TOWARDS AN UNDERSTANDING OF THE EFFECTS OF ENCODING VARIABILITY ON LONG-TERM MEMORY

Milton E. Picklesimer

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctorate of Philosophy in the Department of Psychology and Neuroscience.

Chapel Hill
2016

Approved by:

Neil W. Mulligan

Peter A. Ornstein

Peter C. Gordon

Kelly S. Giovanello

Joseph B. Hopfinger

ABSTRACT

Milton E. Picklesimer: Towards an Understanding of the Effects of Encoding Variability on Long-Term Memory
(Under the direction of Neil W. Mulligan)

Prior research on encoding variability has often employed it as an auxiliary concept in an attempt to explain what gives rise to the spacing effect. However, tests of this hypothesis serendipitously revealed that spacing modulates the effects of encoding variability. It often results in superior performance (relative to encoding constancy) at short repetition lags, and this difference dissipates at longer lags. The chunking hypothesis is an extant but scarcely known theory that can account for this frequent pattern in the literature. This theory's core assumption is that encoding variability can enhance memory when something is recognized as a repetition, and then the information from both presentations is chunked into an enriched memory code (all of which is presumably easier at short lags). However, this core assumption about recognizing repetitions (a.k.a., study-phase retrieval) remains untested. The current study tested this assumption—as well as other ones implied by the chunking hypothesis—using methods akin to a continuous recognition memory paradigm. We also employed a metric of chunking to see if details from both presentations of a target stimulus exhibited statistical dependence between each other in ways predicted by the chunking hypothesis. The results largely supported the predictions of the chunking hypothesis, however, some amendments are needed to account for the effects of associative distance and retrieval difficulty. Regardless, the current study helped elucidate when and why encoding variability enhances memory.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

The repetition of information is generally found to benefit memory (Hintzman, 1976). However, information need not be repeated in the same manner (*encoding constancy*), it can be repeated in ways that encourage a different analysis—causing the learner to processes a different set of attributes about the stimulus. This is known as *encoding variability*. Encoding variability sometimes benefits memory more than encoding constancy. For example, Glenberg and Smith (1981, Exp. 1) found that an encoding variation of the orienting tasks used on words in a list-learning paradigm resulted in better memory for the words. On each trial, subjects were presented with a word and asked to make a judgment about it. The tasks were to rate the item's size (i.e., bigger or smaller than a given reference item) or the item's pleasantness (i.e., more or less pleasant than a given reference item). Items in the encoding constancy conditions were subjected to the same orienting task on both presentations. Items in the encoding variability condition were subjected to the size task on one presentation and the pleasantness task on the other. Glenberg and Smith found that performance on a subsequent free recall test was highest for items subjected to encoding variability. Moreover, encoding variability can enhance important aspects of learning like the transfer of conceptual knowledge to novel contexts (see Bransford, 1979; Butler & Marsh, 2012, diVesta & Peverly, 1984; Nitsch, 1977). However, in other cases, encoding variability has resulted in worse memory than encoding constancy. For example, in a study similar to Glenberg and Smith (1981, Exp. 1), Young and Bellezza (1982, Exp. 4) used a list-learning paradigm where participants were required to make size and/or pleasantness ratings. Young and Bellezza found that encoding variability resulted in worse free

recall than encoding constancy. There have also been instances in which encoding variability did not enhance performance in a transfer paradigm (Dempster, 1987, Exps. 2 – 5).

The inconsistency in encoding variability's effects is all the more surprising when one recognizes it as a form of elaborative encoding. In general, memory improves as one adds features to an existing memory code. The amount of features that are processed at encoding is the amount of elaboration. Elaborative encoding is also thought to be the means by which deeper (i.e., more semantic) processing often results in better memory than shallow processing (Craik & Lockhart, 1972; Craik & Tulving, 1975; Klein & Saltz, 1976). Relative to shallow processing of the superficial features of a stimulus, deeper processing of semantics involves a more elaborate process that consequently appends more features to a memory code. In that sense, processing a stimulus in different ways (i.e., encoding variability) could be considered a form of elaborative encoding. Therefore, given that we know that elaborative encoding can enhance memory, it is all the more surprising that elaborations through encoding variability do not always enhance memory (especially considering that many encoding variations in the literature have been semantic in nature). Therefore, given that we know that the genus elaborative encoding usually enhances memory, it behooves us to understand why various species of encoding variability help in some cases but not in others.

Finally, note that, because encoding variability can enhance memory, this also means encoding constancy is not always the optimal approach. That is, an optimal learning scenario might require a judiciously constructed regimen of encoding constancy and variability. However, even musing on rules of thumb for such a regimen cannot begin until the determinants of positive, null, and negative encoding variability effects have been elucidated.

In order to capitalize on the promise of encoding variability, we first need a basic theoretical understanding of the determinants of its effects so that we may more confidently predict when it will and will not enhance memory. Episodic list learning studies can reveal the effects of an encoding manipulation on the initial stages of memory formation. As it turns out, there is a wealth of such studies that can guide theory development of our understanding of the effects of encoding variability. However, most of these studies have focused on the auxiliary use of encoding variability as a potential explanation for other memory phenomena. To preview, no existing study adequately tests theories of encoding variability in and of itself. However, the current study contains experiments designed to test and develop such a theoretical foundation.

A Brief History of Encoding Variability and The Spacing Effect

Encoding variability is probably most well known for being invoked as an explanation for various memory phenomena. For example, repetitions that invite a different analysis of the information are thought to be more efficacious. In that sense, encoding variability is viewed as an underlying cause of the repetition effect (e.g., Craik & Lockhart, 1972). Traditionally, the spacing effect was attributed to a similar cause (e.g., Bower, 1972; Crowder, 1976). In studies of the spacing effect, items are repeated throughout a list at various intervals. The number of intervening items between the two presentations of a repeated item is referred to as the repetition lag. Items repeated successively are referred to as lag 0 items, meaning there are zero intervening items. Repetitions spaced apart by one intervening item would be called lag 1 items, etc. Also note that, for the prototypical experiment, the repetitions are nominally identical. After presentation of the list, a memory test is administered. Researchers typically find that, as the repetition lag increases, so does performance on the memory test (for reviews, see Benjamin

& Tullis, 2010; Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Hintzman, 1976). The memory advantage for spaced items (lag > 0) over massed items (lag = 0) is the spacing effect.

An encoding variability hypothesis of the spacing effect asserts that, as the presentations of a stimulus become spaced farther apart in time, they become less likely to invite the same analysis of the stimulus—despite the fact that the presentations are nominally identical. This change is thought to be due to drift in one's mental and/or physical context that typically correlates with changes in time. Assuming that the mental or physical context affects the stimulus attributes attended to by the learner, then contextual drift should promote the encoding of a greater multitude of stimulus attributes (i.e., encoding variability) (Bower, 1972).

The encoding constancy (EC) curve in Figure 1 depicts a stylized profile of the typical effects of spacing on nominally identical repetitions of stimuli. The benefits of spacing increase most dramatically for shorter lags and then plateau at longer ones[1]. Moreover, if spacing promotes contextual drift, which in turn promotes encoding variability, then circumventing the process via an overt manipulation of encoding variability should eliminate the spacing effect by lifting memory for all lags in the encoding variability (EV) condition up to the level of memory for the longer lag items in the EC condition. That is, experimentally inducing a change in encoding from presentation 1 to presentation 2 (P_1 and P_2, respectively) should result in memory rivaling that normally produced under the longer lags of EC.

However, most tests of an encoding variability hypothesis of the spacing effect have failed to find the aforementioned data pattern depicted in Figure 1. Failures to produce this pattern have occurred regardless of whether the design overtly required a different analysis at P_1

[1] Several researchers have noted, however, that the tonicity (i.e., monotonic or non-monotonic) of the function relating recall to the repetition lag is also affected by the ratio of the repetition lag to the retention interval (Benjamin & Tullis, 2010; Cepeda et al., 2006; Glenberg, 1976). Nonetheless, the monotonic curve depicted in the EC condition is arguably typical of the kind produced by the combinations of repetition lags and retention intervals used in conventional laboratory studies.

and P_2; such as changing orienting tasks (Bird, Nicholson, & Ringer, 1978, Exp. 1; Glenberg & Smith, 1981, Exp. 1; Shaughnessy, 1976); inviting a different analysis of a target word via a change in the cue words paired with it (Madigan, 1969, Exp. 2); or changing the preceding list items of a target word at P_1 and P_2 (Johnson, Coots, & Flickinger, 1972; Maskarinec & Thompson, 1976, Exp. 1).

Figure 1. Hypothetical Data Pattern Supportive of an EV Hypothesis of the Spacing Effect

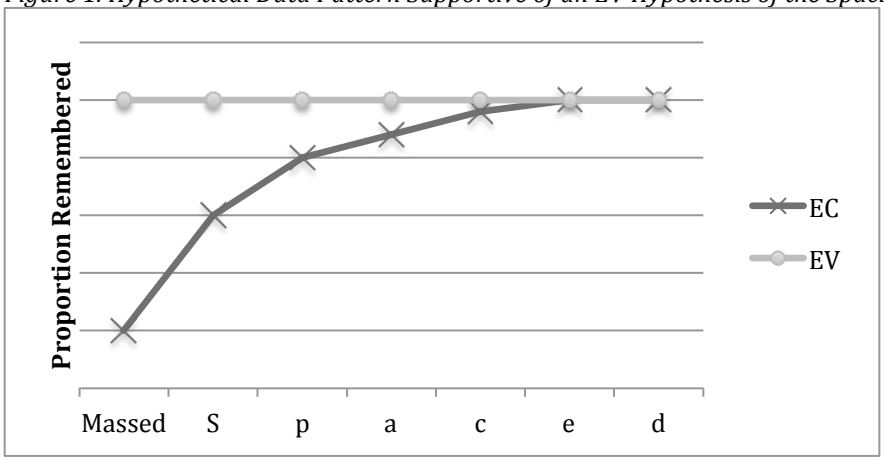


Fig. 1 Notes: Data pattern that would be supportive of an encoding variability hypothesis of the spacing effect. EC = encoding constancy and EV = encoding variability. The massed bin refers to repetitions that were presented successively. The spaced bins refer to non-successive repetitions.

It is clear that support for an EV hypothesis of the spacing effect has been quite variable itself. A more appropriate stance might be to think of encoding variability as an important standalone factor in learning—but whose properties have thus far been poorly understood. It seems forlorn as an explanation of the spacing effect unless one restricts their discussion to a narrow range of spacing effects and encoding variability manipulations. Its failure as an auxiliary explanatory tool has even led some to prematurely conclude that "encoding variability is an attractive theoretical concept but seems empty" (Roediger, Sanches, & Agarwal, 2011). Other research on EV suggests that its effects are moderated by spacing, and not vice versa. More importantly, the interaction between EV and repetition lag may be indicative of one of the determinants of the effects of EV.

The Chunking Hypothesis

Episodic memory experiments often strive to strictly bifurcate periods of study and retrieval so that both acts of memory can be examined separately. This pursuit is problematic in experiments on repetition effects because retrieval is likely to take place during the study phase if someone recognizes that an item has been repeated (i.e., study-phase retrieval). Study-phase retrieval occurs when a repeated stimulus is recognized as a repetition. Interestingly, research on the spacing effect has found that study-phase retrieval declines with spacing (Thios & D'Agostino, 1976; Johnston & Uhl, 1976, Exp. 2). This factor may also be important to understanding the effects of EV on memory. Given that greater spacing of repetitions produces more failure in study-phase retrieval, it is possible that EV may interact with repetition lag in a way that compounds the decline in study-phase retrieval. That is, the effect of temporal distance might be compounded by the effect of associative distance. This may then determine the effect of variably encoded repetitions on memory. In fact, several studies have found that the effects of EV depend on the repetition lag. For one, positive effects of encoding variability are often found under massed repetitions (Bellezza & Young, 1989; D'Agostino & DeRemer, 1973, Exp. 2; Dellarosa & Bourne; 1985; Glenberg, 1979, Exp. 1; Glenberg & Smith, 1981; Hintzman, Summers, & Block, 1975, Exp. 2; Jacoby, 1972; Madigan, 1969, Exp. 2; Maskarinec & Thompson, 1976; McFarland, Rhodes, & Fray, 1979, Exp. 1; Paivio, 1974, Exp. 2; Russo, Mammarella, & Avons, 2002; Exp. 1; Shaughnessy, 1976, Exp. 3; Thios, 1972; Verkoeijen, Rikers, & Schmidt, 2004; Winograd & Raines, 1972, Exp.1). It is also commonly found that, as the repetition lag increases, performance under EV eventually becomes eclipsed by performance under EC. This is because, under EC, memory often increases monotonically with spacing (for an exception, see Glenberg, 1976). Encoding variability, however, often results in a flattened

spacing function (D'Agostino & DeRemer, 1973, Exp. 2; Dellarosa & Bourne, 1985, Exp. 2; Madigan, 1969, Exp.2; Paivio, 1974, Exp. 2; Shaughnessy, Zimmerman, & Underwood, 1974; Verkoeijen et al., 2004). Sometimes EV can also interact with spacing in such a way that memory gets progressively worse with spacing (Thios, 1972; Bellezza & Young, 1989, Exp. 3). To account for this interaction between EV and spacing, Bellezza and Young proposed the chunking hypothesis (Young & Bellezza, 1982; Bellezza & Young, 1989).

The basic dynamics of the chunking hypothesis are relatively simple. The central tenets of this hypothesis are that the effect of encoding variability on memory is a function of two things: 1) the degree to which subsequent presentations of an item cue retrieval of prior presentations and 2) whether or not the newly sampled information gets integrated with that from previous traces, or instead forms an entirely different code within which new traces will be bundled.

Bellezza and Young's hypothesis places an immense amount of stock in the importance of study-phase retrieval. Other theorists had already noted that repetitions of an item that fail to be recognized as such show little-to-no benefit of spacing (Hintzman et al., 1975; Johnston & Uhl, 1976, Exp. 2; Melton, 1967; Thios & D'Agostino, 1976). Bellezza and Young believed that encoding variability was a potentially strong moderator of the likelihood of study-phase retrieval. If one assumes that the likelihood of P_2 cuing retrieval of P_1 is a function of the similarity of their encoding experiences, then one should also predict that encoding variability should make study-phase retrieval more difficult. In short, EV will not result in better memory than EC if an EV manipulation results in a failure of study-phase retrieval. Otherwise, EV can result in memory that is equal to or better than memory following EC (for a formal proof, see Young & Bellezza,

1982). Nonetheless, the effect of EV on study-phase retrieval can have a tremendous bearing on the second aspect of Bellezza and Young's hypothesis, the actual chunking process.

One major way by which Bellezza and Young's hypothesis accounts for positive, negative, and null effects of encoding variability is through a chunking process. If one's encoding experience during a repetition of a nominal stimulus has a sufficient amount of overlap with previous experiences, then the components sampled during the most recent experience can be retrieved and integrated (i.e., chunked) with the code in which previous experiences were bundled. This results in a richer memory code (i.e., a positive effect of encoding variability). If, however, the current encoding experience results in a failure to retrieve previous ones, then a new code is formed. Instead of one enriched code, this results in two relatively impoverished codes (i.e., a negative encoding variability effect). Note, however, that the dynamics of this chunking process require that the effects of repetitions are optimal when encoding is actually a balance between similarities and differences in processing. That is, there should in theory exist an optimal encoding regimen that consists of encoding variability.

The Chunking Hypothesis and Study-Phase Retrieval

There has been no direct test of the influence of study-phase retrieval on encoding variability effects. Suggestive support of a relationship between the two comes from a study by Bellezza and Young (1989, Exps. 1 & 2). They had participants study unrelated word pairs (e.g., *bottle – tower*). The target words, the second word in each pair, were repeated at lags of 0 or 60. Targets in the EC condition were repeated with the same accompanying cue (e.g., *bottle – tower* & *bottle – tower*), whereas targets in the EV condition were repeated with a different (and also unrelated) cue (e.g., *hotel – stone* & *ticket – stone*). Interestingly, recall of the target words was not solely a function of encoding condition or repetition lag. For massed (i.e., lag 0) items, recall

was higher in the EV condition. Conversely, recall of the spaced, lag 60 items was higher in the EC condition. Bellezza and Young's (1989) third experiment used a larger number of repetition lags and found a qualitatively similar interaction between encoding condition and repetition lag. More importantly, these recall results are consistent with the predictions of the chunking hypothesis. That is, study-phase retrieval and chunking should be easiest for massed items, therefore ensuring a memory advantage for massed EV items. However, study-phase retrieval and chunking should become less likely as the effects of temporal distance are compounded by the effects of associative distance. Thus resulting in a diminishing and eventual reversal of the EV advantage as repetition lag increases.

There exist other studies demonstrating that the effects of EV interact with spacing—and these interactions were at least attributed to alleged differences in study-phase retrieval (Greene & Stillwell, 1995; Verkoeijen et al., 2004)[2]. However, some of the explicit core predictions of the chunking hypothesis and other ones implied by it remain untested. Most importantly, if the hypothesis is correct then, as the lag increases, study-phase retrieval should decline more rapidly under EV. In addition, given that study-phase retrieval is affected by temporal *and* associative distance, the deviance of the encoding variation should affect the threshold lag at which EV results in worse memory than EC. For example, if a drastic variation in encoding results in very different experiences at P_1 and P_2, then the cuing potential of P_2 should be weakened, compared to a moderate variation between P_1 and P_2. Therefore the chunking hypothesis implies that the threshold lag for a negative effect of EV on memory should shorten as the encoding variation

[2] The interested reader might also like to know that Greene and Stillwell (1995) and Verkoeijen et al. (2004) arrived at explanations—perhaps by their own devices—very similar to the chunking hypothesis but made no mention of the works of Bellezza and Young (Young & Bellezza, 1982; Bellezza & Young, 1989). The author, MP, believes that these coincidences speak to the self-evident importance of study-phase retrieval.

becomes more deviant. Thus far, the effects of EV on study-phase retrieval and the above interaction between spacing and EV deviance remain untested.

One can also extend aspects of the hypothesis to predict that, for massed repetitions, even the largest deviations in encoding of a nominal stimulus should result in better memory than EC. Study-phase retrieval is essentially guaranteed for successive presentations. Thus, any variation in encoding will result in the sampling and chunking of new components into an older code. Another implication of the chunking hypothesis is that, given two EV manipulations, the more deviant one should result in a greater amount of chunked components under massed repetitions. To preview, the current study can test this implication because it will implement two EV manipulations that differ in deviance.

The appeal of the chunking hypothesis is its relative simplicity and apparent adequacy. However, it does have one major limitation: there are no specified qualitative conditions that lead to the creation of one enriched code versus two impoverished codes. This could lead to a general chunking hypothesis being applied in a post-hoc fashion. That is, if EV results in memory that is greater than or equal to EC, then one might assume that EV resulted in only one code. Conversely, if EV results in memory that is worse than EC, then one might assume that two impoverished codes were made. As one can see, this line of reasoning becomes circular because the relative standing of EV on a conventional dependent measure like accuracy or proportion correct is used to infer the number of codes. If a manipulation of EV results in accuracy or proportion correct that is worse than or no better than EC, it need not be due to the creation of two impoverished codes. However, the alternative analyses covered in the next section could be used to determine the existence of one or two codes (and whether this correlates with memory performance).

In sum, although the chunking hypothesis itself does not use relative standing on conventional dependent measures as a litmus test for the number of codes, there could exist a temptation in empirical analyses to apply a post-hoc account about the number of codes. Therefore caution should be observed when interpreting the data in terms of a chunking hypothesis. One's conclusions should not only be guided by relative standings on conventional dependent measures, but also by reasoning about the manipulations themselves. There may, however, be some alternative analyses of memory performance that could shed light on the number of codes created by various manipulations of EV. Such analyses could help provide a more revealing test of the chunking hypothesis.

The Chunking Hypothesis and The Chunking Process (P_1 & P_2 Memory)

If there is in fact a chunking process driving part of the interaction between EV and spacing, then one should be able to find evidence that 1) information from P_1 and P_2 is being chunked; 2) the amount of chunked information interacts with study-phase retrieval; and 3) that changes in chunking are modulated by spacing and the deviance of the encoding variation. As of now, the requisite data needed to evaluate these notions do not exist.

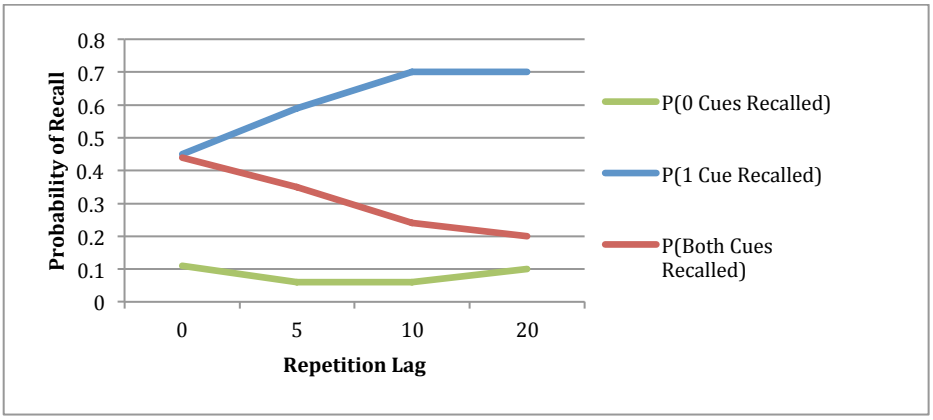
Evidence of EV mediating the probability of an actual chunking process has been scarce. To support the existence of such a process, one would need to demonstrate that information from both P_1 and P_2 has been bound into a composite memory code. Most studies of EV report only the conventional dependent measures of memory (e.g., accuracy and proportion correct). Furthermore, the overwhelming majority of studies using these measures restricted them to information shared by both presentations (e.g., recall of a target word associated with different cues at P_1 and P_2). To determine if EV does affect a chunking process, one must be in a position to test for evidence that some information unique to each presentation has been retained—and

that these pieces of information have become bound. Tests for such evidence have been scarce but, nonetheless, there is at least combined evidence from a few studies showing that the effects of EV on memory are driven by a chunking process whose likelihood changes with lag. These studies are reviewed below so that their results can be used to guide predictions for the current study.

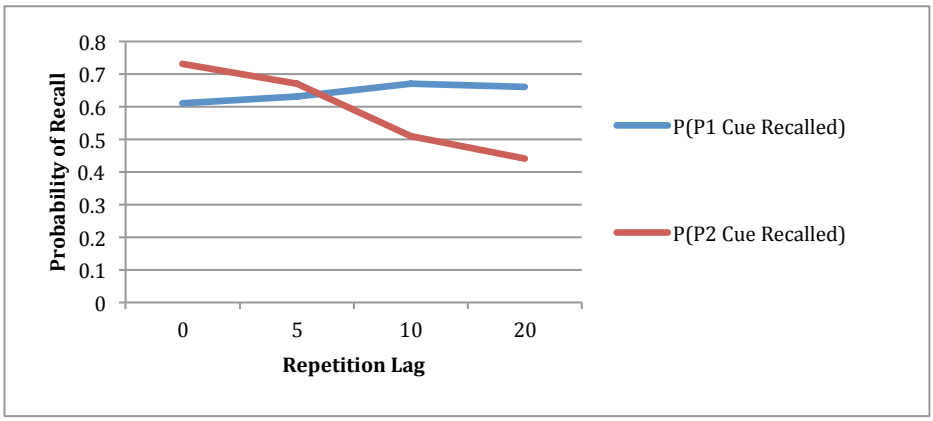
Indirect evidence has been reported by Thios and D'Agostino (1976) in a re-analysis of data from D'Agostino and DeRemer (1973, Exp. 2). This re-analysis at least showed that the retention of P_1 and P_2 info is affected by spacing. In the sole EV condition of D'Agostino and DeRemer (1973, Exp. 2), object phrases were repeated in 2 different sentence frames depicting different actions being performed on the object (e.g., *"The plane hit the flag pole."* and *"The bear climbed the flag pole."*). The sentences were given massed presentations (lag 0) or spaced presentations at lags of 5, 10, or 20 sentences. After encoding, subjects freely recalled as many object phrases as they could. Upon completion, they then wrote any sentence frames they could remember next to their corresponding object phrases. Note that each sentence frame counts as a detail unique to each presentation. The re-analysis of these data by Thios and D'Agostino (1976) plotted various aspects of memory for P_1 and P_2 information and how they changed with spacing. For the sake of brevity, the sentence frames from P_1 and P_2 are referred to as cues and the object phrases as targets. An adaptation of their re-analysis is presented below in Figure 2.

Figure 2. Analyses of P_1 vs. P_2 Recall Adapted from Thios & D'Agostino (1976)

a.



b.



c.

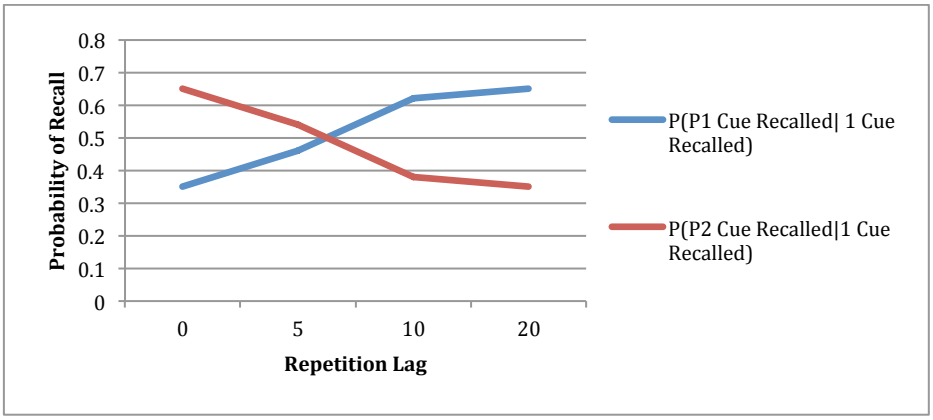


Fig. 2 Notes: A re-analysis of recall data from D'Agostino and DeRemer (1973, Exp. 2) reported in Thios and D'Agostino (1976). Each figure depicts different probability indices of recalling cues as a function of lag, given that the corresponding target was recalled. **a.** Probabilities of recalling 0, 1, or 2 cues as a function of lag. **b.** Probabilities of recalling P_1 *or* P_2 as a function of lag. **c.** Probabilities of recalling P_1 or P_2 as a function of lag given that at least 1 cue was recalled.

Several patterns from these figures suggest that the likelihood of chunking diminishes with spacing. In Figure 2a one can see that the probability of recalling both cues declines with

spacing whereas the probability of recalling only one cue increases. At least one part of the change appears to be driven by a tradeoff enhancement of P_1 info with spacing. Note that in Figure 2b the probability of recalling P_1 increases with spacing whereas the probability of recalling P_2 shows a concomitant decrease. A similar pattern is shown in Figure 2c using a more equitable comparison where P_1 and P_2 recall were conditionalized on the probability of recalling at least one cue in response to a target.

The decline in recall of both cues with spacing provides suggestive support of a chunking process that interacts with the effects of spacing. However, another pattern in the data suggests that an addendum may be needed to the chunking hypothesis. Namely, the opposite trajectories of P_1 and P_2 recall hint at another mechanism that might be driving down P_1—P_2 binding. Several theorists have suggested and/or shown that the locus of the effect of spacing on memory is primarily the enhancement of P_1 information (Benjamin & Tullis, 2010; Braun & Rubin, 1998; Hintzman, 2004; Thios & D'Agostino, 1976). As the repetition lag increases, study-phase retrieval becomes more difficult. When retrieval succeeds at these longer lags, the mnemonic effects are thought to be greater than those for items at shorter lags. Insofar as difficult retrieval is more mnemonically potent than easy retrieval, then, holding retrieval success constant, memory should be better following more difficult retrieval attempts. This is akin to a testing effect being greater for more difficult but successful retrieval attempts (Carpenter & DeLosh, 2006). Therefore, longer lags disproportionately benefit memory for P_1 details because the mnemonic potency of retrieving P_1 increases with lag—but only up to a point (for a simulation, see Benjamin & Tullis, 2010)[3]. However, one might intuit that, because P_2 trials occur closer to the memory test, P_2 details should be easier to retrieve. Contrary to this intuition, it would

[3] Also see Benjamin and Tullis (2010) for additional discussion on the combined influences of study-phase retrieval difficulty and potency on spacing effects.

appear that the retrieval practice effect on P_1 details results in memory that eclipses the "recency effect" for P_2 details. This apparent tradeoff between P_1 recall and P_1—P_2 binding may be a function of time-in-working-memory spent retrieving P_1; furthermore, this tradeoff may be driven by spacing (Braun & Rubin, 1998). That is, as more time is spent retrieving P_1 details, less time is allotted for chunking P_1 and P_2 details. Thus, holding trial length constant, it is possible that chunking success is inversely related to P_1 retrieval time. If so, the chunking hypothesis may require an addendum to its purported mechanisms.

Regardless of the source of the lag-related decline in recalling P_1 and P_2 together, several issues more central to the current proposal remain unresolved by the re-analysis reported in Thios and D'Agostino (1976). For one, it did not compare encoding variations of different deviances. Such a comparison is key to testing the predictions of the chunking hypothesis given that it assumes that P_1—P_2 similarity modulates the probability of study-phase retrieval. In addition, it still remains unclear if P_1 and P_2 details are ever actually chunked into one code. Thios and D'Agostino (1976) provided no resolution to this issue because they did not determine whether or not P_1 and P_2 details were dependent on one another; nor did they determine whether or not such dependence interacted with lag and study-phase retrieval.

Slamecka and Barlow (1979, Exp. 2), however, provided several results relevant to the concerns raised above. To induce encoding variability, Slamecka and Barlow (1979, Exp. 2) paired target words with different cue words at each presentation. The cue variation was designed to evoke meanings of a similar (e.g., *antler – horn* & *tusk – horn*) or different (e.g., *summon – page* & *sheet – page*) nature at each presentation. These were the Same and Different meaning conditions, respectively. After encoding all of the pairs, their subjects took a cued

recall test. All cues were shown on the test and subjects were instructed that they might have to respond with some of the same targets twice.

One analysis of the cued recall data revealed some interesting patterns relevant to the current proposal. This analysis consisted of determining P_1—P_2 dependence by comparing the probabilities of various cued recall outcomes. The conditional probability of recalling a target in response to one cue—given that the target was also recalled in response to the other cue—was compared to the base rate probability of recalling the target in response to one cue. In other words, P(Recall to 2nd Cue|Recall to 1st Cue) was compared to P(Recall to 2nd Cue) to determine if they were different or equal. If the former is greater than the latter, then the two cues (i.e., P_1 and P_2 details) are dependent upon one another. However, if the cues confer no benefits to each other and are therefore independent, then the two probabilities should be statistically equal. In other words, if the first proportion is significantly than the second, then that means chunking occurred. If they are equal, then it means there is no evidence of chunking. Slamecka and Barlow's analysis showed that, in the Same meaning condition, the effects of both cues showed dependence. However, such evidence was not found for the Different meaning condition. This suggests that, at least at the level of the information used for a retrieval cue, the deviance of the encoding variation can affect the dependence (i.e., chunking) of P_1 and P_2 information. Also note that, while these results indicate differences in chunking, the presence of a conditionalized-unconditionalized difference in one case and the absence of it in another was only coincidence. While the results of Slamecka and Barlow (1979, Exp. 2) do address some of the questions left unresolved by Thios and D'Agostino (1976), they fail to answer others. Most importantly, Slamecka and Barlow (1979) only used one lag throughout (24 pairs). This means that is not

possible to tell if, as predicted by the chunking hypothesis, the effect of EV deviance on chunking (i.e., P_1—P_2 dependence) interacts with spacing and presumably study-phase retrieval.

The Current Study

The representative literature reviewed in the introduction shows that research seeking to understand EV in and of itself is sparse. This is evidenced by the large literature testing it as a potential explanation of the spacing effect. Insight may still be gained from the literature on EV and the spacing effect because a survey of said literature also shows that the effect of EV often depends on the repetition lag. The chunking hypothesis (Young & Bellezza, 1982; Bellezza & Young, 1989) was proposed as a set of mechanisms to help explain the dynamics of EV and spacing. However, the most explicit prediction of the chunking hypothesis, that EV and spacing modulate the probability of study-phase retrieval, remains untested. Furthermore, other predictions of the chunking hypothesis remain untested: primarily that the deviance of the encoding variation should interact with the effect of spacing on study-phase retrieval. Another concern is that there exists little proof that EV can result in the chunking of information. Alternative analyses of memory for P_1 and P_2 details should be employed to test for evidence of this. Furthermore, it remains to be determined if such evidence of chunking interacts with study-phase retrieval—which is presumably affected by spacing and the deviance of the encoding variation.

Experiment 1 – The Role of Study-Phase Retrieval

The first experiment directly assessed study-phase retrieval and how it relates to the effects of encoding variability and lag. During the encoding phase, participants overtly made repetition judgments over target words that were shown once or twice. The repeated targets were

presented in either the exact same semantic context or a variation of it that 1) preserved the same meaning or 2) changed to a completely different one. We then tested their memory with a cued recall test specially constructed to examine the effects of encoding variability on chunking.

Piloting. We employed homonyms as our target words. These are words that can take on two or more different meanings but have the same pronunciation and spelling for all meanings. The inherent qualities of these stimuli ensured that our targets had identical perceptual characteristics across meanings. Our first approach to inducing the intended semantic context of a target word was to pair it with a semantically related cue word of the intended meaning (i.e., *purple – maroon* vs. *island – maroon*). After constructing the required number of paired associates (226 to be exact), we piloted our materials in the full version of our paradigm. To ensure that participants were encoding the target words in the intended semantic context, we asked them to form a sentence combining the two words and to state this sentence out loud. To our surprise, we found that they did not always make a sentence using the intended meaning of the target (e.g., for example, for the pair *island – maroon* one participant said "While vacationing on the island, I got so sunburned that I looked maroon."). Moreover, these problems cropped up most frequently for the less dominant meaning of a target word. In the previous example, the island-related meaning of *maroon* is the less dominant of the two. This kind of problem undermined our ability to induce the intended semantic context and also imposed a confound because it disproportionately affected the less dominant meanings of target words.

Given these limitations, we abandoned the use of paired associates. They were apparently too blunt of an instrument to reliably induce certain meanings of homonyms. We did, however, retain homonyms as our target words. To more reliably induce the intended meaning, we created sentences in which we could embed the targets. Previous EV studies have used a

similar approach (for a few examples, see Bobrow, 1970; D'Agostino & DeRemer, 1973; Dellarosa & Bourne, 1985, DeRemer & D'Agostino, 1974; Jacoby, 1972; Postman & Knecht, 1983, Exps. 1 & 2; Thios, 1972). To make use of the many paired associates we constructed, we incorporated both the cue and target words in the sentences we constructed. Thus, we created 4 unique sentences for each target word (2 for each meaning) and 1 unique sentence for each of the 10 primacy words—resulting in a grand total of 226 sentences. The experiments described hereafter employed these sentences as our way of controlling the semantic context. The basic attributes of our materials are described in brief in the methods for Experiment 1. In-depth descriptions are provided in the Appendix.

METHODS

Participants.

Seventy-two introduction to psychology students participated in exchange for course credit. All reported being between the ages of 18-30, having native or native-level fluency in English, and no disorders that would prevent them from reading aloud for several minutes at a time. Six subjects had to be replaced because they appeared to have excessive difficulties following the instructions and/or performing the encoding task (4 in the Similar condition and 2 in the Different condition). This was evidenced by the fact that the number of study-phase trials on which they failed to provide a repetition judgment was more than 2 standard deviations above the mean (the resulting cut-off was 5 or more omissions). Thus, our final sample, including the replacements, is based on 72 subjects who omitted responses on fewer than 5 study-phase trials.

Design.

The experiment used a mixed design, with all factors except one manipulated within-subjects. Repetition lag was manipulated within-subjects and had 3 levels: 0, 6, or 18

intervening items. Presentation frequency was manipulated within-subjects. All subjects experienced items that were presented once (1P) or twice (2P). Within the 2P items, we manipulated the repetition type: encoding constancy (EC) or encoding variability (EV). All subjects had an EC and EV condition. The type of EV, however, was manipulated between-groups. One group experienced encoding variations within the same meaning (the Similar group) and the other group experienced encoding variations between different meanings (the Different group). The pool of useable homonyms for this experiment was limited, so this design feature allowed us to have more target words per subject in all conditions than would be possible if EV type were also manipulated within-subjects.

Materials.

Target Words. The target words were homonyms, words that take on two or more different meanings but have the same pronunciation and spelling for all meanings. The inherent qualities of these stimuli ensured that our targets had identical perceptual characteristics across meanings. Our experimental design required a total of 64 target words. Fifty-six were drawn from the English homonym norms of Armstrong, Tokowicz, and Plaut (2012). The remaining 8 were taken from the materials used by Slamecka and Barlow (1979). The lexical attributes of the targets are described in the Appendix.

Sentences. The sentences were constructed to ensure that the targets only appeared in the 4 requisite sentences written for each target word (2 sentences for each meaning). The sentences were also built around the cue-target pairs (described in the Appendix) to help constrain the semantic details expressed in each sentence. The sentence frames used in the recall test for each target were also piloted to minimize the rate at which participants could guess the correct target

word. After multiple waves of free association piloting, we were able to reduce the correct guessing rate for our bank of sentences to an average of 12%.

Counterbalancing Scheme. Ten targets and their respective sentences were used as primacy items that remained constant for all subjects. The remaining 54 targets were divided up into 3, 18-item sub-lists to facilitate counterbalancing. These 54 targets were counterbalanced across lag, repetition type (EC or EV), and EV type (Similar or Different). The rest of the counterbalancing scheme, and the creation of the study-phase lists, are described at length in the Appendix.

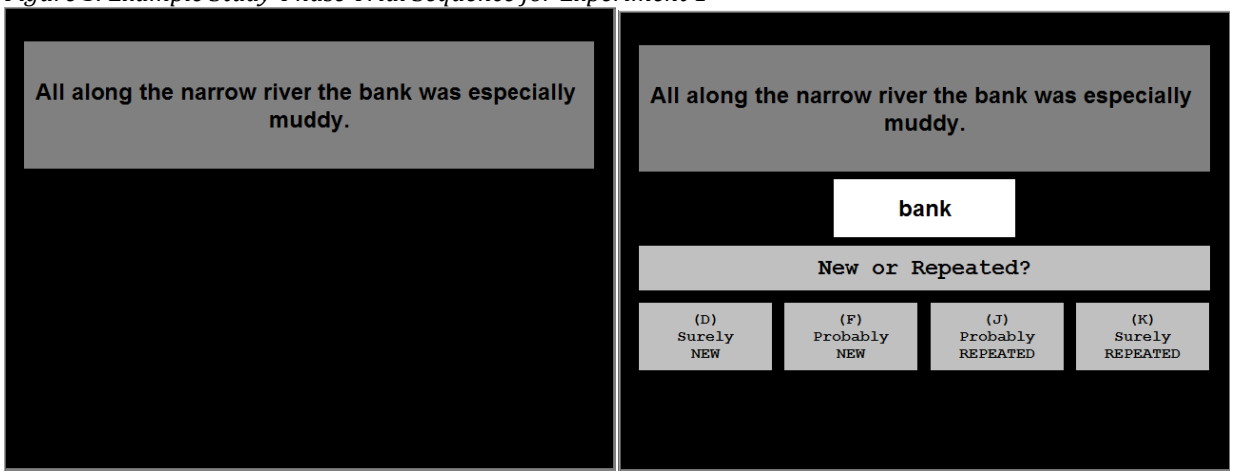
Procedures.

After obtaining informed consent, participants were instructed on the procedures of the first phase, the encoding phase. (However, note that they were not alerted to the fact that a memory test followed this phase). They were told that they would be shown sentences one at a time. They were assured that the sentences were grammatically correct and meaningful, but that they would not link together coherently like the sentences of a story or book. Therefore, they would seem out of context, like excerpts from a story or book. The level of a detail would also vary from sentence to sentence and some of the sentences would be repeated. We gave participants these forewarnings to mitigate any confusion that might arise from reading many unrelated sentences.

They were then told that their two tasks for this experiment were to 1) read the sentence and then 2) make a repetition judgment about one of the words in the sentence. When they read the sentence, they were required to read it aloud and try to comprehend it as quickly and accurately as possible. We required them to read the sentences aloud so that we could verify that they were reading the sentences (and participants were told this). They were also informed that

they would have 7.5 sec to read each sentence. Afterwards, the computer would select one word from the sentence and they would have to decide whether or not they had seen it before in the experiment. (They were told that the computer randomly selected the word but, unbeknownst to them, the word in question was always one of the aforementioned target words). If they had not seen the word before in the experiment, they were to call it "new". If they had seen it before anywhere in the experiment, they were to call it "repeated." Along with their judgment, they had to rate the confidence as "surely" or just "probably." They were informed that they had 3.5 sec to make this judgment for a given word. These instructions were provided alongside an example trial sequence shown in Figure 3. They key mappings were explained to the participants and they were reassured the key mappings would be shown on every trial. They were also reminded to make their repetition decision as quickly and accurately as possible during the 3.5 sec allotted for it. They were also told that the end of each trial would be signaled with a .5 sec crosshair, after which the next sentence would appear. Any remaining questions were answered and then the participant proceeded to the encoding phase. They were not told of the total length of the encoding phase nor were they made aware of the memory test following it.

Figure 3. Example Study-Phase Trial Sequence for Experiment 1



Sentence alone for 7.5 sec

Repetition judgment over target word for 3.5 sec

Fig. 3 Notes: Depicted above is the example of a study-phase trial shown to all participants before the experiment began. After the repetition judgment portion of a trial, a crosshair appeared in the middle of the screen for 500 msec (not depicted).

Cued Recall Phase

Materials.

Test List Construction. In order to compute a chunking index, all target words in a subject's respective EV condition had to be tested twice. In case this testing process had any idiosyncratic effects on memory, the target words in all other conditions were tested twice as well. Targets in the 1P and EC conditions were necessarily tested using the same sentence frame twice. Targets in the EV condition were tested using the P_1 sentence frame and P_2 sentence frame once each. The participants were not alerted to these specifics. Also unbeknownst to the participants, the test was structured in three waves. These waves corresponded to the original order of the three waves from the encoding phase (described in the Appendix). We did this to approximately equate the retention intervals among the experimental conditions because the retention interval can modulate the effects of spacing (Glenberg, 1976; Cepeda et al., 2006). However, the sentences were not tested in their exact original serial positions. Instead we pseudo-randomized the within-wave test trial order with the constraint that any two test trials over the same target had to occur at least three or more trials apart.

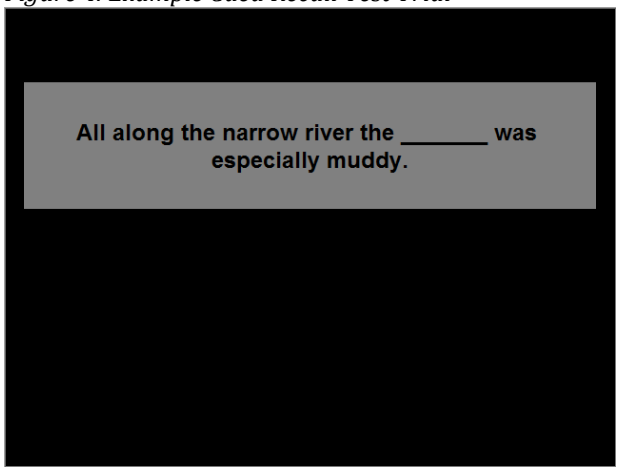
Procedures.

Participants were told that their memory of all previously seen sentences would be tested one at a time. In addition, the sentences would not be tested in the same order in which they were originally presented. They were also told that many of the sentences would be tested twice. Therefore, they should not be alarmed and should just attempt to recall the correct word again. On each test trial, they were presented with the sentence—with the target word deleted—and their task was to recall the missing word out loud. The width of the blank was made the same for

every sentence. The participants were made aware of this and they were also told that sometimes the correct word they needed to recall might not always fit within the blank. They were allotted 7.5 sec for each trial. After which, a crosshair appeared for .5 sec and then the computer automatically advanced to the next trial. After these instructions, participants were given an example of a cued recall trial (shown in Figure 4) with the sentence used in the beginning "All along the narrow river the _____ was especially muddy." They were then reminded that the correct word was *bank*.

After these instructions, the participants did a practice phase of 6 trials of items from the primacy buffer. The practice trials and their order were the same for every participant. After this practice phase, any remaining questions were answered and then the participant proceeded to the full test. The experimenter wrote down all of their responses to reduce the response demands on the participants. No feedback was given during the test. After the test, participants were debriefed.

Figure 4. Example Cued Recall Test Trial



Sentence frame for 7.5 sec

Fig. 4 Notes: Depicted above is the example of a cued recall trial shown to all participants before the test began. After each sentence frame, a crosshair appeared in the middle of the screen for 500 msec (not depicted).

RESULTS

Study-Phase Data

To determine study-phase retrieval accuracy, we treated the repetition judgment data as a recognition memory task and conducted the appropriate analyses. Correctly identified repetitions were coded as hits. Correct identifications of the first presentation of a repeated word or the only presentation of a 1P word were both coded as correct rejections. Incorrectly classifying either kind of trial as repeated item was coded as a false alarm. Finally, incorrectly classifying a repeated item as new was coded as a miss. Trials for which participants failed to respond were excluded from all analyses. When necessary, degrees of freedom were corrected for violations of sphericity.

First Occurrence Detection Rates (Correct Rejection Rates)

Our first analysis of the study-phase retrieval data examined the extent to which our participants could accurately detect the first occurrence of a repeated word or the only occurrence of a 1P word during the study phase. As such, we analyzed the correct rejection rates for both groups. Recall that we also collected confidence ratings during encoding. Thus they were included as a factor as well. We conducted a 2 x 2 (Group: Similar or Different by Confidence: Low or High). There was a significant main effect of group, $F(1, 70) = 11.67$, $MSE = .14$, $p = .001$, $\eta_p^2 = .14$. Correct rejection rates were higher in the Similar group than the Different group. There was also a significant main effect of confidence level, $F(1,70) = 39.12$, $MSE = 2.83$, $p < .001$, $\eta_p^2 = .36$. Correct rejection rates were higher for high-confidence ratings than low-confidence ones. However, these two main effects were qualified by a significant interaction, $F(1,70) = 5.75$, $MSE = .42$, $p = .019$, $\eta_p^2 = .08$. Probing the interaction revealed that the two groups did not significantly differ for low-confidence ratings [$t(70) = 1.23$, $p = .222$], but

that, among high-confidence ratings, correct rejection rates were higher in the Similar group than the Different group, $t(70) = 2.92$, $p = .005$. A summary of the results is shown in Table 1.

Table 1. First Occurrence Detection Rates During the Study Phase

EV Type	Low-Confidence	High-Confidence
Similar	.23 (.16)	.62 (.25)
Different	.27 (.14)	.45 (.25)

Table 1 Notes: means and standard deviations (in parentheses).

Two conclusions can be gathered from these data. First, the overall higher correct rejection rates among the high-confidence ratings suggests that the confidence ratings were used with a degree of metacognitive accuracy. That is, when our participants reported being highly confident that a given word has appeared only once thus far in the experiment, they were indeed more accurate than when the same judgment was made with low confidence. This result shows that the confidence ratings had some utility. Secondly, the interaction between group and confidence level demonstrates that any differences between the encoding conditions might be restricted to higher levels of confidence. Perhaps because performance at higher confidence levels is based on a greater level of memorial information that is further from the decision boundary between repeated vs. new. Consequently, the results of the high-confidence data might simply be less contaminated by noisier inputs. This is an issue we will revisit in the following section on study-phase retrieval accuracy.

Study-Phase Retrieval Accuracy (Hits – False Alarms)

The next set of analyses evaluated our hypotheses about the effects of lag and associative distance on study-phase retrieval. To assess study-phase retrieval accuracy, we did an analysis of recognition memory accuracy for repeated words. Our measure of accuracy was hits (correctly identified repetitions) minus false alarms (words incorrectly judged as repeated). We

first conducted a 2 x 2 x 3 x 2 (Group: Similar or Different by Repetition Type: EC or EV by Lag: 0, 6, or 18 by Confidence: Low or High) ANOVA. We first evaluated the effect of confidence to determine if separate analyses should be conducted for each confidence level. Recall that, in the analysis of correct rejection rates, the pattern of results depended on confidence. Any similar effect of confidence in the current analysis would warrant separate analyses for each confidence level. Moreover, any interaction between confidence and another variable would certainly warrant separate analyses.

Preliminary Analysis Including Both Confidence Levels. There was a very large main effect of confidence, $F(1,70) = 2{,}159.64$, $MSE = 192.20$, $p < .001$, $\eta_p^2 = .969$. Study-phase retrieval accuracy was overall higher for high-confidence ratings than low-confidence ones. Once again demonstrating that our participants used the confidence ratings with a degree of metacognitive accuracy. More importantly, the effect of confidence emerged in 3 interactions. There were two-way interactions between confidence and repetition type [$F(1,70) = 30.32$, $MSE = .75$, $p < .001$, $\eta_p^2 = .30$] and confidence and lag, $F(2,140) = 28.30$, $MSE = .46$, $p < .001$, $\eta_p^2 = .29$. There was also a 3-way interaction between confidence, repetition type, and lag, $F(2,140) = 17.86$, $MSE = .22$, $p < .001$, $\eta_p^2 = .20$. This 3-way interaction alone mandates a partitioned analysis because it means the 2-way interactions between repetition type and lag differed by confidence level. All hit and false alarm rates for both confidence levels are shown in Table 2. As can be seen in Table 2, low-confidence responses were overall much less frequent than high-confidence ones. Moreover, low-confidence hits were very infrequent—which would thus prohibit further analyses of metrics based on them (such as accuracy). As such, the remaining analyses will focus exclusively on the results of the high confidence ratings to evaluate our hypotheses about the effects of lag and associative distance on study-phase retrieval.

Table 2. Study-Phase Retrieval Hit & False Alarm Rates

Group	Repetition Type	Low-Confidence Hit Rates			FA Rate	High-Confidence Hit Rates			FA Rate
		Lag 0	Lag 6	Lag 18		Lag 0	Lag 6	Lag 18	
Similar	EC	.01 (.04)	.03 (.07)	.05 (.10)		.95 (.09)	.96 (.09)	.93 (.13)	
	EV	.01 (.05)	.03 (.06)	.14 (.15)		.95 (.09)	.91 (.14)	.78 (.16)	
					.11 (.10)				.03 (.04)
Different	EC	.02 (.05)	.04 (.08)	.04 (.08)		.96 (.09)	.96 (.10)	.94 (.11)	
	EV	.04 (.08)	.08 (.13)	.12 (.14)		.91 (.14)	.86 (.17)	.77 (.21)	
					.19 (.13)				.08 (.08)

Table 2 Notes: EC = encoding constancy, EV = encoding variability. FA = false alarm. Cell values are means and standard deviations (in parentheses).

High-Confidence Responses Only. The next analysis was restricted to high-confidence responses and entailed a 2 x 2 x 3 (Repetition Type: EC or EV by Group: Similar or Different by Lag: 0, 6, or 18) ANOVA. The main effect of lag was significant, $F(2,140) = 24.88$, $MSE = .28$, $p < .001$, $\eta_p^2 = .26$. Only the linear contrast of this main effect emerged as significant, $F(1,70) = 45.70$, $MSE = .56$, $p < .001$, $\eta_p^2 = .40$. This indicates that, overall, study-phase retrieval accuracy decreased proportionately with lag. The main effect of repetition type was also significant, $F(1,70) = 29.80$, $MSE = .64$, $p < .001$, $\eta_p^2 = .30$. Study-phase retrieval accuracy was overall higher for items in the EC condition than EV conditions. The main effect of group was also significant, $F(1,70) = 5.97$, $MSE = .39$, $p = .017$, $\eta_p^2 = .08$. Of the two groups, study-phase retrieval accuracy was overall higher in the Similar group than the Different one. There was also a significant interaction between lag and repetition type, $F(2,140) = 18.49$, $MSE = .18$, $p < .001$, $\eta_p^2 = .21$.

To evaluate this interaction, we first separately examined the effect of lag within each repetition type. There was no significant effect of lag for EC repetitions, $F(2,142) = 2.00$, *MSE*

= .01, p = .139. There was, however, a significant, linear effect of lag for EV repetitions, $F(1,71)$ = 64.48, MSE = .87, p < .001, η_p^2 = .48. This indicates that, for EV repetitions, study-phase retrieval accuracy decreased with lag. Finally, we also evaluated the second component to the lag by repetition type interaction: the lag-by-lag differences between EC and EV repetitions. We did this by calculating the confidence intervals of the differences between the EC and EV conditions at each lag. At lag 0, the confidence interval of the EC-EV difference was (-.01 to .06). At lag 6, the confidence interval of the difference was (.01 to .09). At lag 18, the confidence interval of the difference was (.12 to .20). The confidence interval at lag 0 encompasses 0, thus indicating no difference between EC and EV at that lag. The interval at lag 6 did not encompass 0, indicating significantly higher accuracy in the EC condition at that lag. The same was true for lag 18. In addition, the confidence intervals of the differences for lags 6 and 18 do not overlap, indicating that the EC-EV difference at lag 18 was significantly different from that of lag 6. Finally, none of the remaining interactions were significant. The interaction between repetition type and group was non-significant, $F(1,70)$ = 2.34, MSE = .05, p = .131. For all other interactions, F was < 1.

In sum, the above results show that, although overall study-phase retrieval numerically declined with lag for both the EC and EV conditions, this decline was only significant in the EV conditions. The EV and EC conditions did not differ at lag 0, but did differ at lags 6 and 18. Of the two groups, study-phase retrieval accuracy was overall higher in the Similar group than the Different one. In addition, the rate of decline across lags did not significantly differ between the two EV types (although there was a numerical trend towards a more rapid decline in the EV Different condition). Regardless, these results show that lag and associative distance did impair study-phase retrieval. The results are displayed in Figure 5.

Figure 5. Study-Phase Retrieval Accuracy for High-Confidence Ratings

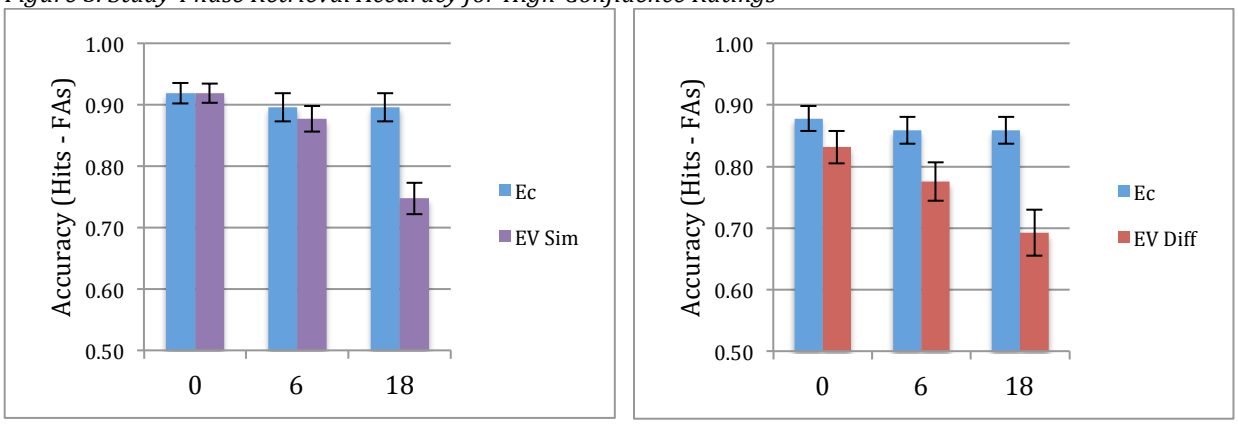


Fig. 5 Notes: Error bars indicate +/- 1 standard error of the mean. EC = encoding constancy, EV = encoding variability. FAs = False Alarms.

Cued Recall Data

Cued Recall OR Memory. Our first set of analyses of the cued recall data focused on memory for the target word. As such, we analyzed cued recall OR memory (Ross & Landauer, 1978). This was the probability that the correct target word was recalled in response to either the P_1 *or* the P_2 sentence frame. This metric allowed for an equitable and sensitive comparison of target word memory across all conditions.

We first compared OR memory for 1P words between those assigned to the Similar and Different groups. Note that memory for the 1P (and EC) targets was in fact tested twice; thus permitting an OR memory analysis of them. We found no significant difference in 1P OR memory between the groups, $t(70) = .95$, $p = .347$. These results are displayed in Table 3. Next we analyzed OR memory for the repeated items using a 2 x 2 x 3 (Repetition Type: EC vs. EV by Group: Similar vs. Different by Lag: 0, 6, or 18) ANOVA. There was a significant main effect of repetition type, $F(1,70) = 11.68$, $MSE = .27$, $p = .001$, $\eta_p^2 = .14$. Overall, OR memory

was higher for EV items than EC ones. There was no main effect of group, $F < 1$. There was a significant main effect of lag, $F(2,140) = 5.44$, $MSE = .13$, $p = .005$, $\eta_p^2 = .07$. A test of the least significant difference revealed that, overall, OR memory for lag 0 zero items was significantly lower than OR memory for lag 6 items ($SE = .02$, $p = .007$) or lag 18 items ($SE = .02$, $p = .007$). The latter two, however, did not significantly differ ($SE = .02$, $p = .746$). Note, however, that the main effect of lag seems to be mostly driven by a spacing effect in 3 out of 4 repetition conditions; the EC conditions of both groups and the EV condition of the Different group. This might explain why the interaction between repetition type and lag approached but did not achieve conventional significance levels, $F(2,140) = 2.20$, $MSE = .05$, $p = .115$. However, we do have an *a priori* motivation to test for an EV > EC difference at lag 0. When collapsing across both EV types, we did find a significant EV > EC difference at lag 0, $t(71) = 3.14$, $p = .002$.

Next, there was a significant interaction between repetition type and group, $F(1,70) = 6.11$, $MSE = .14$, $p = .016$, $\eta_p^2 = .02$. A follow-up analysis splitting the data by group revealed that there was an effect of repetition type for the Different group [$F(1,35) = 15.49$, $MSE = .40$, $p < .001$, $\eta_p^2 = .31$], but not the Similar group, $F < 1$. Thus, there was statistically an overall EV > EC difference in the Different group, but not the similar group. Finally, neither of the remaining interactions were significant (both F's < 1). The results of the cued recall OR memory analysis are displayed in Figure 6.

Table 3. Cued Recall OR Memory for 1P Items

	Similar Group	Different Group
1P OR Memory	.64 (.16)	.60 (.15)

Table 3 Notes. Cell values are means and standard deviations (in parentheses).

Figure 6. Cued Recall OR Memory for Repeated Items

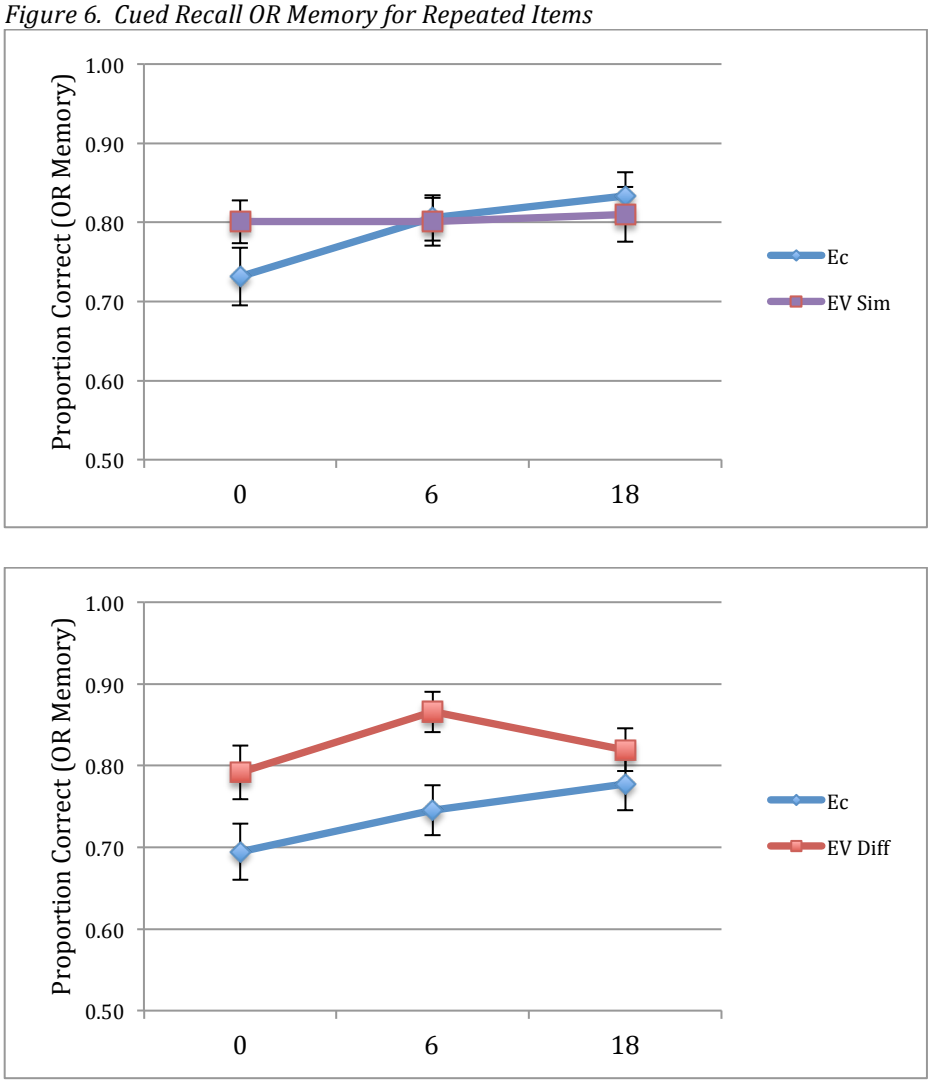


Fig. 6 Notes: Error bars indicate +/- 1 standard error of the mean. Lag is plotted along the x-axis. EC = encoding constancy, EV Sim = encoding variability in the Similar condition, EV Diff = encoding variability in the Different condition.

Chunking Rates. The next set of analyses examined the extent to which lag and associative distance affected the chunking of information across presentations of a target word. To estimate the amount of chunking for both EV types at each lag, we calculated a chunking index (proposed by Bellezza & Young, 1989, Exp. 3). This index had two inputs: 1) a measure

of information recalled from both presentations of a target word and 2) an estimate of the baseline recall rate that would be expected if recall were based on two independent memory traces. The measure of information recalled from both presentations was cued recall AND memory—which was the proportion of target words for which a participant correctly recalled the word in response to the first recall cue (i.e., the first sentence frame) *and* the second recall cue (i.e., the second sentence frame). The independence baseline was the probability of recalling a truly independent memory trace (e.g., a 1P item) twice (i.e., 1P OR memory squared). Thus, the chunking index was: cued recall AND memory – (1P OR memory)^2. This index was computed for all 3 lags of the EV conditions for all subjects[4]. Note that this metric would be uninterpretable for the EC items because the AND metric is ambiguous for them. The recall cues (i.e., the sentence frames) are identical for P_1 and P_2 in the EC condition. Therefore, there is no way of knowing whether a participant recalled a given target word from P_1, P_2, or both. As such, we only computed a chunking index for the EV conditions.

We analyzed the chunking index results with a 2 x 3 (EV Type: Similar vs. Different by Lag: 0, 6, or 18) ANOVA. There was a marginally significant main effect of lag, $F(2,140) = 3.03$, $MSE = .12$, $p = .051$. In general, chunking levels trended towards increasing with lag. There was, however, a significant main effect of EV type, $F(1,70) = 33.29$, $MSE = 1.75$, $p < .001$, $\eta_p^2 = .32$. Chunking levels were overall higher in the Similar condition than the Different one. Finally, lag and EV type did not interact, $F < 1$. This indicates that the overall advantage of

[4] We employed this metric instead of the independence metric used by Slamecka and Barlow (1979). The Bayesian metric used by Slamecka and Barlow (1979) required that the recall rates from one of the presentations be in the denominator, even in the equation's reduced form. One limitation of this is that this independence estimate can be undefined if the recall rate for the presentation used in the denominator is zero. As such, the metric employed by Slamecka and Barlow (1979) would be useless if the P_1 or P_2 recall rate for a given subject was zero. The metric used by Bellezza and Young (1989, Exp. 3) did not suffer from this limitation.

the Similar condition did not significantly change with lag. The results of the chunking analysis are shown in Figure 7. Also, cued recall AND levels are shown in Table 4.

Table 4. Cued Recall AND Memory

	Lag		
EV Type	0	6	18
Similar	.56 (.21)	.60 (.23)	.63 (.26)
Different	.32 (.20)	.42 (.28)	.38 (.25)

Table 4 Notes. Cell values are means and standard deviations (in parentheses).

Figure 7. Chunking Levels by EV Type and Lag

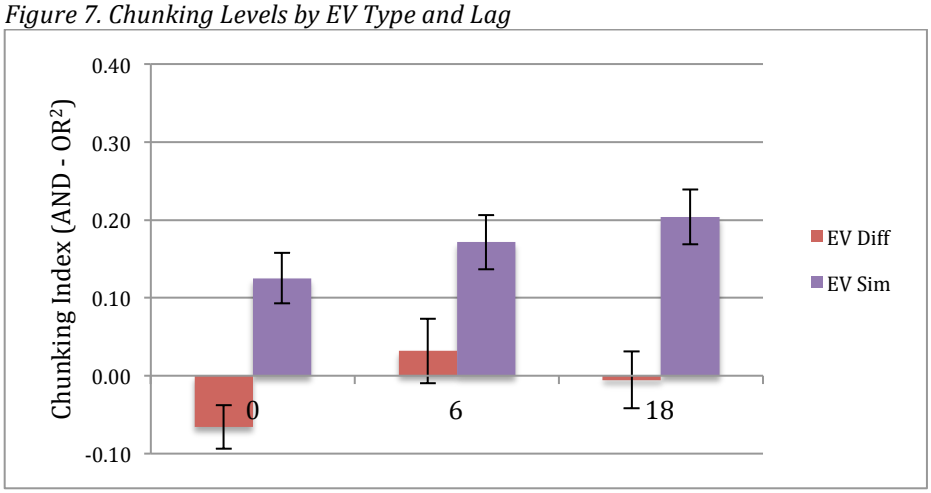


Fig. 7 Notes: Error bars indicate +/- 1 standard error of the mean. EV Sim = encoding variability in the Similar condition, EV Diff = encoding variability in the Different condition.

Interim Summary & Discussion

Study-Phase Retrieval. We restricted our primary analyses to high-confidence responses because low-confidence responses were so infrequent. Our preliminary analyses including data from both confidence levels revealed large differences between the low-confidence and high-confidence data. Also, the difference in accuracy between the two levels showed the utility of our decision to collect confidence data (and also indicated that the ratings were used with some

metacognitive accuracy). Moreover, our preliminary analyses also justified a subsequent exclusive focus on the high-confidence responses to evaluate our hypotheses.

We hypothesized that study-phase retrieval would decline as a function of lag and associative distance. We believed that study-phase retrieval would decline with lag for the EC and EV conditions of both groups—but more so for the EV conditions. This is because the EV conditions have the added effect of associative distance impairing study-phase retrieval. In addition, study-phase retrieval in the Different EV condition should decline more rapidly than in the Similar EV condition because the former entails a larger change in associative distance than the latter. In short, we found results in support for many of these hypotheses. However, some of the specific interactions we hypothesized did not fully materialize. We did find a negative effect of lag on study-phase retrieval, however it only emerged as significant in the EV conditions, not each group's respective EC condition (although there was a subtle numerical trend towards a decline in both groups' EC conditions). We also found that associative distance impaired study-phase retrieval because it was overall lower in the Different group. Although there was a trend towards a faster decline in the Different EV condition than the Similar EV condition, the hypothesized interaction was not significant. Nonetheless, the results showed that increases in lag and associative distance do impair study-phase retrieval.

Some other results suggest that the effect of associative distance might have also been more systemic than we anticipated. For one, first occurrence detection rates (i.e., correct rejection rates) were overall higher in the Similar group than the Different group (at least for the more frequent category of high-confidence responses). With regard to study-phase retrieval accuracy, we also found a similar overall Similar > Different advantage. Recall that the type of EV did not interact with lag as we had hypothesized. Instead, only the main effect of Similar >

Different emerged as significant. Note that these results indicate that the deviance of the encoding variation had episode-wide—as opposed to within-condition—effects on one's ability to detect first occurrences and repetitions. This could be due to an overall increase in interference in the Different EV condition; which would explain why some of the hypothesized interactions did not materialize. This is an issue we will revisit in the General Discussion, but it does not wholly undermine our ability to make inferences about the effects of lag and associative distance on study-phase retrieval (and later memory).

Cued Recall. We hypothesized that lag and associative distance would also interact with cued recall rates. We hypothesized that the EC condition of each respective group would exhibit a prototypical spacing effect in cued recall where recall increases as a function of lag. For both EV types, we hypothesized that EV would exceed EC at lag 0 and that this difference would dissipate with lag. In addition, we hypothesized that recall rates would decline faster with lag in the Different EV condition due the effects of more heavily impaired study-phase retrieval on later memory.

The EC condition of both groups did exhibit a positive effect of lag on recall, replicating a prototypical spacing effect on recall. However, the hypothesized interaction between lag and repetition type did not emerge. However, we did have *a priori* reasons to specifically find a lag 0 advantage in the EV conditions. This difference did emerge as significant when collapsing across both EV types. One additional, surprising result was that an EV advantage persisted until lag 6 for the Different EV condition. One possibility is that these two effects, the mitigated lag 0 advantage and the precarious lag 6 advantage, were a reactive side effect of our overt assessment of study-phase retrieval. This is an issue that we will address in Experiment 2.

Chunking Rates. We hypothesized that chunking rates are dependent on study-phase retrieval success. Therefore, chunking rates should peak at the shortest lag, lag 0, and in the less deviant encoding variation, the Similar EV condition. As such, chunking rates in the EV conditions should be maximal where study-phase retrieval is maximal (i.e., at lag 0 in the Similar EV condition). We only found support for one of those hypotheses. The main effect of EV type indicated that chunking rates were overall higher in the Similar EV condition. There was a marginally significant main effect of lag but, contrary to our hypotheses, it trended towards an increase in chunking with lag. Therefore, we also did not find that chunking levels were maximal in the hypothesized intersection of conditions. It is possible that this was also a reactive side effect of our overt assessment of study-phase retrieval. Perhaps, when study-phase retrieval is not overtly required, the peak of the chunking function will be pushed farther towards the left as was hypothesized. This is another issue that we will address in Experiment 2.

One possible reason for the trend towards a positive effect of lag on chunking could be that study-phase retrieval does promote chunking, but that it interacts with the frequency of retrieval (and possibly the difficulty of it). In an incidental learning paradigm, devoid of any overt assessment of study-phase retrieval, study-phase retrieval is assumed to be more spontaneous and less frequent. It is also presumably more likely to occur at shorter lags (e.g., lag 0). In a paradigm such ours, study-phase retrieval is presumably less spontaneous and more frequent. It could also be that successful study-phase retrievals at longer lags are more potent chunking agents, but because such retrievals occur at a lower frequency in incidental learning paradigms, their effects on the results are offset by the effects of the more frequent lag 0 study-phase retrievals. Conversely, in a paradigm such as our Experiment 1, successful study-phase retrievals at longer lags are more frequent, thereby contributing their more potent effects more

often to the results. As such, the peak of the chunking function would be shifted farther towards the right. If so then, in an incidental learning paradigm, with no overt assessment of study-phase retrieval (such as the following Experiment 2), the peak should be shifted farther towards the left.

Experiment 2 – Omitting Repetition Decisions

Experiment 1 assessed the effects of lag and associative distance on overt study-phase retrieval to see if these effects related to later memory. This was necessary given that, although there is some evidence from the literature that study-phase retrieval is the proximate cause of the effects of encoding variability on memory, no studies on encoding variability have assessed it. However, we recognize that one reason that overt assessments of study-phase retrieval are rare in the literature is because such assessments might have reactive effects on memory. Consequently, some of the results of the first experiment may have been reactivity effects. Therefore, Experiment 2 was conducted to determine the extent to which the results of Experiment 1 were due to reactivity. Note that we do not concede that our results were entirely due to reactivity effects. In essence, Experiment 1 only turned up the gain on variable already thought to be at play in incidental learning paradigms. It did not turn on a variable thought to be absent from them. Hence Experiment 2 is really assessing the extent to which the design of Experiment 1 enhanced processes (e.g., study-phase retrieval) that routinely contribute to memory performance.

Experiment 2 retained many of the same design features of Experiment 1. The materials and memory test were completely the same. The critical change was removing the overt study-phase retrieval assessment. However, we also had to utilize an orienting task that would serve as a comparable "control condition" for Experiment 1. We required an orienting task that drew the

participant's attention to the target word (as was done in Experiment 1), but did not overtly require them to compare the target on the current trial to a previous presentation of it. We also required an orienting task that invited semantic processing because our materials and the encoding variations were semantic in nature. We also needed an orienting task that could be easily mapped onto a 4-option rating scale in case the idiosyncrasies of the decision processes associated with such a scale affected the results. The participants of Experiment 2 would also still read the sentences in isolation for 7.5 sec (like the participants of Experiment 1), but now the 3.5 sec period following that would be replaced with a 4-choice semantic orienting task instead of a 4-choice repetition decision.

Piloting. During initial piloting for Experiment 2, we used 4-choice pleasantness ratings as an orienting task (very unpleasant to very pleasant). This is because pleasantness ratings have been shown to increase semantic processing relative to other non-semantic orienting tasks (Hyde & Jenkins, 1973). Also, unlike Experiment 1, we told participants of the upcoming memory test. We feared there would be floor effects in some conditions had we not told them. However, this approach with these two combined factors (pleasantness ratings and intentional encoding) resulted in ceiling effects in most conditions. Therefore, in our next wave of piloting, we shifted to completely incidental learning and did not warn participants of the upcoming memory test; although we still retained the pleasantness rating task. However, this approach still produced ceiling effects in most conditions. Finally, we retained the incidental learning setting, but shifted to a less semantic orienting task. We chose frequency ratings (i.e., rating the extent to which a word is used in daily speech). We chose this task because it has been shown to have intermediate effects on semantic processing. It improves memory and semantic processing more than a non-semantic task such as counting the number of e's and g's in a word, but not as much

as making pleasantness ratings (Hyde & Jenkins, 1973). This approach resulted in intermediate levels of memory performance. As a result, we employed it for the duration of Experiment 2.

METHODS

Participants.

The participants met the same inclusion criteria as those from Experiment 1. There were 72 participants total. Some were introductory psychology students given course credit in exchange for participation, and some were young adults paid $10 in exchange for participation. Like Experiment 1, we first calculated study-phase response omissions to see if anyone met outlier criteria for exclusion and replacement. We excluded and replaced anyone whose study-phase response omissions were more than 2 standard deviations above the mean (the resulting cut-off was 5 omissions in this experiment as well). Five participants had to be replaced (3 in the Similar group and 2 in the Different group). Thus, our final sample, including the replacements, is based on 72 subjects who omitted responses on fewer than 5 study-phase trials.

Design.

All aspects of the current experiment's design were identical to Experiment 1 except we did not assess study-phase retrieval. As such, we also did not collect confidence ratings during the study-phase and confidence was not included as a factor in our analyses.

Materials.

The materials were identical to those used in Experiment 1. The exact same study and test lists were used, only the orienting task was replaced.

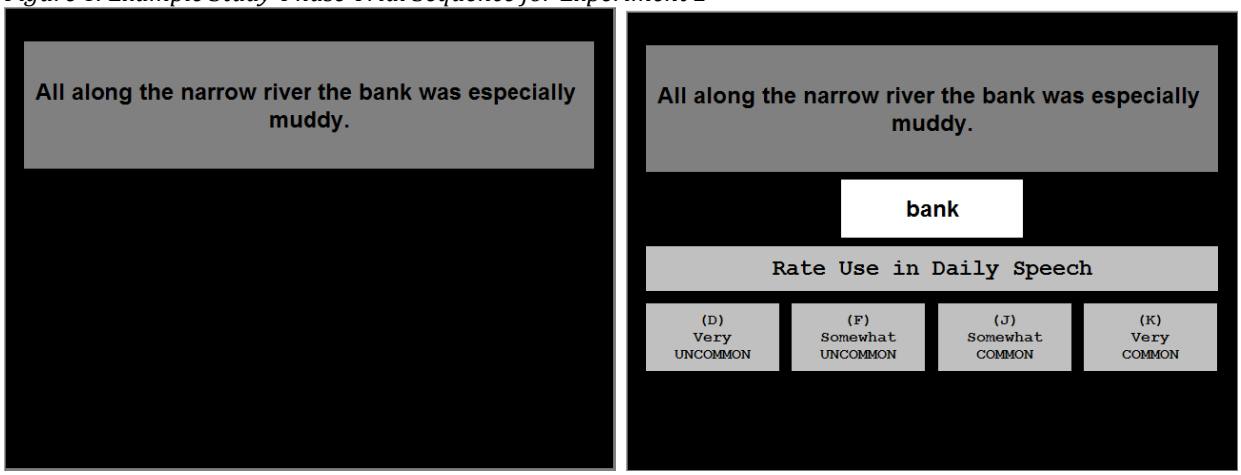
Procedures.

Like Experiment 1, participants in the current experiment were not told about the upcoming memory test. They were also given all of the same instructions before the study and

test phases, except for those regarding the orienting task. Participants in the current experiment were told that they were to judge how often the word was used in daily speech. Their judgment was to be based on how often they used the word as well as how often they heard it used by others. They rated the word from being very uncommon to very common using the D, F, J, and K keys in ascending order. (An example study-phase trial is shown in Figure 8).

If a participant asked whether or not they should constrain their judgment to the meaning in the current sentence, we told them that they should. For example, if shown a sentence using the river-related meaning of bank, they should base their judgment on that meaning. If shown a sentence using the money-related meaning of bank, they should base their judgment on that meaning. Note that participants were only told this if they asked about it beforehand. We took this approach to avoid sensitizing all of our participants to homonymy before the experiment. Also note that, because this part of the instructions comes up before they are exposed to the words and meanings in their condition, this kind of inquisitiveness necessarily ended up being randomly assigned to the 2 conditions. Regardless, the majority of our participants did not ask about these aspects of the words and judgment task.

Figure 8. Example Study-Phase Trial Sequence for Experiment 2



Sentence alone for 7.5 sec

Frequency rating over target word for 3.5 sec

Fig. 8 Notes: Depicted above is the example of a study-phase trial shown to all participants before the experiment began. After the frequency rating portion of a trial, a crosshair appeared in the middle of the screen for 500 msec (not depicted).

RESULTS

Cued Recall Data

Cued Recall OR Memory. Our first wave of analyses of the cued recall data focused on comparing memory for the target word. As such, we analyzed cued recall OR memory (a metric previously described in Experiment 1). We first compared OR memory for 1P words between those assigned to the Similar and Different groups. We found no significant difference in 1P OR memory between the groups, $t(70) = .04$, $p = .968$. These results are displayed in Table 5. Next we analyzed OR memory for the repeated items using a 2 x 2 x 3 (Repetition Type: EC vs. EV by Group: Similar vs. Different by Lag: 0, 6, or 18) ANOVA. There was a significant main effect of repetition type, $F(1,70) = 13.13$, $MSE = .23$, $p = .001$, $\eta_p^2 = .16$. There was no significant main effect of group, $F(1,70) = 1.69$, $MSE = .10$, $p = .198$. There was a significant main effect of lag, $F(2,140) = 5.75$, $MSE = .15$, $p = .004$, $\eta_p^2 = .08$. However, the main effects of repetition type and lag were qualified by a significant repetition type by lag interaction, $F(2,140) = 6.99$, $MSE = .17$, $p = .001$, $\eta_p^2 = .09$. We first evaluated the within-condition portion of this interaction by analyzing the effect of lag within each repetition type. For EC repetition types, there was a significant, linear effect of lag, $F(1,71) = 22.02$, $MSE = .61$, $p < .001$, $\eta_p^2 = .24$. As lag increased, so did OR memory. For EV repetition types, there was no significant effect of lag, $F < 1$. Next, we evaluated the between-condition portion of this interaction by analyzing the differences between EC and EV at each lag. OR memory for EV items was significantly higher than that for EC items at lag 0, $t(71) = 4.65$, $p < .001$. The two, however, did not significantly differ at lag 6 [$t(71) = .92$, $p = .361$] or lag 18 [$t(71) = .32$, $p = .750$]. Finally, none of the

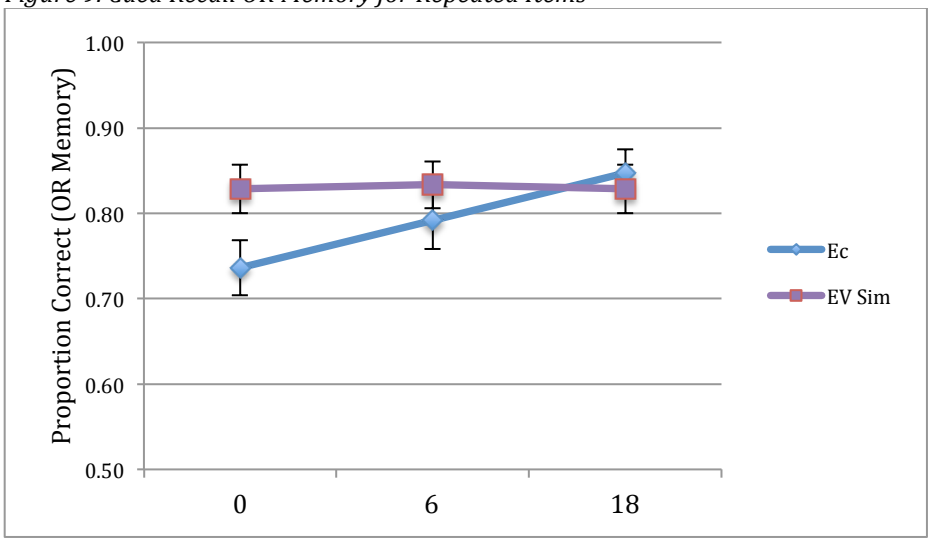
remaining interactions, (repetition type by group; lag by group; or repetition type by lag by group) were significant (all *F*'s < 1). The OR memory results for the repeated items are displayed in Figure 9.

Table 5. Cued Recall OR Memory for 1P Items

	Similar Group	Different Group
1P OR Memory	.67 (.16)	.68 (.16)

Table 5 Notes. Cell values are means and standard deviations (in parentheses).

Figure 9. Cued Recall OR Memory for Repeated Items



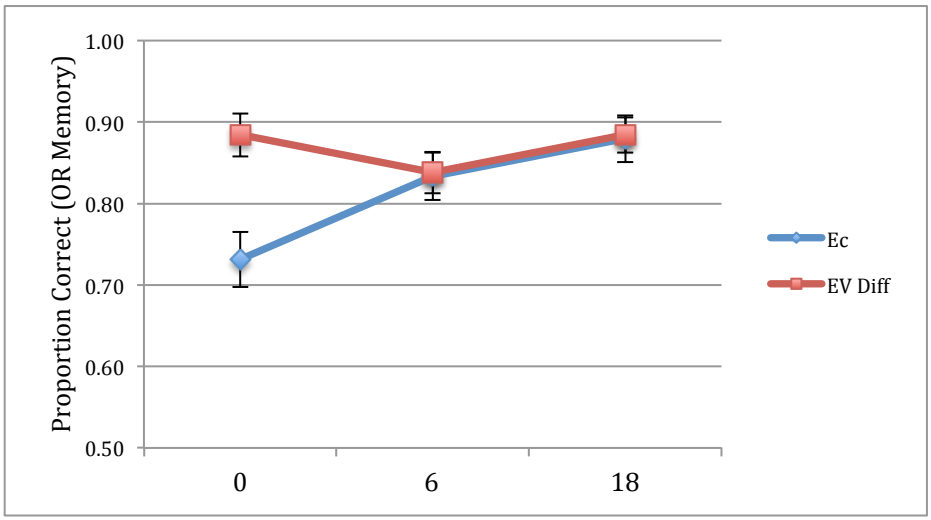


Fig. 9 Notes: Error bars indicate +/- 1 standard error of the mean. Lag is plotted along the x-axis. EC = encoding constancy, EV Sim = encoding variability in the Similar condition, EV Diff = encoding variability in the Different condition.

Chunking Rates. The next wave of analyses examined the extent to which lag and associative distance affected the chunking of information across presentations of a target word. Our chunking rate analysis was based on the same chunking index described in Experiment 1. We analyzed the chunking index results with a 2 x 3 (EV Type: Similar vs. Different by Lag: 0, 6, or 18) ANOVA. There was a significant main effect of EV type, $F(1,70) = 8.99$, $MSE = 1.16$, $p = .004$, $\eta_p^2 = .11$. Chunking rates were overall higher in the Similar EV condition than the Different one. There was no significant main effect of lag, $F(2,140) = 1.43$, $MSE = .06$, $p = .244$ There was also no significant interaction between EV type and lag, $F < 1$ (although there was a trend such chunking declined with lag in the Similar condition and remained low and flat in the Different one). The results of the chunking analysis are displayed in Figure 10. Also, cued recall AND levels are shown in Table 6.

Figure 10. Chunking Levels by EV Type and Lag

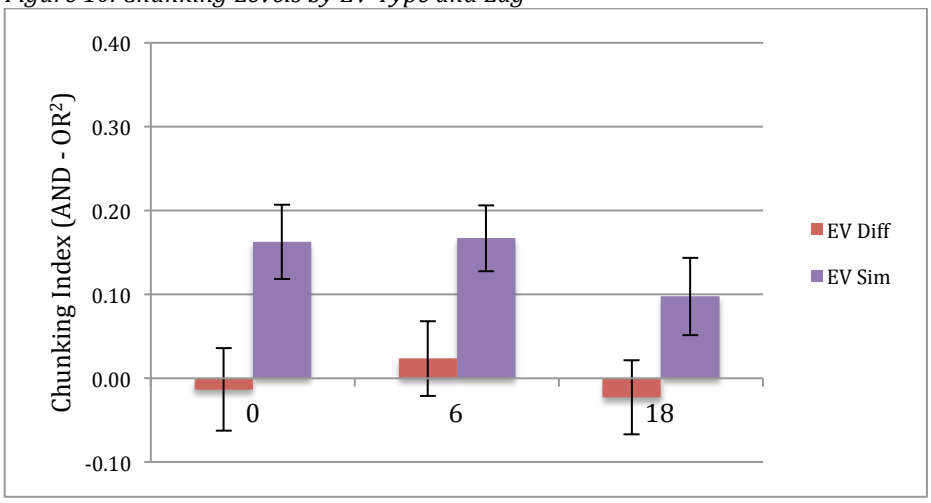


Fig. 10 Notes: Error bars indicate +/- 1 standard error of the mean. EV Sim = encoding variability in the Similar condition, EV Diff = encoding variability in the Different condition.

Table 6. Cued Recall AND Memory

	Lag		
EV Type	0	6	18
Similar	.64 (.25)	.65 (.23)	.58 (.22)
Different	.47 (.25)	.50 (.22)	.46 (.27)

Table 6 Notes. Cell values are means and standard deviations (in parentheses).

Interim Summary & Discussion.

Experiment 2 removed the overt assessment of study-phase retrieval to examine the extent to which our assessment in Experiment 1 was reactive. Also, to serve as an adequate control condition for Experiment 1, the 4-option repetition judgment was replaced with a 4-option semantic orienting task. We used a frequency judgment task, and incidental learning conditions, to achieve intermediate levels of semantic processing and memory performance.

Results. With the design of Experiment 2, we found a lag by repetition type interaction on cued recall OR memory. EV exceeded EC at lag 0, and this difference dissipated to non-significance at the other two lags. We also found a spacing effect on OR memory for EC items but not EV ones. These effects did not interact with group (although there was a numerical trend towards a larger lag 0 advantage in the Different EV condition). Note that these results are similar to what has been found in other studies on EV and spacing. We found the prototypical EV lag 0 advantage—which dissipated at longer lags. In Experiment 1, there was a significant EV lag 0 advantage but the interaction between lag and repetition type did not materialize. Also, in Experiment 2, we found a prototypical spacing effect for EC items but not EV ones—another asymmetry often found in the literature. These results are important because they show that,

without any overt assessment of study-phase retrieval, our paradigm produces effects typical of the literature.

With regards to the chunking index results, we found an overall effect of EV type. Chunking rates were higher in the Similar EV condition than the Different one. There was no significant effect of lag and no significant lag by EV type interaction. We hypothesized that, in Experiment 2, the peak of the chunking function would be pushed farther to the left. There was a numerical trend towards this effect but the requisite effects did not materialize. Therefore, the results of the current experiment only provided suggestive support for our hypothesis about the effects of Experiment 1's methods on chunking rates.

Comments on Reactivity. With the completion of Experiment 2, we can determine the extent to which the methods of Experiment 1 were reactive. Keep in mind that some reactivity was expected. In a sense, Experiment 1 was only thought to turn up the gain on a variable already present in incidental learning paradigms; not turn on a variable thought to be absent from them. As such, the differences between Experiments 1 and 2 should be differences of degrees, not types. A comparison of the cued recall results from both experiments largely supports this notion (although there were some exceptions). For cued recall OR memory, both experiments numerically exhibited a lag by repetition type interaction where EV exceeded EC at lag 0, but then this difference dissipated with lag. However, this interaction was only significant in Experiment 2. Nonetheless, both experiments exhibited a similar lag by repetition type interaction; it simply emerged as significant in one experiment but not the other. Therefore, in this respect, the two experiments only differed in degree, not type. Also, both experiments showed a spacing effect on OR memory for EC items. In Experiment 2, this effect emerged in the presence of a prototypical lag by repetition type interaction. In Experiment 1, there was

statistically an overall positive effect of lag on OR memory. Three out of the 4 repetition conditions showed positive effects. The respective EC conditions of both groups showed an effect (as well as the Different EV condition). This is a minor issue we will revisit in the General Discussion. Regardless, we found prototypical spacing effects for EC items in both experiments. This effect was accompanied with an interaction in Experiment 2 but not in Experiment 1.

With respect to chunking rates, both experiments showed an overall advantage for the Similar EV type. This outcome is somewhat consistent with a prediction of the chunking hypothesis, which is that, ceteris paribus, chunking rates should be higher for intermediate levels of EV as opposed to maximal levels of EV. Chunking rates for the Different EV type remained relatively flat and no different from zero across both experiments (another issue we will also revisit in the General Discussion). However, in neither experiment did we find a significant effect of lag or a significant lag by EV type interaction. Experiment 1 also showed a surprising trend where, at least for the Similar EV type, chunking increased somewhat with lag. Experiment 2 showed an opposite trend consistent with our original hypotheses. Although both effects were trends that did not statistically materialize, their presence warrants further discussion about the potential reactivity of Experiment 1's methods. Also, because the apparent asymmetry was restricted to the Similar EV type, we will restrict our discussion to it. Given that the between-experiment differences in chunking for this EV type resembles a crossover interaction, one might be tempted to conclude that Experiment 1 qualitatively changed the chunking process. However, a closer inspection of the results suggests that this is not the case. First, note that, across the two experiments, the chunking rates for this EV type are quite similar at lag 0 (Exp 1: M = .13, SD = .19; Exp 2: M = .16, SD = .27) and lag 6 (Exp 1: M = .17, SD = .21; Exp 2: M = .17, SD = .24). The most pronounced difference is at lag 18 (Exp 1: M = .20,

SD = .21; Exp 2: *M* = .10, *SD* = .28). Thus, it would appear that the difference between the two experiments was restricted to the longest lag. An exploratory cross-experiment comparison of these data points revealed a significant interaction between lag and experiment, $F(1, 70) = 4.46$, $MSE = .19$, $p = .038$, $\eta_p^2 = .06$. However, further probing of the interaction revealed that the difference between the two experiments at lag 18 only approached significance, $t(70) = 1.83$, $p = .071$. Nonetheless, these results provide suggestive support for our earlier proposals about lag-related differences in the mnemonic potency of study-phase retrieval and cross-experiment differences in study-phase retrieval frequency. The methods of Experiment 1 arguably increased the frequency of study-phase retrieval—especially at longer lags. If retrieval at these lags is also more mnemonically potent, then an experiment in which such retrievals are also more frequent will exhibit higher chunking rates at longer lags. Also note that these collective results indicate that what initially appeared to be a qualitative difference between the experiments actually appears to be a quantitative difference after closer inspection. Chunking rates were comparable at lags 0 and 6 across both experiments; they differed most at the longest lag, lag 18. Furthermore, this difference can be accounted for by theoretically sensible reasons that suggest both experiments relied on a common study-phase variable whose presence was only higher in one experiment than the other. Therefore, we can justify using the study-phase results of Experiment 1 to make inferences about the effects of study-phase retrieval on later in memory in incidental learning paradigms such as our Experiment 2 and other ones akin to it.

GENERAL DISCUSSION

The Current Study

The primary goal of the current study was to understand the mechanisms by which encoding variability affects long-term explicit memory. Several studies have observed EV

advantages under conditions that arguably facilitated study-phase retrieval—such as when repetition lags are short or when subjects have been repeatedly familiarized with the core materials before being exposed to encoding variations of them. Of the theories of the effects of encoding variability on long-term explicit memory, only the chunking hypothesis takes into account the role of study-phase retrieval (Young & Bellezza, 1982; Bellezza & Young, 1989). Study-phase retrieval has also been shown to modulate the spacing effect (Thios & D'Agostino, 1976). Given that spacing modulates the effects of encoding variability (Bellezza & Young, 1989; Thios, 1972), it is therefore likely that study-phase retrieval modulates the effects of encoding variability as well. Although Bellezza and Young found indirect evidence consistent with their chunking hypothesis, no study of encoding variability has directly assessed what Bellezza and Young assumed to be *the* proximate cause of EV effects: study-phase retrieval. The current study directly assessed study-phase retrieval to see how it relates to the effects of encoding variability. Our first experiment directly assessed study-phase retrieval accuracy with methods akin to a continuous recognition memory paradigm. We then removed this assessment in our second experiment to determine the extent to which our assessment in Experiment 1 was reactive. (As mentioned in the Interim Summary & Discussion of Experiment 2, the reactivity of our study-phase retrieval assessment was found to be minimal.) In addition, we also assessed the effects of lag and associative distance on study-phase retrieval and later memory. Ample evidence from the literature shows that these two factors modulate the effects of encoding variability on later memory, which suggests they are also potential modulators of study-phase retrieval. It is also worth noting that the effect of lag on EV has been studied very often but the effect of associative distance on EV has seldom been studied (for examples, see Bobrow, 1970; Hintzman et al., 1975, Exp. 2; Rowe, 1973, Exp. 1; Thios, 1972).

To examine these issues, we manipulated repetition type, lag, and the associative distance of the encoding variation. We selected homonyms as the target words so that we could maintain the same nominal stimuli and manipulate two extremes of encoding variability around them. All subjects had an encoding constancy control condition to allow us to examine the relative effects of encoding variability. The associative distance of the encoding variation was manipulated between groups and consisted of variations within the same meaning of the target word (i.e., the Similar group) or variations between completely different meanings of the target word (i.e., the Different group). These levels were selected to create a strong manipulation of associative distance and because these two levels of EV have been used frequently in the literature (albeit often in separate studies).

Furthermore, in order to induce the intended semantic context, we embedded the target words in sentence frames (as opposed to merely pairing them with semantically related cue words). This allowed for a more reliable inducement of the intended semantic context and ensured greater consistency of encoding across subjects. Using sentence frames helped reduce noise due to inter-subject and inter-item variability in establishment of the intended semantic context—variability that would have been more pronounced had we used more blunt instruments such as word pairs. Therefore, our study goes beyond many prior EV studies because we created a more stable manipulation of the intended semantic context. To our knowledge, very few studies in the EV literature have used sentence frames to manipulate encoding variations of target words (for a few examples, see Bobrow, 1970; D'Agostino & DeRemer, 1973; Dellarosa & Bourne, 1985, DeRemer & D'Agostino, 1974; Jacoby, 1972; Postman & Knecht, 1983, Exps. 1 & 2; Thios, 1972).

Hypotheses I: Study-Phase Retrieval. Based on the tenets of the chunking hypothesis, we hypothesized that lag and associative distance would be negatively correlated with study-phase retrieval accuracy. Study-phase retrieval accuracy should decrease with lag and be overall higher for EC items. Also, lag and repetition type should interact such that the EC > EV difference becomes larger with lag. We found results in support of these hypotheses. One remaining implication of the chunking hypothesis is that associative distance (i.e., EV type) should interact with lag such that study-phase retrieval accuracy should decline more rapidly for the Different EV type than the Similar one. Although there was a numerical trend towards this effect, we only found a main effect such that overall accuracy was higher in the Similar group than the Different one. (This issue is discussed at length in the following section). Nonetheless, the majority of the results for the study-phase retrieval data supported the predictions one should make based on the tenets of the chunking hypothesis.

Hypotheses II: Recall & Chunking Rates. Based on previous studies and the tenets of the chunking hypothesis, we hypothesized that the EV advantage on cued recall would be most pronounced at lag 0. We also hypothesized that this advantage would occur in parallel with maximal chunking rates on our chunking index. In addition, chunking rates should be higher for the intermediate EV type, the Similar condition. A crucial aspect of the chunking hypothesis is that memory advantages due to EV occur because components from multiple presentations of the target get chunked into one memory code. This chunking process is assumed to create a memory code richer than the mere sum of components from two independent presentations of a target. Therefore, a test of the chunking hypothesis required an index of chunking rates that took these assumptions into account. We employed a chunking index proposed by Bellezza and Young (1989, Exp. 3).

To examine our hypotheses about chunking rates, we refer to the results of Experiment 2—when study-phase retrieval was not overtly assessed. In short, we found that the EV advantage in cued recall did peak at lag 0 for both EV types. Chunking rates were also higher for the Similar EV type. However, chunking rates were only positively related to cued recall for the Similar EV type. They remained flat and no different from zero at all lags for the Different EV type. This result was unexpected. Study-phase retrieval success is thought to promote chunking rates, and although study-phase retrieval was lower for the Different EV type, it was not at chance.

These results suggest that associative distance might contribute more heavily to the effects of EV than previously thought. If study-phase retrieval was still well above chance for the Different EV type, but chunking rates for this EV type were overall lower (and at zero), then study-phase retrieval might not be sufficient for chunking. Perhaps, if the associative distance is too great, chunking is difficult and/or impossible. It might be difficult to chunk components under such conditions because 1) the large change in associative distance induces more interference and/or 2) it might also be impossible to chunk under such conditions because the components at each presentation are largely irrelevant to each other.

Support for the first explanation was found in the study-phase results of Experiment 1. Detection rates for first occurrences were lower in the Different group. Study-phase retrieval accuracy was also overall lower in this group. These results suggest that the extreme level of associative distance in this group had episode-wide effects—as opposed to within-condition effects—on detecting repetitions of target words. These episode-wide effects could be due to an overall increase in interference due to monitoring extremes in encoding variations. (This might also explain why we only found a main effect of group on study-phase retrieval and not the

hypothesized lag by EV type interaction.) There is also support for the second explanation offered above that attributes the results to irrelevance. For example, there is little in common between the river-related meaning of *bank* versus the money-related meaning of *bank*. As a result, there is little to be chunked regardless of whether or not one recognizes that *bank* has been repeated. These factors might explain why chunking rates were flat and no different from zero for the Different EV type in both experiments. Nonetheless, these results show that 1) extremes of encoding variability can have deleterious effects on memory and 2) that study-phase retrieval may not be sufficient for chunking. To test the latter hypothesis, future studies could try to equate study-phase retrieval between two levels of associative distance to see if differences in chunking rates still persist.

Given our latter explanation for the chunking results, one might question why one of our manipulations of EV type was so extreme? Keep in mind that we chose this kind of manipulation because 1) it creates the strongest manipulation of associative distance while maintaining the same nominal stimulus and 2) several studies of EV have used variations between meanings with homographs and homonyms (Bobrow, 1970; D'Agostino & DeRemer, 1973; Hintzman et al., 1975, Exp. 2; Johnston, Coots, & Flickinger, 1972; Maki & Hasher, 1975, Exp. 2; Rowe, 1973, Exp. 1; Slamecka & Barlow, 1979; Thios, 1972; Winograd & Raines, 1972, Exp. 1). In addition, other studies have used manipulations that created extreme changes in semantic context that are arguably on par with variations between meanings of homographs and homonyms. For example, some studies have manipulated EV by pairing a target word with unrelated cues on both presentations of the target such as *tomato – air* and *crown – air* (Bellezza & Young, 1989; Greene & Stillwell, 1995; Shaughnessy et al., 1974; Young & Bellezza, 1982, Exp. 3). In sum, several studies in the literature have used the same extreme manipulations of

EV used in the current study or used manipulations conceptually similar to it. As such, there was empirical motivation for examining both extremes of EV within one study.

Implications & Future Directions

Chunking Hypothesis. The results of the current study largely support the predictions one would make based on the tenets of the chunking hypothesis. EV advantages in cued recall were largest at lag 0 and chunking rates were higher for the intermediate EV type. EV also resulted in a decline in study-phase retrieval as lag increased and study-phase retrieval was overall lower for the more deviant EV type. Thus, preceding differences in study-phase retrieval related to later memory performance in theoretically predicted ways. However, the relationships were not entirely in line with the predictions of the chunking hypothesis.

For one, there were instances in which chunking was inversely related to study-phase retrieval. Recall that, in Experiment 1, study-phase retrieval, in the Similar condition, was lowest at lag 18 but chunking was highest at this lag. (However, in Experiment 2, there was the predicted mutual decline in study-phase retrieval and chunking across lag.) Nonetheless, this unique result of Experiment 1 suggests that the mnemonic potency of study-phase retrieval may modulate chunking rates. Furthermore, difficult study-phase retrievals may be more potent catalysts for chunking than easy retrievals. If so, future versions of the chunking hypothesis must take into account both retrieval success and difficulty. Future studies will have to employ a proxy measure of retrieval difficulty to see how this relates to chunking rates.

One of the current study's original goals was to analyze study-phase RTs as a proxy of retrieval difficulty. We also intended to examine any corresponding changes in P_1 and P_2 recall rates to see if their changes across lag behaved in theoretically predicted ways. However, feasibility limitations to our counterbalancing scheme prohibited analyses of both study-phase

RT and P_1 vs. P_2 recall. In order to have interpretable results of study-phase RTs and P_1 vs. P_2 recall, the sentences should ideally have been counterbalanced across the P_1 and P_2 positions. The current study only counterbalanced the target words across encoding condition (1P, EC, and EV), EV type (Similar vs. Different), and lag (0, 6, or 18). The sentences for the intended meaning of a target word were randomly assigned to P_1 or P_2. This design at least permitted us inferences about the above 3 factors. This design also resulted in a counterbalancing scheme requiring 72 subjects. A permissible analysis of study-phase RTs and P_1 vs. P_2 recall would require that we counterbalance all 4 sentences for a given word (2 for the first meaning and 2 for the second) across both presentations, all 3 encoding conditions, both EV types, and all 3 lags. That would create a counterbalancing scheme requiring 288 subjects. Given that we could not accommodate such a design, we decided to ultimately forgo the analyses of P_1 and P_2 recall. Thus, we also decided to forgo the analyses of the study-phase RT data because they were inextricably bound to our hypothesis tests about P_1 vs. P_2 recall dynamics.

Associative distance also had a larger influence on the results than predicted. Recall that study-phase retrieval accuracy was above-chance in all conditions, but chunking rates were at zero for all lags of the Different EV type. This suggests that study-phase retrieval is not sufficient for chunking—a tenet the original chunking hypothesis did not take into account. Our results further suggest that intermediate levels of EV are optimal for learning regardless of study-phase retrieval success.

Concept Learning and Transfer of Learning. The results of the current study might also help explain why, in some cases and not others, EV enhances other feats of memory like concept learning and transfer of learning. Some studies have found that EV enhances transfer of conceptual knowledge (Nitsch, 1977; di Vesta & Peverly, 1984) while others have not

(Dempster, 1987, Exps. 2—5). The current study would predict that EV enhances transfer when the encoding conditions facilitate chunking of the information from variously encoded concepts. Dempster (1987, Exps. 2—5) found that EV did not enhance transfer, but his design also arguably did not facilitate chunking because the lags were quite long (38 items) and no overt study-phase retrieval requirement was made. However, the studies by Nitsch (1977) and di Vesta and Peverly (1984) arguably facilitated chunking. Lag was not consistent or systematically varied in these studies, but their encoding conditions overtly required their subjects to practice retrieving the concept name-definition-example relationships. Sometimes the subjects were cued with a concept name and asked to provide the definition. Other times they were cued with a definition or example and asked to recall the concept name. Interestingly, Nitsch (1977) also found that participants exhibited less confusion (and optimal transfer) if, in the first half of the study, they practiced learning the concepts with examples in the original learning context (i.e., had a preliminary EC phase) and then moved on to practicing with examples in different learning contexts (i.e., had a follow-up EV phase). These combined results suggest that EV benefits concept learning and transfer under conditions that facilitate study-phase retrieval of the original learning context (and possibly study-phase retrieval across multiple new contexts). Moreover, the fact that EV can have such pronounced effects on transfer of conceptual knowledge suggests that it may more heavily affect memory for superordinate information. Nonetheless, these are tentative conclusions that will require further study.

Future Directions

Retrieval Processes & Chunking. Although the chunking index we employed in the current study provided a much-needed metric of chunking success, we implicitly made assumptions about its meaning. To interpret the metric as an index of chunking, we had to

assume that the dependence between P_1 and P_2 was created during encoding, and then that information was reinstated at retrieval in a way that aided recall. To test this assumption (at least the latter part of it) one should create an overt manipulation of P_1—P_2 proximity at retrieval to see if this modulates memory performance in ways predicted by the chunking hypothesis. This retrieval-based analysis of chunking would ideally present the cues devoid of the target. If one cue facilitates recognition of the other when the target is absent, it stands to reason that their underlying commonalities (i.e., the target) caused this. Moreover, if manipulating the proximity of the cues at retrieval modulates their mutual facilitation, then it also stands to reason that P_1—P_2 dependence is created during encoding, and that reinstatement of this information at retrieval facilitates memory performance.

To examine these retrieval factors, future studies could devise a recognition memory test where the cues are presented devoid of the targets. Cues from the study phase could be preceded by ones that were old and do or do not share the same target. One should also examine new cues preceded by other new cues that are new and do or do not share the same target word. The appropriate comparisons between conditions would reveal the extent to which recognition of the old cue was facilitated by a preceding old, related cue. Note that, such a paradigm would require at least twice as many stimuli as the current study because of the inherent requirements of having equal pools of old and new items in a recognition memory paradigm. As such, future studies on this issue should probably abandon the use of homonyms (and only use one type of EV). This would allow for a larger pool of useable words. Homonyms were the ideal kind of stimuli for the current study because associative distance had to be taken into account for several reasons. However, the pool of useable homonyms was quite limited and could not provide a sufficient number of stimuli for the requisite recognition memory paradigm needed for examining retrieval

processes and chunking. Therefore, if one uses only one level of associative distance, a larger pool of words other than homonyms can be used. In sum, our interpretation of the chunking index was theoretically sensible and motivated by previous research, however, testing for P_1—P_2 dependence at retrieval would add reassurance to our interpretation of the chunking index. In addition, such a test could also determine if there is cross-methodological support for the chunking hypothesis.

Item and Relational Processing, EV, & Study-Phase Retrieval. It is often found that one of the most mnemonically potent mixtures of encoding processes is a combination of item-specific and inter-item relational processing (Einstein & Hunt, 1980; Hunt & Einstein, 1981; Hunt & McDaniel, 1993; Grimaldi, Poston, & Karpicke, 2015)[5]. (Henceforth referred to as item and relational processing, respectively). Item processing refers to encoding the attributes unique to an item, whereas relational processing refers to encoding the attributes common to multiple items (i.e., the superordinate information). When subjects engage in an item-processing task such as making pleasantness ratings and then engage in a relational processing task such as sorting items by category, recall is better than performing two cycles of the same task (Hunt & Einstein, 1981). Also, subjects given a surprise free recall test following a combination of item and relational processing can outperform those who are forewarned of the free recall test, but not given any orienting tasks and simply left up to their own devices during encoding (Einstein & Hunt, 1980, Exp. 1; Hunt & Einstein, 1981, Exp. 4).

Interestingly enough, instances where item—relational processing result in better memory than item—item or relational—relational processing are technically instances where EV regimens are superior to EC ones. Other researchers have also noted that item—relational

[5] Others have also noted that memory is optimal when there is a balance between processing of similarities and differences (Nairne, 2002; Nairne, 2006; Poirier et al., 2012; Goh & Lu, 2012).

processing regimens are in fact a form of EV (Huff & Bodner, 2014). However, Huff and Bodner (2014) asserted that EV is only superior to EC when the EV regimen consists of item—relational processing. However, their theorizing—and experiments—failed to account for study-phase retrieval (and chunking or associative distance). The results of the current study, however, suggest that it is not necessary to have an item—relational encoding regimen to achieve a benefit of EV. Note that the current study only used encoding manipulations that were item-specific in nature and still found positive effects of EV. The items were completely unrelated and no inter-item relational processing was encouraged. The EV manipulations also involved varying the encoding context at the level of individual target words. Moreover, the benefits of EV were related to study-phase retrieval, chunking, and associative distance. Thus, the current study alone refutes Huff and Bodner's (2014) assertions. A regimen of item and relational processing does not appear to be necessary for a benefit of EV. It is also worth noting that, for similar reasons, the current study refutes the assumptions of Component Levels Theory (Glenberg, 1979). One of the primary assertions of this theory was that the benefits of EV differ as a function of the locus of the EV manipulation (i.e., whether it is at the level of individual items or at the level of associations between multiple items).

Future studies could examine the extent to which the benefits of item—relational processing are due to study-phase retrieval and other factors. Older experiments in the literature strongly suggest that they are. First, note that, in a typical study of item and relational processing, the item and relational tasks are performed successively for each item. Therefore, the lag between the item and relational tasks for a given item is effectively 0. If short lags are critical to the relative benefit of item—relational processing, then the difference between regimens of item—relational processing and regimens of item—item and relational—relational

processing should dissipate at longer lags. As it turns out, Hunt and Einstein (1981, Exp. 2) found precisely this result. However, they did not attribute the results to study-phase retrieval (or any other determinants of EV benefits identified herein). They assumed that spacing in and of itself incurs some encoding variability—which would disproportionately benefit the item—item and relational—relational conditions because the item—relational condition was presumably already saturated with encoding variability. Future studies could adapt the methods of the current study to an item—relational processing paradigm and determine the extent to which study-phase retrieval, chunking, and associative distance play a role. The results of the current study, combined with Hunt and Einstein's (1981) second experiment, suggest that these are key determinants of the benefits of item—relational processing regimens.

Conclusions

For many years, encoding variability was a concept used in an attempt to explain the spacing effect. Studies on this matter serendipitously revealed that spacing modulated the effects of encoding variability. To account for these findings, the chunking hypothesis was proposed. Afterwards, research on encoding variability fell largely dormant for many years. Moreover, the critical assumption of the chunking hypothesis—that study-phase retrieval is the proximate cause of EV effects—was never tested. The current study tested this assumption with methods akin to a continuous recognition memory paradigm. We also manipulated associative distance within one study. The tenets of the chunking hypothesis motivated predictions about this variable, and differences in associative distance across the literature might explain discrepant conclusions about EV between studies. In short, we found that the effects of encoding variability were related to study-phase retrieval, associative distance, and chunking in largely theoretically predicted ways. More importantly, the chunking hypothesis can account for the effects of

encoding variability in ways that other theories cannot. This will hopefully bring about clarity to a topic that has remained dormant for too long. The current study can also help the field understand when and why encoding variability can enhance memory and other educationally relevant outcomes such concept learning and transfer of knowledge. Perhaps our latest incarnation of the chunking hypothesis can also guide the field's theorizations on related frameworks (e.g., item—relational processing) and educationally relevant topics. At the very least, it can provide sorely needed rules of thumb as to when encoding variability will and will not enhance memory.

APPENDIX

Supplemental Methods Section

Encoding Phase

Materials.

For our target words, we sought ones that could be perceptually identical across presentations, but potentially be processed in semantically unrelated ways. These constraints are necessary to having a strong manipulation of associative distance. In addition, we wanted to minimize item selection effects and cross-contamination of the experimental conditions. Hence, the many other criteria imposed below to meet the above goals.

Target Words. For our target words, we used homonyms. These are words that can take on two or more different meanings but have the same pronunciation and spelling for all meanings. The restrictions inherent in these stimuli ensured that our targets had identical perceptual characteristics across meanings. To acquire such words, we primarily drew from the English homonym norms of Armstrong, Tokowicz, and Plaut (2012). A number of selection criteria were imposed to satisfy our goals above. First and foremost, we sought homonyms that had truly divergent meanings. Therefore, we could not employ polysemes. These are homonyms that have nearly identical meanings, but technically qualify as homonyms because of their different grammatical forms. For example, some words qualify as homonyms because they maintain the same spelling and pronunciation when they take on different parts of speech, even though each part is centered on a common meaning (e.g., "That was not my *intent*" and "I am *intent* on doing this"). Polysemes would undermine our manipulation of associative distance, thus they had to be excluded. We also excluded homonyms that are capitonyms. These words have different meanings and identical spellings, but are also orthgraphically different across

meanings (e.g., *ray* vs. *Ray*). Capitonyms would not be perceptually identical across meanings, so they had to be excluded. To mitigate item selection effects, we also excluded homonyms that could be imbued with disproportionate salience in a word learning experiment. For example, some words could be disproportionately salient because one of their meanings is vernacular, unfamiliar, insulting, and/or gross (e.g., words like *troll*, *beetle*[6], *husky*, or *boil*, respectively). To ensure that our manipulation was at the level of single items, we also needed target words that were ostensibly unrelated to each other. Therefore, we could only use one word in a cluster of potentially related homonyms such as *bat*, *batter, pitch,* and *pitcher* or *bluff* and *cape*. If potential targets were found to be related, one was randomly selected from the cluster. After imposing the above criteria, we obtained 56 homonyms from the Armstrong et al. (2012) database. However, to get a sufficient number of words for all of our experimental conditions, 8 additional homonyms (not found in Armstrong et al., 2012) were drawn from the materials of Slamecka and Barlow (1979) that also met our selection criteria.

Cue Words. In our initial design, we used paired associates. Each target word had to have four cue words—two associated with one meaning and two associated with another. We retained the paired associates we originally created and built sentences around them as our materials in the current study. The selection of the cue words is described in this section. Because we required a large number of cue words that met multiple constraints, it was impossible to obtain the requisite number from a normative database. However, we wanted to impose some normative basis for selecting our materials to minimize the researcher's biases in the creation of them. Therefore, most of the cue words were obtained from a normative database

[6] As an example of an unfamiliar alternative meaning, one of this word's alternative meanings was "a wooden mallet or club used to stamp and finish linen, compress paving stones, or mash vegetables."

via a process described below. Once that database was exhausted, the author, MP, selected the rest of the requisite cue words.

The majority of the cue words were drawn from the Nelson, McEvoy, and Schreiber (2004) free association norms. To mitigate the effects of semantic guessing on our cued recall test, we avoided using the strongest associates in the norms or any word with an associative strength to the homonym greater than .15. Below this threshold, we selected the strongest associates for a given meaning. We first sought words from the side of the norms where the homonym appeared as a freely associated target in response to a cue word. When we could not find the requisite cue words from this side of the norms, we then expanded our search to the other side of the norms, where the homonym appeared as a cue to which people made responses[7]. When there were not enough words from either side of the norms to produce enough cues for a target, the remainders were generated by the author, MP. Finally, the entire bank of targets and cues was constructed such that there were no duplicates (i.e., no cue for a target appeared as a target elsewhere, nor as a cue associated with a different target).

The Nelson et al. (2004) norms also heavily dictated which meanings of a homonym we employed. Specifically, we employed the two meanings for which there were the most associates to the homonym. If only one meaning was represented in the Nelson et al. (2004) norms, the author, MP, selected the second meaning to be used for the homonym. Meaning dominance was also determined by the Nelson et al. (2004) norms. That is, whichever meaning had the strongest associate to the homonym in the norms was determined to be the strongest meaning.

[7] When looking solely at the norms for when a homonym was produced as a target in response to cue, there were not always enough words to obtain two cues for two meanings of a homonym. Thus, we had compromise by looking at both sides of the Nelson et al. (2004) norms to maintain some normative basis to the creation of our materials.

Sentences. The sentences were constructed to ensure that the targets only appeared in the 4 requisite sentences written for each target word. This was necessary given that the repetition judgments were always over the target words. However, the enormity of the sentence bank made it impossible to ensure that the cue words also only appeared in the sentences intentionally built around a cue-target pairing. Keep in mind that the cue words only served as a normative constraint around which we built our sentences. Thus, the frequency with which they appeared in our sentence bank need not be as strictly controlled. The sentence frames used in the recall test for each target were also piloted to minimize the rate at which participants could guess the correct target word. After multiple waves of free association piloting, we were able to reduce the correct guessing rate for our bank of sentences to an average of 12%.

Counterbalancing Scheme. To facilitate counterbalancing, we divided up the targets into four sub-lists. One group of 10 targets was allocated to a sub-list used for primacy items. These primacy items were constant for all subjects. The remaining 54 targets were used as critical items for the experiment. The 54 critical items were divided up into 3 sub-lists comprised of 18 targets each. We divided the targets up such that each sub-list was matched on average log SUBTLEX contextual diversity (Brysbaert & New, 2009). Contextual diversity refers to the number of semantic contexts in a corpus a word appears in (operationally defined as the number of passages or documents in which a word appears in a corpus). Note that this metric does not refer to the sheer number of times a word appears in a corpus (for further explanation, see Brysbaert & New, 2009). Regardless, it is a lexical attribute that is at least related to word frequency metrics like that from the Kucera-Francis norms (Kucera & Francis, 1967). However, contextual diversity is a better predictor of lexical-decision times (Brysbaert & New, 2009). Norms on the related concept of contextual variability also turn out to be better predictors of item

recognition memory (Steyvers & Malmberg, 2003) and source memory (Marsh, Cook, & Hicks, 2006). Given the better predictive validity of contextual diversity as opposed to mere word frequency, we decided to control for the former instead of the latter in the current study. [Log SUBTLEX contextual diversity (Sub-List 1: $M = 2.54$, $SD = .51$; Sub-List 2: $M = 2.53$, $SD = .40$; Sub-List 3: $M = 2.56$, $SD = .57$; Primacy Items: $M = 2.61$, $SD = .66$)].

The targets in the 3 critical sub-lists were counterbalanced across all the encoding conditions (i.e. 1P, EC, EV Similar, and EV Different) and all 3 lags. To control for the effect of the meaning dominance of the target words, half of the targets shown to each subject appeared in their more dominant meaning for all experimental conditions while the remaining half appeared in their less dominant meaning. One exception was for the Different EV condition. Given that, in this condition, the meanings of the EV targets completely changed from P_1 to P_2, all subjects in the Different group saw half of the EV targets appear in their more dominant meaning first (i.e., on P_1) and vice versa.

Study List Construction. The beginning of the encoding phase consisted of a primacy buffer that contained the same targets, embedded in the same sentences, and shown in the same order for all subjects. Half of the primacy items were targets appearing in their more dominant meaning. This buffer consisted of 14 trials comprised of 6 1P items and 4 repeated items (2 each in the lag 0 and lag 6 conditions). The repeated items in the primacy buffer were only EC items. We maintained this constraint because the type of EV differed between groups.

For the critical targets in the rest of the encoding phase, we constructed a skeleton list with a series of placeholder trials for the 7 experimental conditions (1P, EC x Lag, & EV x Lag). This skeleton list was constructed to ensure that the items for any given experimental condition were not disproportionately concentrated in one part of the encoding phase. The placeholders

established in this skeleton list remained constant for all subjects but its constituent items necessarily changed according to the subject's position in the counterbalancing scheme. The skeleton list consisted of 3 waves of 30 placeholder trials. Each wave contained placeholder trials for 18 target items: 6 1P items and 6 items from the two repeated conditions (with equal numbers in each assigned to the 3 lags). These waves of trials were constructed such that there were no consecutive placeholders for items from the same experimental condition—with the exception of massed repetitions, of course. In addition, half of the placeholders for each experimental condition of each wave were reserved for targets appearing in their more dominant meaning; the other half were reserved for target words appearing in their less dominant meaning. This design feature ensured that meaning dominance was not confounded with any of the experimental conditions. These placeholders were then populated with randomly selected sentences from the respective sub-list used for an encoding condition's location in the counterbalancing scheme.

In addition, we imposed other constraints to improve the comparability of the Similar and Different groups. First, within each segment of the counterbalancing scheme, identical sentences were used in the 1P and EC conditions for both groups. For the respective EV conditions of both groups, only their P_1 sentences were different. For example, if someone in the Similar group had the target word *pupil* in their EV condition, they might see it embedded in the sentence "The student in the front row became the teacher's favorite pupil" at P_1, and in the sentence "Being a fast learner made her a star pupil in the class" at P_2. We would then create the corresponding version for someone in the Different group to use the sentence "During the eye exam, the doctor noticed that one pupil was abnormal" at P_1, and the sentence "Being a fast learner made her a star pupil in the class" at P_2. Note that the two EV types we employed necessarily required using

at least some different sentences for each group. However, our yoking procedure ensured that half of the sentences for each group's EV condition were identical. Thereby ensuring that any differences between the two groups were not confounded with the use of completely different sentence frames. In addition, because the EV P_2 sentences were identical, that ensured that the repetition judgments in the EV conditions of both groups were at least made on the same study-phase retrieval cues.

REFERENCES

Armstrong, B. C., Tokowocz, N., & Plaut, D. C. (2012). eDom: Norming software and relative meaning frequencies for 544 English homonyms. *Behavior Research Methods*, *44 (4)*, 1015 – 1027.

Bellezza, F. S., & Young, D. R. (1989). Chunking of repeated events in memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *15*(5), 990 – 997.

Bellezza, F. S., Winkler, H. B., & Andrasik, F. (1975). Encoding processes and the spacing effect. *Memory & Cognition*, *3*(4), 451 – 457.

Benjamin, A. S., & Tullis, J. (2010). What makes distributed practice effective? *Cognition*, *61*, 228 – 247.

Berry, C. J., Shanks, D. R., Speekenbrink, M., & Henson, R. N. A. (2012). Models of recognition, repetition priming, and fluency: Exploring a new framework. *Psychological Review*, *119(1)*, 40 – 79.

Bird, C. P., & Nicholson, A. J., & Ringer, S. (1978). Resistance of the spacing effect to variations in encoding. *American Journal of Psychology, 91*, 713-721.

Bobrow, S. A. (1970). Memory for words in sentences. *Journal of Verbal Learning and Verbal Behavior, 9*, 363-372.

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

Bower, G. H., Lesgold, A. M., Tieman, D. (1969). Grouping operations in free recall. *Journal of Verbal Learning and Verbal Behavior*, *8*, 481 – 493.

Bransford, J. D. (1979). Acquiring new knowledge and skills. *Human cognition: Learning, understanding, remembering* (pp. 205-245). Belmont, CA: Wadsworth.

Braun, K., & Rubin, D. C. (1998). The spacing effect depends on an encoding deficit, retrieval, and time in working memory: Evidence from once-presented words. *Memory*, *6*(1), 37 – 65.

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41 (4)*, 977 – 990.

Butler, A. C., & Marsh, E. J. (2012). *Retrieval variability promotes superior transfer of learning*. Poster presented at the annual scientific meeting of the Psychonomic Society, Minneapolis, MN.

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

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

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and 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.

D'Agostino, P. R., & DeRemer, P. (1973). Repetition effects as a function of rehearsal and encoding variability. *Journal of Verbal Learning and Verbal Behavior, 12*(1), 108-113. doi: 10.1016/S0022-5371(73)80066-0

Dellarosa, D., & Bourne, L. E. (1985). Surface form and the spacing effect. *Memory & Cognition*, *13(6)*, 529-537.

Dempster, F. N. (1996). Distributing and managing the conditions of encoding and practice. In E. L. Bjork & R. A. Bjork (Eds.), *Human memory* (pp.197 – 236). San Diego, CA: Academic Press.

DeRemer, P., & D'Agostino, P. R. (1974). Locus of distributed lag effect in free recall. *Journal of Verbal Learning and Verbal Behavior, 13*(2), 167-171.

di Vesta, F. J., & Peverly, S. T. (1984). The effects of encoding variability, processing activity, and rule-examples sequence on the transfer of conceptual rules. *Journal of Educational Psychology, 76*(1), 108-119. doi: 10.1037/0022-0663.76.1.108

Duncan, K., Sadanand, A., & Davachi, L. (2012). Memory's penumbra: Episodic memory decisions induce lingering mnemonic biases. *Science*, *337*, 485 – 487.

Einstein, G. O., & Hunt, R. R. (1980). Levels of processing and organization: Additive effects of individual-item and relational processing. *Journal of Experimental Psychology: Human Learning and Memory, 6*(5), 588—598.

Glenberg, A. M. (1976). Monotonic and nonmonotonic lag effects in paired-associate and recognition memory paradigms. *Journal of Verbal Learning and Verbal Behavior*, *15*, 1 – 16.

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

Glenberg, A. M., & Smith, S. M. (1981). Spacing repetitions and solving problems are not the same. *Journal of Verbal Learning and Verbal Behavior, 20*(1), 110-119. doi: 10.1016/S0022-5371(81)90345-5

Goh, W. D., & Lu, S. H. X. (2012). Testing the myth of the encoding-retrieval match. *Memory & Cognition, 40,* 28—39.

Greene, R. L., & Stillwell, A. M. (1995). Effects of encoding variability and spacing on frequency discrimination. *Journal of Memory and Language, 34*(4), 468-476. doi: 10.1006/jmla.1995.1021

Grimaldi, P. J., Poston, L., & Karpicke, J. D. (2015). How does creating a concept map affect item-specific encoding? *Journal of Experimental Psychology: Learning, Memory, & Cognition, 41*(4)*,* 1049—1061.

Hintzman, D. L., Summers, J. J., & Block, R. A. (1975). Spacing judgments as an index of study-phase retrieval. *Journal of Experimental Psychology: Human Learning and Memory, 1*, 31-40.

Hintzman, D. L. (1976). Repetition and Memory. *The psychology of learning and motivation. Vol. 10*. In G. H. Bower (Ed.), *Repetition and memory* (pp.47-91). New York: Academic Press.

Hintzman, D. L. (2004). Judgment of frequency versus recognition confidence: Repetition and recursive reminding. *Memory & Cognition*, *32*(2), 336 – 350.

Huff, M. J., & Bodner, G. E. (2014). All varieties of encoding variability are not created equal: Separating variable processing from variable tasks. *Journal of Memory and Language, 73,* 43—58.

Hunt, R. R., & Einstein, G. O. (1981). Relational and item-specific information in memory. *Journal of Verbal Learning & Verbal Behavior, 20,* 497—514.

Hunt, R. R., McDaniel, M. A. (1993). The enigma of organization and distinctiveness. *Journal of Memory & Language, 32,* 421—445.

Hyde, T. S., & Jenkins, J. J. (1973). Recall for words as a function of semantic, graphic, and syntactic orienting tasks. *Journal of Verbal Learning and Verbal Behavior*, *12*, 471—480.

Jacoby, L. L. (1972). Context effects on frequency judgments of words and sentences. *Journal of Experimental Psychology, 94*(3), 255-260.

Johnston, W. A., Coots, J. H., & Flickinger, R. G. (1972). Controlled semantic encoding and the effect of repetition lag on free recall. *Journal of Verbal Learning and Verbal Behavior, 11*(6), 784-788. doi: 10.1016/S0022-5371(72)80013-6

Johnston, J. A., & Uhl, C. N. (1976). The contributions of encoding effort and variability to the spacing effect on free recall. *Journal of Experimental Psychology: Human Learning and Memory*, *2(2)*, 153 – 160.

Jones, G. V. (1976). A fragmentation hypothesis of memory: Cued recall of pictures and of sequential position. *Journal of Experimental Psychology: General*, *105*, 277 – 293.

Klein, K., & Saltz, E. (1976). Specifying the mechanisms in a levels-of-processing approach to memory. *Journal of Experimental Psychology: Human Learning and Memory*, *2(6)*, 671 – 679.

Kucera, H., & Francis, W. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Madigan, S. A. (1969). Intraserial repetition and coding processes in free recall. *Journal of Verbal Learning and Verbal Behavior, 8*, 828-835.

Maki, R. H., & Hasher, L. (1975). Encoding variability: A role in immediate and long-term memory? *The American Journal of Psychology, 88(2)*, 217—231.

Marsh, R. L., Cook, G. I., & Hicks, J. L. (2006). The effect of context variability on source memory. *Memory & Cognition*, *34(8)*, 1578—1586.

Maskarinec, A. S., & Thompson, C. P. (1976). The within-list distributed practice effect: Tests of the varied context and varied encoding hypotheses. *Memory & Cognition, 4*, 741-746.

McFarland, C. E., Rhodes, D. D., & Frey, T. J. (1979). Semantic-feature variability and the spacing effect. *Journal of Verbal Learning and Verbal Behavior, 18*(2), 163-172. doi: 10.1016/S0022-5371(79)90100-2

Melton, A. W. (1967). Repetition and retrieval from memory, *Science*, *158*, 532.

Nairne, J. S. (2002). The myth of the encoding-retrieval match. *Memory, 10,* 389—395.

Nairne, J. S. (2006). Modeling distinctiveness: Implications for general memory theory. In R. R. Hunt & J. B. Worthen (Eds.), *Distinctiveness and memory* (pp. 27—46). New York, NY: Oxford University Press.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of Florida free association, rhyme, and word fragment norms. *Behavior Research Methods*, *Instruments, & Computers*, *36 (3)*, 402 – 407.

Nitsch, K. E. (1977). Structuring decontextualized forms of knowledge. (Doctoral Dissertation, Vanderbilt University, 1977). *Dissertation Abstracts International, 38,* 3479B-3968B.

Paivio, A. (1974). Spacing of repetitions in the incidental and the intentional free recall of pictures and words. *Journal of Verbal Learning and Verbal Behavior*, *13*, 497 – 511.

Poirier, M., Nairne, J. S., Morin, C., Zimmerman, F. G. S., Koutmeridou, K., & Fowler, J. (2012). Memory as discrimination: A challenge to the encoding-retrieval match principle. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 38*(1)*,* 16—29.

Postman, L., & Knecht, K. (1983). Encoding variability and retention. *Journal of Verbal Learning and Verbal Behavior, 22(2),* 133-152. doi: 10.1016/S0022-5371(83)90101-9

Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1(3)*, 181 – 210.

Roediger, H. L., Sanches, J. B., & Agarwal, P. K. (2011). *Does variable encoding affect learning and retention relative to constant encoding?* Talk presented at the annual scientific meeting of the Psychonomic Society, Seattle, WA.

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 and Verbal Behavior*, *22*, 133 - 152.

Rowe, E. J. (1973). Frequency judgments and recognition of homonyms. *Journal of Verbal Learning and Verbal Behavior, 12*(4), 440-447. doi: 10.1016/S0022-5371(73)80024-6

Russo, R., Mammarella, N., & Avons, S. E. (2002). Toward a unified account of spacing effects in explicit cued-memory tasks. *Journal of Experimental Psychology Learning, Memory, and Cognition, 28*(5), 819-829. doi: 10.1037/0278-7393.28.5.819

Shaughnessy, J. J. (1976). Persistence of the spacing effect in free recall under varying incidental learning conditions. *Memory & Cognition, 4*, 369-377.

Shaughnessy, J. J., Zimmerman, J., & Underwood, B. J. (1974). The spacing effect in the learning of word pairs and the components of word pairs. *Memory & Cognition, 2*, 742-748.

Slamecka, N. J., & Barlow, W. (1979). The role of semantic and surface features in word repetition effects. *Journal of Verbal Learning and Verbal Behavior*, *18*, 617 – 627.

Starns, J. J., & Hicks, J. L. (2005). Source dimensions are retrieved independently in multidimensional monitoring tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 1213 – 1220.

Starns, J. J., & Hicks, J. L. (2008). Context attributes in memory are bound to item information, but not to one another. *Psychonomic Bulletin & Review*, *15(2)*, 309 – 314.

Steyvers, M., & Malmberg, K. N. (2003). The effect of normative context variability on recognition memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *29(5)*, 760—766.

Thios, S. J. (1972). Memory for words in repeated sentences. *Journal of Verbal Learning and Verbal Behavior, 11*, 789-793.

Thios, S. J., & D'Agostino, P. R. (1976). Effects of repetitions as a function of study-phase retrieval. *Journal of Verbal Learning and Verbal Behavior*, *15*, 529-537.

Verkoeijen, P. P. J. L., Rikers, R. M. J. P., & Schmidt, H. G. (2004). Detrimental influence of contextual change on spacing effects in free recall. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *30(4)*, 796 – 800.

Winograd, E., & Raines, S. R. (1972). Semantic and temporal variation in recognition memory. *Journal of Verbal Learning and Verbal Behavior, 11*(1), 114-119. doi: 10.1016/S0022-5371(72)80067-7

Young, D. R., & Bellezza, F. S. (1982). Encoding variability memory organization, and the repetition effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8*, 545-559.