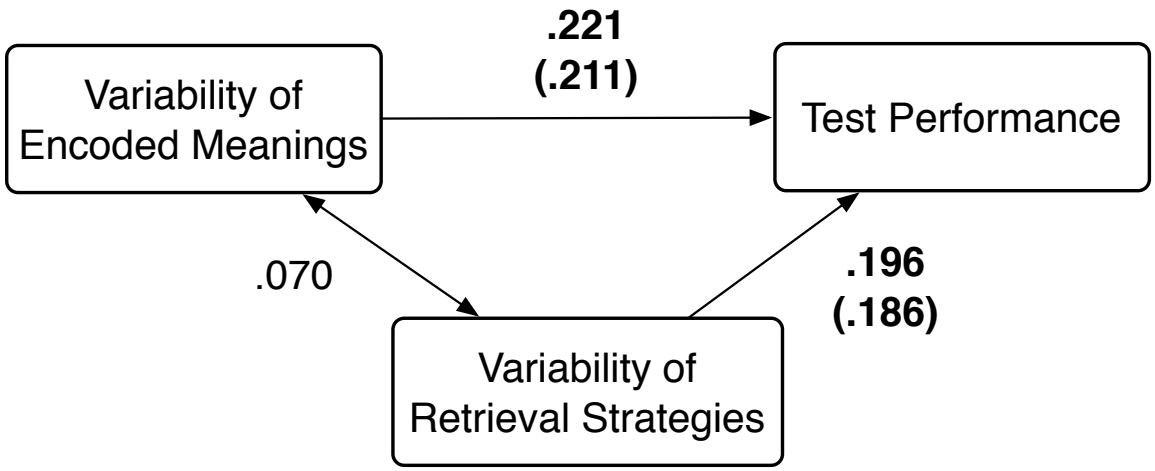

Figure 11. Zero-order and (partial) correlations for variability of encoded meanings, variability of retrieval strategies, and test performance. Data are combined from Experiments 4 and 5, $N = 125$. Values in **bold** reliably differ from zero.

References

Anderson, J. R., & Bower, G. H. (1972). Configural properties in sentence memory. *Journal of Verbal Learning & Verbal Behavior, 11*(5), 594-605. doi:10.1016/S0022-5371(72)80043-4

Anderson, J. R., & Bower, G. H. (1973). *Human associative memory*. Oxford, England: V. H. Winston & Sons.

Anderson, R. C., & Pichert, J. W. (1978). Recall of previously unrecallable information following a shift in perspective. *Journal of Verbal Learning & Verbal Behavior, 17*(1), 1-12. doi:10.1016/S0022-5371(78)90485-1

Aroian, L. A. (1944). The probability function of the product of two normally distributed variables. *Annals of Mathematical Statistics, 18*, 265–271. doi:10.1214/aoms/1177730442

Ballard, D., Hayhoe, M., Pook, P., & Rao, R. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences, 20*, 723-767. doi:10.1017/S0140525X97001611

Balota, D. A., & Paul, S. T. (1996). Summation of activation: Evidence from multiple primes that converge and diverge within semantic memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(4), 827-845. doi:10.1037/0278-7393.22.4.827

Balota, D.A., Yap, M.J., Cortese, M.J., Hutchison, K.A., Kessler, B., Loftis, B., Neely, J.H., Nelson, D.L., Simpson, G.B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods, 39*(3), 445-459.

Bell, G., & Gemmell, J. (2009). *Total recall: How the e-memory revolution will change everything*. New York, NY: Dutton.

Benjamin, A. S., & Bird, R. D. (2006). Metacognitive control of the spacing of study repetitions. *Journal of Memory and Language, 55*, 126–137. doi:10.1016/j.jml.2006.02.003

Benjamin, A. S. & Tullis, J. G. (2010). What makes distributed practice effective? *Cognitive Psychology, 61*(3), 228-247. doi:10.1016/j.cogpsych.2010.05.004

Bilodeau, I. M., & Schlosberg, H. (1951). Similarity in stimulating conditions as a variable in retroactive inhibition. *Journal of Experimental Psychology, 41*(3), 199-204. doi:10.1037/h0056809

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

Bjork, R. A., & Richardson-Klavehn, A. (1989). On the puzzling relationship between environmental context and human memory. In C. Izawa (Ed.), *Current issues in cognitive processes: The Tulane Flowerree Symposium on Cognition* (pp. 313-344). Hillsdale, NJ: Erlbaum.

Bower, G. H. (1972). Stimulus sampling theory of encoding variability. In A. W. Melton and E. Martin (Eds.), *Coding Processes in Human Memory* (pp. 85-123), Washington, DC: Winston.

Brady, F. (2004). Contextual interference: A meta-analytic study. *Perceptual and Motor Skills, 99*(1), 116-126. doi:10.2466/PMS.99.4.116-126

Brady, F. (2008). The contextual interference effect and sport skills. *Perceptual and Motor Skills, 106*(2), 461-472. doi:10.2466/PMS.106.2.461-472

Burnkrant, R. E., & Unnava, H. R. (1987). Effects of variation in message execution on the learning of repeated brand information. In M. Wallendorf and P. Anderson (Eds.) *Advances in Consumer Research* (Vol. 14, pp. 173-176). Provo, UT : Association for Consumer Research.

Catalano, J. F., & Kleiner, B. M. (1984). Distant transfer in coincident timing as a function of variability of practice. *Perceptual and Motor Skills, 58*(3), 851-856. doi: 10.2466/pms.1984.58.3.851

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132*(3), 354-380. doi:10.1037/0033-2909.132.3.354

Chelonis, J. J., Calton, J. L., Hart, J. A., & Schachtman, T. R. (1999). Attenuation of the renewal effect by extinction in multiple contexts. *Learning and Motivation, 30*(1), 1-14. doi:10.1006/lmot.1998.1022

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning & Verbal Behavior, 11*(6), 671-684. doi:10.1016/S0022-5371(72)80001-X

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General, 104*(3), 268-294. doi:10.1037/0096-3445.104.3.268

Dallett, K., & Wilcox, S. G. (1968). Contextual stimuli and proactive inhibition. *Journal of Experimental Psychology, 78*(1), 475-480. doi:10.1037/h0026461

Ebbinghaus, H. (1913). *Memory: A contribution to experimental psychology* (H. A. Rueger & C. E. Bussenius, Trans.). New York: Teachers College. (Original work published in German in 1885.)

Eich, E., & Metcalfe, J. (1989). Mood dependent memory for internal versus external events. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*(3), 443-455. doi:10.1037/0278-7393.15.3.443

Eich, J. E., Weingartner, H., Stillman, R. C., & Gillin, J. C. (1975). State-dependent accessibility of retrieval cues in the retention of a categorized list. *Journal of Verbal Learning & Verbal Behavior, 14*(4), 408-417. doi:10.1016/S0022-5371(75)80020-X

Estes, W. K. (1955). Statistical theory of spontaneous recovery and regression. *Psychological Review, 62*(3), 145-154. doi:10.1037/h0048509

Finley, J. R., & Benjamin, A., S. (in press). Adaptive changes in encoding strategy with experience: Evidence from the test expectancy paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition.* doi:10.1037/a0026215

Finley, J. R., Brewer, W. F., & Benjamin, A. S. (2011). The effects of end-of-day picture review and a sensor-based picture capture procedure on autobiographical memory using SenseCam. *Memory, 19*(7), 796-807. doi:10.1080/09658211.2010.532807

Finley, J. R., Tullis, J. G., & Benjamin, A. S. (2010). Metacognitive control of learning and remembering. In M. S. Khine & I. Saleh (Eds.), *New science of learning: cognition, computers and collaboration in education*. Springer.

Fisher, R. P., & Craik, F. I. (1977). Interaction between encoding and retrieval operations in cued recall. *Journal of Experimental Psychology: Human Learning and Memory, 3*(6), 701-711. doi:10.1037/0278-7393.3.6.701

Fisher, R. P., Geiselman, R. E., Raymond, D.S., Jurkevich, L. M., Warhaftig, M. L. (1987). Enhancing enhanced eyewitness memory: Refining the cognitive interview. *Journal of Police Science and Administration, 15*, 291-297.

Foss, D. J., & Harwood, D. A. (1975). Memory for sentences: Implications for human associative memory. *Journal of Verbal Learning & Verbal Behavior, 14*(1), 1-16. doi:10.1016/S0022-5371(75)80002-8

Gartman, L. M., & Johnson, N. F. (1972). Massed versus distributed repetition of homographs: A test of the differential-encoding hypothesis. *Journal of Verbal Learning & Verbal Behavior, 11*(6), 801-808. doi:10.1016/S0022-5371(72)80016-1

Gilbert, J. A. E., & Fisher, R. P. (2006). The effects of varied retrieval cues on reminiscence in eyewitness memory. *Applied Cognitive Psychology, 20*(6), 723-739. doi:10.1002/acp.1232

Glanzer, M., & Duarte, A. (1971). Repetition between and within languages in free recall. *Journal of Verbal Learning & Verbal Behavior, 10*(6), 625-630. doi:10.1016/S0022-5371(71)80069-5

Glenberg, A. M. (1979). Component-levels theory of the effects of spacing of repetitions on recall and recognition. *Memory & Cognition, 7*(2), 95-112. doi:10.3758/BF03197590

Goldsmith, M. & Koriat, A. (2008). The strategic regulation of memory accuracy and informativeness. In A. Benjamin and B. Ross (Eds.), *Psychology of learning and motivation, Vol. 48: Memory use as skilled cognition* (pp. 1-60). San Diego, CA: Elsevier.

Godden, D. R., & Baddeley, A. D. (1975). Context-dependent memory in two natural environments: Land and underwater. *British Journal of Psychology, 66*, 325-331. doi:10.1111/j.2044-8295.1975.tb01468.x

Goodwin, D.W., Powell, B., Bremer, D., Hoine, H. & Stern, J. (1969). Alcohol and recall: state-dependent effects in man, *Science, 163*, 1358. doi:10.1126/science.163.3873.1358

Greenberg, D. L., & Verfaellie, M. (2010). Effects of fixed- and varied-context repetition on associative recognition in amnesia. *Journal of the International Neuropsychological Society, 16*(4), 596-602. doi:10.1017/S1355617710000287

Hall, J. W., Grossman, L. R., & Elwood, K. D. (1976). Differences in encoding for free recall vs. recognition. *Memory & cognition, 4*(5), 507-513. doi:10.3758/BF03213211

Harbison, J.I., Dougherty, M. R., Davelaar, E., & Fayyad, B. (2009). The lawfulness of decisions to terminate memory search. *Cognition, 111*(3), 397-402. doi:10.1016/j.cognition.2009.03.002.

Herzog, S. M., & Hertwig, R. (2009). The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psychological Science, 20*(2), 231-237. doi:10.1111/j.1467-9280.2009.02271.x

Hintzman, D. L. (1974). Theoretical implications of the spacing effect. In R. L. Solso (Ed.), *Theories in cognitive psychology: The Loyola Symposium* (pp. 77-99). Potomac, MD.: Erlbaum.

Hintzman, D. L., & Stern, L. D. (1978). Contextual variability and memory for frequency. *Journal of Experimental Psychology: Human Learning and Memory, 4*(5), 539-549. doi:10.1037/0278-7393.4.5.539

Hourihan, K. L. & Benjamin, A. S. (2010). Smaller is better (when sampling from the crowd within): Low memory span individuals benefit more from multiple opportunities for estimation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 1068-1074. doi:10.1037/a0019694

Jones, G. V. (1976). A fragmentation hypothesis of memory: Cued recall of pictures and of sequential position. *Journal of Experimental Psychology: General, 105*(3), 277-293. doi:10.1037/0096-3445.105.3.277

Jones, W., & Teevan, J. (eds.) (2007). *Personal information management*. University of Washington Press.

Kerr, R., & Booth, B. (1978). Specific and varied practice of motor skill. *Perceptual and Motor Skills, 46*, 395–401. doi:10.2466/pms.1978.46.2.395

Koffka, K. (1935). *Principles of gestalt psychology*. Oxford, England: Harcourt, Brace.

Kohler, W. (1947). *Gestalt psychology* (2nd ed). Oxford, England: Liveright.

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review, 103*, 490-517. doi:10.1037/0033-295X.103.3.490

Kornell, N., & Metcalfe, J. (2006). Study efficacy and the region of proximal learning framework. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 32*, 609-622. doi:10.1037/0278-7393.32.3.609

Kornell, N., & Son, L. K. (2009). Learners' choices and beliefs about self-testing. *Memory, 17*(5), 493–501. doi:10.1080/09658210902832915

Lee, T. D., Magill, R. A., & Weeks, D. J. (1985). Influence of practice schedule on testing schema theory predictions in adults. *Journal of Motor Behavior, 17*(3), 283-299.

Lockhead, G. R. (1972). Processing dimensional stimuli: A note. *Psychological Review, 79*(5), 410-419. doi:10.1037/h0033129

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers, 28*(2), 203-208. doi:10.3758/BF03204766

MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology, 58*, 593–614. doi:10.1146/annurev.psych.58.110405.085542

Magill, R. A., & Hall, K. G. (1990). A review of the contextual interference effect in motor skill acquisition. *Human Movement Science, 9*, 241-289. doi:10.1016/0167-9457(90)90005-X

Mäntylä, T. (1986). Optimizing cue effectiveness: Recall of 500 and 600 incidentally learned words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12*(1), 66-71. doi:10.1037/0278-7393.12.1.66

Mäntylä, T., & Nilsson, L. (1983). Are my cues better than your cues? uniqueness and reconstruction as prerequisites for optimal recall of verbal materials. *Scandinavian Journal of Psychology, 24*(4), 303-312. doi:10.1111/j.1467-9450.1983.tb00504.x

Mäntylä, T., & Nilsson, L. (1988). Cue distinctiveness and forgetting: Effectiveness of self-generated retrieval cues in delayed recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*(3), 502-509. doi:10.1037/0278-7393.14.3.502

Martin, E. (1968). Stimulus meaningfulness and paired-associate transfer: An encoding variability hypothesis. *Psychological Review, 75*(5), 421-441. doi:10.1037/h0026301

Mathes, A. (2004, December). *Folksonomies: Cooperative classification and communication through shared metadata*. Retrieved from http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html

McDaniel, M.A., & Einstein, G.O. (2007). *Prospective memory: An overview and synthesis of an emerging field.* Thousand Oaks, CA: Sage.

McLeod, P. D., Williams, C. E., & Broadbent, D. E. (1971). Free recall with assistance from one and from two retrieval cues. *British Journal of Psychology, 62*(1), 59-65. doi:10.1111/j.2044-8295.1971.tb02011.x

Melton, A. W. (1970). The situation with respect to the spacing of repetitions and memory. *Journal of Verbal Learning and Verbal Behavior, 9*, 596-606. doi:10.1016/S0022-5371(70)80107-4

Memon, A., Meissner, C. A., & Fraser, J. (2010). The cognitive interview: A meta-analytic review and study space analysis of the past 25 years. *Psychology, Public Policy, and Law, 16*(4), 340-372. doi:10.1037/a0020518

Mensink, G., & Raaijmakers, J. G. (1988). A model for interference and forgetting. *Psychological Review, 95*(4), 434-455. doi:10.1037/0033-295X.95.4.434

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning & Verbal Behavior, 16*(5), 519-533. doi:10.1016/S0022-5371(77)80016-9

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

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 26, pp. 125–141). New York: Academic Press.

Nelson, T. O., & Narens, L. (1994). Why investigate metacognition? In J. Metcalfe, & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 1–25). Cambridge, MA: The MIT Press.

Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology, 25*, 241-286. doi:10.1006/ceps.2000.1040

Ross, B. H., & Landauer, T. K. (1978). Memory for at least one of two items: Test and failure of several theories of spacing effects. *Journal of Verbal Learning & Verbal Behavior, 17*(6), 669-680. doi:10.1016/S0022-5371(78)90403-6

Rubin, D. C., & Wallace, W. T. (1989). Rhyme and reason: Analyses of dual retrieval cues. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*(4), 698-709. doi:10.1037/0278-7393.15.4.698

Shea, C. H., & Kohl, R. M. (1990). Specificity and variability of practice. *Research Quarterly for Exercise and Sport, 61*, 169-177.

Shea, J. B., & Morgan, R. L. (1979). Contextual interference effects on the acquisition, retention, and transfer of a motor skill. *Journal of Experimental Psychology: Human Learning and Memory, 5*(2), 179-187. doi:10.1037/0278-7393.5.2.179

Simpson, G. B., & Kang, H. (1994). Inhibitory processes in the recognition of homograph meanings. In D. Dagenbach, T. H. Carr, D. Dagenbach & T. H. Carr (Eds.), *Inhibitory processes in attention, memory, and language*. (pp. 359-381). San Diego, CA, US: Academic Press.

Smith, S. M. (1984). A comparison of two techniques for reducing context-dependent forgetting. *Memory & Cognition, 12*(5), 477-482. doi:10.3758/BF03198309

Smith, S.M. (1988). Environmental context-dependent memory. In G. Davies and D. Thomson (Eds.) *Memory in context: Context in memory* (pp. 13-33). New York, NY: Wiley.

Smith, S. M. (2007). Context and human memory. In H. L. Roediger, III, Y. Dudai, and S. M. Fitzpatrick (Eds.) *Science of Memory: Concepts*, Oxford University Press, pp. 111-114.

Smith, S.M., Glenberg, A.M., & Bjork, R.A. (1978). Environmental context and human memory. *Memory & Cognition, 6*(4), 342-353. doi:10.3758/BF03197465

Smith, S.M., & Vela, E. (2001). Environmental context-dependent memory: A review and meta-analysis. *Psychonomic Bulletin and Review, 8*, 203-220. doi:10.3758/BF03196157

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. In S. Leinhardt (Ed.), *Sociological methodology* (pp. 290–312).Washington, DC: American Sociological Association.

Son, L. K. (2004). Spacing one's study: Evidence for a metacognitive control strategy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 601–604. doi:10.1037/0278-7393.30.3.601

Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 204–221. doi:10.1037/0278-7393.26.1.204

Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertip. *Science, 333*, 776-778. doi:10.1126/science.1207745

Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations*. New York, NY, US: Doubleday & Co.

Tulving, E. (1974). Recall and recognition of semantically encoded words. *Journal of Experimental Psychology, 102*(5), 778-787. doi:10.1037/h0036383

Tulving, E., & Osler, S. (1968). Effectiveness of retrieval cues in memory for words. *Journal of Experimental Psychology, 77*(4), 593-601. doi:10.1037/h0026069

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review, 80*(5), 352-373. doi:10.1037/h0020071

Tulving, E., & Watkins, M. J. (1975). Structure of memory traces. *Psychological Review, 82*(4), 261-275. doi:10.1037/h0076782

Twilley, L. C., Dixon, P., Taylor, D., & Clark, K. (1994). University of Alberta norms of relative meaning frequency for 566 homographs. *Memory & Cognition, 22*(1), 111-126. doi:10.3758/BF03202766

Vander Wal, T. (2007, February 2). *Folksonomy coinage and definition*. Retrieved from http://vanderwal.net/folksonomy.html

Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science, 19*(7), 645-647. doi:10.1111/j.1467-9280.2008.02136.x

Walker, W. H., & Kintsch, W. (1985). Automatic and strategic aspects of knowledge retrieval. *Cognitive Science: A Multidisciplinary Journal, 9*(2), 261-283. doi:10.1207/s15516709cog0902_3

Wallsten, T. S., Budescu, D. V., Erev, I., & Diederich, A. (1997). Evaluating and combining subjective probability estimates. *Journal of Behavioral Decision Making. Special Issue: Stochastic and Cognitive Models of Confidence, 10*(3), 243-268. doi:10.1002/(SICI)1099-0771(199709)10:3<243::AID-BDM268>3.0.CO;2-M

Watkins, M. J., & Gardiner, J. M. (1979). An appreciation of generate–recognize theory of recall. *Journal of Verbal Learning & Verbal Behavior, 18*(6), 687-704. doi:10.1016/S0022-5371(79)90397-9

Watkins, M. J., & Tulving, E. (1975). Episodic memory: When recognition fails. *Journal of Experimental Psychology: General, 104*(1), 5-29. doi:10.1037/0096-3445.104.1.5

Watson, J. M., Balota, D. A., & Roediger, H. L. (2003). Creating false memories with hybrid lists of semantic and phonological associates: Over-additive false memories produced by converging associative networks. *Journal of Memory and Language, 49*(1), 95-118. doi:10.1016/S0749-596X(03)00019-6

Whitten, W. B., & Leonard, J. M. (1981). Directed search through autobiographical memory. *Memory & Cognition, 9*(6), 566-579. doi:10.3758/BF03202351

Williams, D. M. (1977). *The process of retrieval from very long term memory* (Unpublished doctoral dissertation). University of California, San Diego.

Williams, D. M., & Hollan, J. D. (1981). The process of retrieval from very long term memory. *Cognitive Science: A Multidisciplinary Journal, 5*(2), 87-119. doi:10.1207/s15516709cog0502_1

Williams, M. D., & Santos-Williams, S. (1980). Method for exploring retrieval processes using verbal protocols. In R. S. Nickerson (Ed.), *Attention and Performance VII* (pp. 671 – 689). Hillsdale, NJ: Lawrence Erlbaum.

Winograd, E., & Conn, C. P. (1971). Evidence from recognition memory for specific encoding of unmodified homographs. *Journal of Verbal Learning & Verbal Behavior, 10*(6), 702-706. doi:10.1016/S0022-5371(71)80078-6

Wulf, G., & Schmidt, R. A. (1997). Variability of practice and implicit motor learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*(4), 987-1006. doi:10.1037/0278-7393.23.4.987

Young, C. J. (2004). Contributions of metaknowledge to retrieval of natural categories in semantic memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(4), 909-916. doi:10.1037/0278-7393.30.4.909

Appendix A

Stimuli used in Experiments 1-4

	Meaning A		Meaning B	
Target Word	Cue Word 1	Cue Word 2	Cue Word 1	Cue Word 2
bat	swing	hit	fangs	cave
bow	ribbon	gift	arrow	quiver
bug	bother	irritate	hornet	ant
card	sympathy	valentine	jack	deck
cast	crew	play	crutch	sling
cell	jail	inmate	biology	nucleus
change	affect	shift	quarter	cents
charm	wit	persuade	luck	bracelet
check	mark	signature	money	payment
chest	trunk	cooler	hairy	ribs
china	dishes	porcelain	asia	dragon
cold	weather	glacier	medicine	sick
company	guest	visitor	industry	personnel
court	tennis	basketball	justice	verdict
crane	bird	stork	construction	erect
cricket	grasshopper	insect	england	sport
date	year	birth	blind	couple
draft	recruit	army	beer	bar
fall	crash	ledge	semester	summer
fire	gun	trigger	candle	heat
fly	travel	jet	maggot	buzz
foot	kick	boot	mile	yard
form	structure	shape	document	contract
foul	baseball	soccer	vulgar	unpleasant
free	liberty	independent	sample	volunteer
glare	stare	eyes	sun	light
head	ears	face	leader	boss
horn	antler	antelope	trombone	brass
husky	big	hefty	sled	dog
interest	hobby	curiosity	loan	rate
iron	crease	press	rust	metal
marble	granite	statue	glass	toy
march	parade	soldier	month	spring
might	strong	power	probability	maybe
nail	polish	hand	screw	hook

(continued)

Target Word	Meaning A		Meaning B	
	Cue Word 1	Cue Word 2	Cue Word 1	Cue Word 2
note	music	pitch	letter	bulletin
nut	cracker	fruit	crazy	freak
organ	heart	lung	keyboard	piano
park	slide	pigeon	curb	driveway
pawn	chess	queen	sell	shop
pen	scribble	writer	cage	pig
picket	strike	demonstration	fence	stake
pipe	tobacco	smoke	leak	wrench
pit	cherry	olive	hole	ditch
pitcher	jug	pour	throw	catch
point	peak	spike	there	finger
pressure	gauge	atmosphere	stress	anxiety
rare	steak	medium	unusual	scarce
reflect	prism	mirror	pause	think
right	direction	turn	proper	ideal
ring	alarm	telephone	wedding	ruby
seal	porpoise	flipper	envelope	wrap
shed	tools	shack	hair	fur
sink	drown	ship	toilet	counter
steer	bull	cattle	drive	navigator
straw	hay	scarecrow	tube	sip
tick	lice	flea	clock	watch
tie	vest	jacket	fasten	string
toast	tribute	champagne	butter	crumb
well	oil	water	health	able

Appendix B

Stimuli used in Experiment 5

	Meaning A				Meaning B			
Target Word	Cue Word 1	Cue Word 2	Cue Word 3	Cue Word 4	Cue Word 1	Cue Word 2	Cue Word 3	Cue Word 4
bat	swing	hit	pitch	softball	fangs	cave	vampire	nocturnal
bow	arrow	quiver	archer	hunting	ribbon	gift	sash	wrapping
card	sympathy	valentine	note	thanks	jack	deck	deal	bluff
cast	crutch	sling	plaster	fracture	crew	play	production	characters
cell	biology	nucleus	bacteria	neuron	jail	inmate	dungeon	prison
change	quarter	cents	coin	nickel	affect	shift	revolution	swap
check	status	evaluate	inspect	review	money	payment	cash	deposit
chest	trunk	cooler	drawer	dresser	flat	ribs	breast	burly
china	dishes	porcelain	delicate	plates	asia	dragon	dynasty	emperor
cold	weather	glacier	russia	north	medicine	sick	sinus	remedy
company	industry	personnel	organization	department	guest	visitor	alone	host
court	justice	verdict	lawsuit	legal	tennis	basketball	racquetball	volleyball
crane	construction	erect	hoist	lift	bird	stork	heron	feather
date	blind	couple	prom	romantic	year	birth	event	appointment
fall	semester	summer	leaf	term	crash	ledge	leap	dive
fire	candle	heat	coal	lighter	gun	trigger	cannon	shoot
fly	travel	jet	pilot	helicopter	maggot	bug	mosquito	pest
foot	mile	yard	meter	measurement	kick	boot	pedal	sole
free	liberty	independent	roam	america	sample	volunteer	charity	clinic
head	ears	face	nose	eyes	leader	boss	chief	principal
horn	antler	antelope	tusk	ram	trombone	brass	blow	tuba
interest	relevance	curiosity	appeal	concern	loan	rate	credit	lend

(continued)

Target Word	Meaning A				Meaning B			
	Cue Word 1	Cue Word 2	Cue Word 3	Cue Word 4	Cue Word 1	Cue Word 2	Cue Word 3	Cue Word 4
iron	rust	metal	copper	mineral	crease	press	wrinkle	starch
might	probability	maybe	uncertain	perhaps	strong	power	greatness	force
nail	screw	hook	tack	stake	polish	hand	toes	file
nut	cracker	fruit	crunchy	acorn	crazy	freak	insane	weird
organ	heart	lung	donor	intestine	keyboard	piano	church	music
park	slide	pigeon	bench	grove	curb	driveway	lot	asphalt
pawn	chess	queen	king	rook	sell	shop	store	broker
pen	scribble	writer	letters	cursive	cage	pig	corral	coop
pipe	tobacco	smoke	cigar	snuff	leak	wrench	plumber	steel
pit	hole	ditch	trench	gravel	cherry	olive	peach	plum
point	there	finger	aim	direct	peak	spike	apex	tip
pressure	stress	anxiety	tense	trouble	gauge	atmosphere	valve	vapor
reflect	pause	think	consider	contemplate	prism	mirror	light	image
right	proper	ideal	just	ethics	clockwise	turn	way	side
ring	wedding	ruby	gold	jewel	alarm	telephone	chime	buzz
seal	porpoise	flipper	otter	walrus	envelope	wrap	cover	tape
shed	tools	shack	hutch	barn	hair	fur	slough	molt
sink	toilet	counter	clog	bathtub	drown	ship	swim	dunk
steer	drive	navigator	guide	wheel	bull	cattle	livestock	rodeo
straw	hay	scarecrow	wicker	hut	tube	sip	cup	suck
tie	vest	jacket	tuxedo	shirt	fasten	string	rope	bondage
toast	butter	crumb	muffin	eggs	tribute	champagne	salute	drink
well	oil	water	oasis	spring	health	able	prosper	fine

Appendix C

Study Strategies Listed in Questionnaire in Experiment 4

Strategy Label	Full Text Used in Questionnaire
Cue-target Association	Made just one association (one meaning) between a target word and cue words.[a]
Multiple Cue-target Associations	Made multiple associations (more than one meaning) between a target word and cue words.[a]
Inter-item Association	Made associations between target words.[b] / Made associations across the list (for example, multiple target words).[a]
Target Focus	Focused more on the target words.[a]
Mental Imagery	Used mental imagery (formed a picture in your head).
Rote Rehearsal	Repeated words over and over in your head.
Verbalization	Spoke words out loud or under your breath.
Narrative	Put words into a sentence, phrase, or story.
Personal Significance	Related words to something personally significant.
Observation	Just read or looked at the words.
Multiple Target Meanings	Thought of multiple meanings for a target word.[b]

Note. Strategy labels are for reference and were not used in the questionnaire. [a]Used only in study-target-with-cues condition. [b]Used only in study-targets-only condition.

Appendix D

Test Strategies Listed in Questionnaire in Experiment 4

Strategy Label	Full Text Used in Questionnaire
Free Recall + Match	Tried to remember any target words from the study phase, then tried to see if any of those fit with the cue word(s).
Previous Answer + Match	Tried to remember any previous answers you had given on the test, then tried to see if any of those fit with the cue word(s).
Generate-Recognize	Tried to think of any words at all that fit with the cue word(s), and then tried to remember if any of those words were target words you had studied.
Generate	Tried to think of any words at all that fit with the cue word(s), but DID NOT try to remember if any of those words were target words you had studied.
Serial Cue Use	[When there were two cue words on a single trial]: Focused on one cue word at a time.
Cue Relationship, Simultaneous	[When there were two cue words on a single trial]: Tried to figure out a relationship between the two cue words.
Cue Relationship, Sequential	Tried to figure out a relationship between the current cue word(s) and cue word(s) from previous trials.
Direct Search: One Cue	Used a single cue word by itself to directly search your memory for a target word.
Direct Search: Both Cues	Used multiple cue words together to directly search your memory for a target word.
Recall of Old Cues	Tried to remember cue words from the study phase that fit with the current cue word(s), then used those to try to remember the target word.[a]

Note. Strategy labels are for reference and were not used in the questionnaire.
[a]Used only in study-target-with-cues condition.

Appendix E

Study Strategies Listed in Questionnaire in Experiment 5

Strategy Label	Full Text Used in Questionnaire
Cue-target Association	Made an association between the two cue words and the target word.
Separate Cue-target Associations	Made separate associations for each of the two cue words with the target word.
Single Cue Focus	Picked just one of the two cue words to study with the target word.
Inter-item Association	Made associations across the list (for example, multiple target words).
Target Focus	Focused more on the target words.
Mental Imagery	Used mental imagery (formed a picture in your head).
Rote Rehearsal	Repeated words over and over in your head.
Verbalization	Spoke words out loud or under your breath.
Narrative	Put words into a sentence, phrase, or story.
Personal Significance	Related words to something personally significant.
Observation	Just read or looked at the words.

Note. Strategy labels are for reference and were not used in the questionnaire.

Appendix F

Test Strategies Listed in Questionnaire in Experiment 5

Strategy Label	Full Text Used in Questionnaire
Free Recall + Match	Tried to remember any target words, then tried to see if any of those fit with the two new cue words.
Generate-Recognize	Tried to think of any words at all that fit with the two new cue words, and then tried to remember if any of those words were target words you had studied.
Generate	Tried to think of any words at all that fit with the two new cue words, but DID NOT try to remember if any of those words were target words you had studied.
Recall of Old Cues	Tried to remember old cue words that fit with the new cue words, then used the old cue words to try to remember the target word.
Inter-cue Association	Tried to figure out a relationship between the two new cue words.
Serial Cue Use	Focused on one new cue word at a time.
Direct Search: One Cue	Used a single new cue word by itself to directly search your memory for a target word.
Direct Search: Both Cues	Used both new cue words together to directly search your memory for a target word.

Note. Strategy labels are for reference and were not used in the questionnaire.