Washington University in St. Louis
Washington University Open Scholarship

All Theses and Dissertations (ETDs)

1-1-2011

Often Wrong but Never in Doubt: Categorized Lists Produce Confident False Memories

Kurt DeSoto

Follow this and additional works at: https://openscholarship.wustl.edu/etd

Recommended Citation
DeSoto, Kurt, "Often Wrong but Never in Doubt: Categorized Lists Produce Confident False Memories" (2011). *All Theses and Dissertations (ETDs)*. 540.
https://openscholarship.wustl.edu/etd/540

This Thesis is brought to you for free and open access by Washington University Open Scholarship. It has been accepted for inclusion in All Theses and Dissertations (ETDs) by an authorized administrator of Washington University Open Scholarship. For more information, please contact digital@wumail.wustl.edu.

WASHINGTON UNIVERSITY

Department of Psychology

Often Wrong but Never in Doubt:

Categorized Lists Produce Confident False Memories

by

Kurt Andrew DeSoto

A thesis presented to the
Graduate School of Arts and Sciences
of Washington University in
partial fulfillment of the
requirements for the
degree of Master of Arts

December 2011

Saint Louis, Missouri

Abstract

In the *categorized list procedure*, subjects study words from semantic categories, then take a recognition test on those items. Subjects are likely to recognize common (high *output dominance*) category members even when they are not studied. Across three experiments, we sought to extend the categorized list procedure, further develop an explanation of why false recognition of category members occurs in this procedure, and modulate false recognition of category members by manipulating encoding and retrieval phases. Experiment 1 extended previous categorized list research, showing that subjects are likely to false alarm with high confidence to high output dominance category members that were not studied. Experiment 2 attempted to improve subjects' metacognitive awareness of the deceptive nature of categorized lists by eliminating unrelated lures from the test list. This manipulation decreased the overall false alarm rate, but did not eliminate the relationship between output dominance and false alarm rate. Experiment 3 sought to reduce false recognition by providing subjects additional study and test events on the material. Providing subjects feedback on their tests followed by an additional study period to relearn the material was generally successful in eliminating high confidence false alarms at final test for frequent category members that were not studied. The results converge to imply that (1) the categorized list procedure is a robust method for investigating false recognition, (2) false recognition for high output dominance items may be related to source monitoring errors during testing caused by processes at encoding (e.g., Dewhurst, 2001), and (3) these high confidence false memories are relatively resistant to manipulations intended to minimize them.

Acknowledgments

I gratefully thank my major professor, Dr. Henry Roediger, III, as well as the other members of my Thesis Examination Committee, Dr. Larry Jacoby and Dr. Ian Dobbins, for their insight, direction, and advice throughout this project. I also wish to recognize the Washington University Psychology Department Behavior, Brain, & Cognition program, the Memory Lab, the Memory & Cognition Lab, and particularly Adam Putnam, Yana Weinstein, John Nestojko, Pooja Agarwal, and Megan Smith for their collegial enthusiasm and comments. Dr. Leonel Garcia-Marques is appreciated for contributing the inspiration for Experiment 2, and Dr. John Wixted is to thank for suggesting improved ways to assess and display the data. I would like to thank Jane McConnell, Jean Ortmann-Sotomayor, and Brittany Butler for administrative assistance on this project, as well as Paige Madara, Zoe Tabachnick, Robyn Husa, and Kelly Young for their help collecting data. Last, I am indebted to my family -- Kurt, Donna, and Aimee DeSoto -- for their support, and my gratitude especially extends to Rebecca Koenig for her unyielding optimism, encouragement, and companionship.

Table of Contents

List of Tables

List of Figures

Often Wrong but Never in Doubt:

Categorized Lists Produce Confident False Memories

Although the general issue of false memory has been of considerable applied interest for forensic investigators and eyewitness testimony experts, the topic has also enjoyed a spotlight in the cognitive laboratory. Research carried out in this area generally occurs as follows: subjects come into the laboratory and study a set of materials such as pictures, words, sentences, or stories and then take a final recall or recognition test. Often, the characteristics or composition of the studied materials cause false memories to occur; items that were not studied (pictures, sentences, or words) are recognized or recalled, sometimes at high rates. Examining the memory errors that arise in these experiments not only helps explain why these errors occur, but also may provide information about the cognitive processes underlying normal (i.e., non-illusory) memory performance (Roediger, 1996).

The study of false memory is particularly compelling because subjects not only remember events and words that were not studied, but also sometimes remember them with extreme confidence. Metacognition — that is, individuals' monitoring and control over their own thinking — often breaks down in certain experiments, resulting in strikingly self-assured wrong answers. These high confidence false memories have some significant implications for applied and theoretical memory science.

A variety of methods to study high confidence false memories in the research laboratory have been developed to address both theoretical and applied issues. Among them are the misinformation procedure (Loftus, 1975; Loftus, Miller, & Burns, 1978); the

false fame procedure (Jacoby, Woloshyn, & Kelley, 1989); the plurality recognition test (Rotello, Macmillan, & Van Tassel, 2000; Mickes, Hwe, Wais, & Wixted, 2011; Wixted, 2011); the use of deceptive sentences (Brewer & Sampaio, 2006; Brewer, Sampaio, & Barlow, 2005; Sampaio & Brewer, 2009), and the Deese-Roediger-McDermott procedure (DRM; Deese, 1959; Roediger & McDermott, 1995). Although how these studies are conducted appears similar, the theoretical mechanisms by which memory errors arise in these procedures seem to differ (see Marsh, Eslick, & Fazio, 2008, for a review). Researchers have relied on these methods for a number of purposes, ranging from making claims about the forensic relevance of evidence to developing theoretical models of different memory processes.

One evolving method for investigating false memories uses categorized lists — lists of words organized by semantic category (e.g., Dewhurst & Anderson, 1999; Smith, Ward, Tindell, Sifonis, & Wilkenfeld, 2000). A "birds" list, for instance, may contain a variety of exemplars ranging from common members of the category, like "robin," to less common members, like "ostrich." Studies using these materials have found that participants often falsely recall or falsely recognize items that were not presented, especially items that are common to or typical of the category (Smith et al., 2000). Although categorized list methods have revealed interesting patterns of data, as discussed below, the use of these materials and the theories underlying them remain somewhat underdeveloped, and many questions remain unanswered.

In this thesis, we suggest a revised procedure to examine high confidence false memories using categorized lists. Besides developing a refined categorized list method

to study high confidence false memories, this thesis seeks to evaluate (1) the theoretical mechanisms that underlie the false memories evoked by these materials and (2) whether effects demonstrated by these methods can be modulated experimentally, as predicted by theory. First, we summarize the various ways categorized lists have been used in the study of false memory. Next, we discuss the theoretical explanations of why false recall and false recognition occur for categorized list materials. Last, we describe a series of three experiments that begin to extend and enhance the methods and understanding of high confidence false memories.

The Categorized List Procedure

The plethora of studies carried out using the DRM procedure (Deese, 1959; Roediger & McDermott, 1995) as a tool attest to its wide applicability and theoretical utility (see Gallo, 2006, 2010, for a review). In this procedure, subjects study a list of words (e.g., "bed," "rest," "awake") that are semantically associated to a critical lure word that is never presented (i.e., "sleep"). At test, subjects are highly likely to incorrectly recognize or recall the critical lure word as being listed, often with high confidence. The structure of the DRM procedure can be limiting, however, in that only one critical lure word generally exists per list and that all the studied words are considered to be similar for purposes of the procedure.

Categorized list methods (e.g., Dewhurst & Anderson, 1999; Smith et al., 2000) address some of the limitations of the DRM procedure. These methods tap into natural categories, which have been shown by the work of Rosch (1973) and others to be integral to learning, classification, language, and culture (Rosch & Mervis, 1975; see Dry &

Storms, 2010, for a review). In these experiments, subjects typically study lists of words comprising different semantic categories such as birds, four-legged animals, or vegetables (see the Appendix for an example). These categories often contain items that are common members of the category (e.g., "robin," "dog," or "carrot") as well as items that are rarer for the category (e.g., "ostrich," "moose," or "turnip"). The composition of these lists is manipulated during the study phase; for instance, the most common category member might be removed. At test, however, subjects might be highly likely to recall or recognize that typical category member even though it was not presented (e.g., Smith et al., 2000). We will discuss these findings in more detail later in this section.

Categorized list materials are usually drawn from a norming study executed by Battig and Montague (1969) or an updated version by Van Overschelde, Rawson, and Dunlosky (2004). In the Van Overschelde et al. study, three groups of undergraduates (N = 600) from geographically diverse universities were recruited. These subjects heard a category name and were given 30 seconds to type as many members of the category as possible. The responses were ranked for each category from the most to least frequently produced member of that category. These categorized lists are deemed to be ordered by *output dominance,* such that the most frequently mentioned category members have higher output dominance than less frequently mentioned category members.[1] Responses that occurred with a frequency of 5% or less were excluded from this rank ordering. Moreover, as noted in the original Battig and Montague norms, "All legible responses

[1] The term used to describe an item's position in the norms varies from paper to paper, possibly because this variable is not named in the Van Overschelde et al. (2004) or the original Battig and Montague (1969) norms. We use *output dominance* here consistent with the work of Smith and colleagues (2000), but Dewhurst and colleagues (1999) use the phrase *instance frequency* to contrast it with traditional *printed frequency*. When the categorized lists come from prototype studies as in Rosch (1975), the term *typicality* may be appropriate for describing this variable.

[were] included, even those which are obviously inappropriate to the category name" (p. 3). Thus, the word "spider" could be found on the "insects" list, even though, strictly speaking, it is not a member of that category.

Categorized lists are a useful tool for examining high confidence false memories because they offer at least two distinct advantages over other methods: they provide a greater number of experimental observations per list and permit also allow a high degree of manipulation of the stimuli. Although false memories evoked by DRM lists (sometimes called *associative lists* to contrast them with categorized lists) tend to be quite strong, each associative list contains only one critical (i.e., unpresented) item. Thus, many associative lists must be used to collect enough false memory observations. Moreover, even with a sufficient number of lists (at least 55 DRM lists can be found in the literature), using this method could be inefficient in certain cases.

A similar problem arises with the more forensically relevant methods, which generally collect only one observation of interest per subject (e.g., was the perpetrator correctly identified in the lineup?). The methodological issue here is that these procedures lead researchers to compute point-biserial coefficients to assess confidence-accuracy relationships, which can give an incomplete or incorrect picture of the relationship between these two variables (see Roediger, Wixted, & DeSoto, 2011, for a review of these issues).

Categorized lists consist of the category name (a superordinate) followed by subordinates, category members rank-ordered by some measure of category membership. This structure allows experimental manipulation that is not possible in the methods

described above. If a categorized list word is removed from the stimulus set to become a critical item, for instance, the "strength" of the critical item is directly related to its position in the norms. This also means that more than one item can be removed from a categorized list at study. The same materials can also be used to compare and contrast false memories for items of different positions in the norms.

The use of categorized lists to study false memories was pioneered in two lines of study: one by Smith and colleagues (Smith et al., 2000; Smith, Gerkens, Pierce, & Choi, 2002; Smith, Tindell, Pierce, Gilliland, & Gerkens, 2001) and a second by Dewhurst and colleagues (Dewhurst, 2001; Dewhurst & Anderson, 1999; Dewhurst et al., 2009; Dewhurst & Farrand, 2004). Smith et al. (2000) selected nine categorized lists from previous research (in this case, Rosch, 1975). Consistent with the DRM procedure, the item with the highest output dominance — that is, the item that was most frequently mentioned by subjects in the norming studies — was omitted from each category at presentation. Subjects studied each list and were given a recall test either just after each list was studied or after all lists had been presented. For each recall test, a category name was displayed and subjects recalled as many items as they could from the corresponding list. Results demonstrated that this category cued recall procedure evoked intrusions of the omitted (i.e., highest output dominance) item in recall, especially when the recall test was given after all lists had been presented relative to when the recall test was given after the presentation of each list. This pattern has also been demonstrated in research with DRM materials (McDermott, 1996).

Smith et al. (2000) also noted that the use of categorized lists made it "possible to systematically observe the effects of gradations in the strength of items from the category" (p. 389). To demonstrate this observation, the researchers collected their own norming data for 10 different categories of 30 items each. They then assigned alternating items to two counterbalancing lists (A and B). Half of subjects studied the ten A lists and the other half studied the ten B lists. Subjects were cued by category and asked to recall as many items as possible that were presented from the given category. After subjects had recalled the items from each of the 10 lists, they were then asked to assign a confidence rating from 1 (*complete guess*) to 10 (*absolutely certain the response was on the studied list*) for each recalled word.

Results from this experiment showed that items higher in output dominance (i.e., the more frequently mentioned items in the norms) were both falsely and correctly recalled more regularly than items lower in output dominance. No relation between confidence ratings and intrusions (i.e., false recall) was found. These findings were further confirmed by several hierarchical multiple regression analyses, leading Smith et al. (2000) to conclude that false recall is related to output dominance. According to these researchers, category items that easily come to mind when given a category cue — i.e., category items for which there is high conceptual or retrieval fluency (akin to perceptual fluency; see Jacoby & Dallas, 1981) — are likely to be represented in both false and correct recall.

The lack of relationship between confidence and false recall in this experiment, however, is surprising. Smith et al. (2000) theorized, "Neither item accessibility nor

distinctiveness contribute to one's confidence that falsely recalled items are accurate" (p. 391). As we shall soon see, the same conclusion cannot be made for the relationship between confidence and accuracy in recognition memory for categorized list words.

Meade and Roediger (2006, 2009) also used categorized lists in their research. Instead of omitting only the item with the highest output dominance from each list, the first five items were removed. Items were presented to subjects list by list and the recall test was cued by category. The researchers found frequent intrusions of the omitted words from the studied categories in recall; this outcome occurred with greater frequency for older than for young adults. Older adults' poorer performance on this task was attributed to their greater difficulties in source monitoring (i.e., distinguishing the original source of a memory). These studies also illustrate the flexibility of categorized lists; omitting five words per list to serve as critical items would be much more difficult to do with the DRM procedure.

Another line of research on the categorized list procedure is found in the work of Dewhurst and colleagues (Dewhurst, 2001; Dewhurst & Anderson, 1999; Dewhurst et al., 2009). Dewhurst and Anderson (1999) also compared errors in the categorized list procedure — which they call the *category repetition procedure* — to those in the DRM procedure. They drew on Rajaram's (1996) distinctiveness-fluency framework, based on original contributions made by Jacoby (1991), to make predictions about errors in memory for categorized lists. According to Jacoby, two independent processes operate in (and support) recognition memory: recollection and familiarity. Recollection is typified by conscious, controlled processes that rely on attention, whereas familiarity judgments

are made with little processing effort and are often based on characteristics of perceptual processing (such as speed). The distinctiveness-fluency framework integrates these processes, dictating that when the focus of encoding is on the distinctiveness of individual items, subjects rely more heavily on recollection when responding (i.e., answer with "remember" responses; see Tulving, 1985); however, when the focus of encoding is to make the to-be-studied material more fluent (whether perceptually or conceptually), familiarity-based (i.e., "know") responses increase. Dewhurst and Anderson proposed that studying multiple items of semantic categories enhanced the conceptual fluency of that category, leading to familiarity-based false memory errors: "When a member of a category was presented at test, activation that resulted from the preceding encounter with the category enhanced the familiarity" (p. 671). Dewhurst and Anderson also found a high incidence of false remember responses, however, in line with Roediger and McDermott (1995). Drawing on similar theories to explain such errors in the DRM procedure, the researchers suggested that false remember responses were due to errors of source monitoring (Johnson, 2006; Lindsay, 2008). They argued that subjects implicitly generated related category members during encoding and then at test erroneously attributed their recollections to studied words instead of words generated during encoding.

Dewhurst (2001) further evaluated the implicit generation theory in a second study. He hypothesized that subjects should be more likely to generate high output dominance category members relative to lower output dominance category members, since those members are, after all, more frequently mentioned members of the category in

the controlled norming studies (e.g., Van Overschelde et al., 2004). To assess this hypothesis, Dewhurst manipulated output dominance of target and lure items so that half of the targets and lures were high output dominance items and the other half were low output dominance targets and lures. Subjects studied six items each from 25 categorized lists — one high dominance, one low dominance, and four of medium dominance — and were tested on four items from each category: two high dominance items (one target, one lure) and two of low dominance (one target, one lure), as well as 50 nonstudied items that were unrelated to the studied categories. Results showed that subjects made more "remember" and "know" false alarms to high output dominance lures than low output dominance lures. (In correct recognition, there were more remember responses to low frequency items and more know responses to high frequency items.) He concluded that these findings supported the hypothesis that some false alarms occurring in the category list procedure resulted from "high degrees of match between lures and nonstudied items generated at encoding" (p. 157). Other false alarms, as described above, were results of source memory errors.

One important issue at this juncture is the exact nature of the implicit generation processes that occur during encoding. Are category members generated intentionally (overtly) or automatically? If the latter, do items achieve conscious representation or is activation implicit and unconscious? Are subjects aware that they engage in these processes? A 2004 study by Dewhurst and Farrand informs some of these topics. In this study, subjects studied multiple categorized list items and were asked to provide verbal descriptions for their remember, know, and guess responses for false and correct

recognition responses. The researchers found that 36% of false remember responses were given verbal descriptions consistent with associations made or items generated at encoding. For example, when one subject falsely recognized (and provided a remember response to) the word "Saturn" when presented earlier with a list of planets, he or she remarked, "There were other planets on the list — helped me remember it." Similarly, another subject falsely recognized the word "cousin," stating that he or she remembered "because I saw brother and sister as well" (p. 408).

This introspective evidence is supported by a number of other studies. In his book on the DRM procedure, Gallo (2010) cites studies by Norman and Schacter (1997), Huron, Servais, and Danion (2001), and Read (1996) that support the idea of *conscious activation*; namely, subjects "might consciously think of the related lure at study and potentially rehearse this word along with the list words" (p. 81). In sum, evidence suggests that subjects are aware of this process occurring, but the degree of intentionality of these processes may vary.

More recently, in an unpublished study, Roediger and DeSoto (2011) extended the work of Meade and Roediger (2006, 2009) by including confidence judgments in the categorized list procedure. Subjects studied 15 words from 10 categorized lists (items ranked 6-20 from sets of items 1-25 in the norms). The five most typical (1-5) and five least typical (21-25) items from each list were omitted from the study phase; 150 words in total were studied. Subjects were then given a recognition test over all 250 words from the 10 studied lists (items 1-25, 15 studied), as well as 50 unrelated words from new categories. Roediger and DeSoto conducted an item-level analysis and found a weak

positive association between confidence and accuracy (r = .29) across all items. This association was revealed to be the result of two subcomponents that differed markedly, however. A strong positive relationship existed between confidence and accuracy for the 150 studied words (r = .70), but a negative relationship existed for the 50 highly typical lure words (r = -.54). In the latter case, high confidence was associated with the greatest false alarm rates (i.e., both false alarms and confidence were higher for items 1-5 in the norms than for items 21-25). Conversely, subjects rarely false alarmed to items of low output dominance (21-25), and when these false alarms did occur, subjects were significantly less confident in them. In summary, subjects tended to be extremely confident when false alarming to high output dominance items and less confident on low output dominance items.

When Roediger and DeSoto examined the false alarm rates for items 1-5 across the different categorized lists, they observed a linear pattern: subjects appeared more likely to false alarm to item 1 than item 2, more likely to false alarm to 2 than 3, and so on. This pattern was consistent with earlier findings by Smith et al. (2000) and Dewhurst (2001) and provided the theoretical basis for the current experiments. To foreshadow, the pattern observed by Roediger and DeSoto across items 1-5 will also be replicated across the entire range of output dominance values (i.e., 1-20) in Experiments 1, 2, and 3 of this thesis. This expanded procedure permits a greater range of items with which to study the relation between confidence and accuracy.

Explanations of False Memories in the Categorized List Procedure

What causes false memories for categorized list words? In associative (e.g., DRM) lists, the primary culprit is associative strength (Stadler, Roediger, & McDermott, 1999; Roediger, Watson, McDermott, & Gallo, 2001), especially the associative strength from list items to the critical item. Associative strength is a measure of associative relatedness, which is described as "a normative description of the probability that one word will call to mind a second word" due to contiguity or co-occurrence in the language (Thompson-Schill, Kurtz, & Gabrieli, 1998, p. 440). Indeed, in DRM lists, the studied items that converge on the lure are usually strongly related to the lure itself, and the greater the propensity for list items to be associated to the critical lure, the higher the false recall or false recognition (Roediger et al., 2001). However, word association norms suggest that associative strength is not responsible for errors made on the categorized lists used in the study by Roediger and DeSoto (2011); associative strength is low from studied items to lures in these materials (Nelson, McEvoy, & Schreiber, 1998; Nunes & DeSoto, raw data). This difference may arise partly because DRM lists consist of horizontal associates, whereas categorized lists consist of vertical associates (see Park, Shobe, & Kihlstrom, 2005, for a discussion of the distinction). In other words, associative list items are all interrelated, but categorized lists feature hierarchical relationships between the category name and its items (for example, "bird" and "cardinal"). Thus, associative strength alone fails to explain why high confidence false alarms occur, which is strong evidence that associative relationships do not drive the high confidence false alarms seen in the categorized list procedure.

A second possible explanation of false alarms in the categorized list procedure is that performance is driven by or related to the word frequency effect, which is the observation that infrequent words are more regularly recognized than common words (e.g., Balota & Neely, 1980). Evidence reported in this thesis and elsewhere (e.g., Smith et al., 2000), however, suggest that false alarms are not entirely explained by frequency in the language. This consideration will be thoroughly addressed throughout this thesis.

A third explanation is offered by fuzzy-trace theory (Brainerd & Reyna, 2005; Reyna & Brainerd, 1995; see Brainerd, Reyna, & Zember, 2011, for more recent work). This opponent-process account of memory suggests that two kinds of memory traces drive successful remembering: verbatim and gist. A verbatim trace represents a specific surface form that is connected and interrelated to other verbatim traces through gist (meaning-based) traces. According to fuzzy-trace theory, when verbatim traces are processed, true memories are supported while false memories are suppressed. When gist traces are processed, though, both true and false memories are supported. It is possible that studying categorized lists processes gist traces for those materials to a greater extent than verbatim traces, leading to increased false recognition for lure words that match well with the gist trace. Such a finding would be supported by Dewhurst's (2001) observation that both "know" hits and false alarms were greater for high output dominance category members than low output dominance category members. Although a fuzzy-trace account of categorized list errors may seem plausible, little research has been conducted on this topic (but see Marx & Henderson, 1996).

Thus, we turn to the more established account as proposed by the research of Dewhurst and colleagues (Dewhurst, 2001; Dewhurst & Anderson, 1999; Dewhurst et al., 2009) and Roediger and DeSoto (2011). When subjects study lists of categorized words, two things happen: (1) spreading activation occurs, strengthening the representation of studied and nonstudied items in memory; and (2) subjects may implicitly generate nonstudied associates of the category. Furthermore, subjects are more likely to generate high output dominance associates than low output dominance associates (Dewhurst & Farrand, 2004). At test, subjects are likely to false alarm to nonstudied category items as a result of either of those two processes enumerated above: (1) due to the summation of small degrees of match between the test item and nodes activated by prior spreading activation (e.g., Arndt & Hirshman, 1998; Clark & Gronlund, 1996) or (2) due to a source monitoring error that the test item was actually presented as opposed to implicitly generated (thought about) during encoding. Note that although spreading activation generally occurs via associative networks, some research has shown that it can also occur for categorized materials (e.g., Pusen, Erickson, Hue, & Vyas, 1988). Meanwhile, source monitoring errors that cause false recognition result from contextual or episodic information being associated with the cue. Because the subject mentally generated a nonstudied category associate, and perhaps even rehearsed it during the study phase, there is a greater possibility that that category member will be tagged with additional context, and thus falsely recognized.

According to Dewhurst (2001), spreading activation and source memory errors are associated with know and remember responses, respectively. Because the

experiments in this thesis collect confidence ratings instead of remember/know responses, however, we will be unable to distinguish between these two types of memory errors. For present purposes, we take Dewhurst's account to mean that there should be a greater number of false alarms for higher output dominance category members relative to lower output dominance category members.

Present Research

If the above account is accurate, we might predict that the patterns demonstrated in the experiments discussed earlier would hold throughout the range of output dominance positions. The Dewhurst (2001) study only used one high output dominance and one low output dominance item from each list, however, and the Roediger and DeSoto (2011) experiment used only the five highest output dominance items as lures and did not use those lures as studied items, so it was not possible to evaluate the relationship between the confidence and accuracy of those items. A fuller assessment of this account would require examining the response rates throughout the range of the *typicality gradient* (Barsalou, 1987) — in other words, the entire range of output dominance values, as did Smith et al. (2000). Our goal was to combine the general methodology used by Smith with the theoretical account provided by Dewhurst to better understand recognition memory in the categorized list procedure.

Thus, the goals of the research underlying this thesis were: First, to test a revised procedure intended to conceptually replicate Dewhurst (2001), Smith et al. (2000), and Roediger and DeSoto (2011) and show that false recognition in the categorized list procedure arises as a function of the output dominance of the lure item across a wider

range in the norms (items 1-20 versus 1-5). Specifically, subjects were randomly assigned to study half of the items from different categorized lists of 20 items — either the even or the odd items. These results are then aggregated across subjects to give both a correct recognition rate as well as a false recognition rate for each word as well as each position across all the categorized lists. Second, to use confidence judgments, instead of remember-know judgments, to investigate metacognitive processes at play during this task. Confidence is a more typical judgment for most people and is more relevant to context outside the lab. Third, to further develop a theoretical account of how categorical false recognition occurs. Fourth, to determine whether these false recognition effects are explained entirely by the printed frequency of the materials (i.e., the word frequency effect, e.g., Balota & Neely, 1980). Lastly, in later experiments, we hope to identify methods to reduce high confidence false alarms in the categorized list procedure. Such methods relate to the accounts of why errors arise and possibly provide additional support for these accounts.

Experiment 1

Experiment 1 served primarily as a vehicle to revise and improve the categorized list procedure (which was also used in Experiments 2 and 3 of this thesis). As described earlier, categorized list methods have previously varied from study to study, so we felt it important to design a new categorized list procedure that integrates the lines of research and contributions by both Dewhurst (2001) and Smith et al. (2000). Specifically, we synthesized Dewhurst's operationalization of output dominance with Smith's approach of assessing performance over a wide range of these values.

To test this revised categorized list procedure, we evaluated whether it would replicate earlier work with categorized lists by showing that false and (and perhaps correct) recognition for category members is a direct function of the output dominance of those members. In this experiment, subjects studied lists of 10 words (either the even or odd list positions, counterbalanced across subjects) from 12 categorized lists and were then given a yes/no recognition test over the (1) studied words, (2) alternate words (nonstudied words from the same lists that were studied by the subjects in the other counterbalancing condition), and (3) words from nonstudied categories. After judging whether each word was old or new, subjects indicated their confidence in the recognition judgment on a scale from 0-100.

To assess cognitive performance for targets (i.e., studied words) and lures (i.e., related and unrelated lures), we the calculated response rates and confidence ratings for the different item types as well as the effects of output dominance on correct and false recognition (i.e., hit and false alarm rates). To assess metacognitive performance for these items, we calculated absolute accuracy (i.e., calibration) and relative accuracy (i.e., resolution). Last, we investigated the effects of output dominance on measures of corrected recognition as measured by two different models.

We predicted, based on theories and prior results by Dewhurst (2001), Smith et al. (2000), and Roediger and DeSoto (2011), that false recognition would be a declining function of output dominance; namely, subjects would be more likely to false alarm to higher output dominance category members relative to lower output dominance members. We were less certain about the pattern for correct recognition. The standard

word frequency effect in recognition memory would suggest that an item class that receives the greatest proportion of correct rejections (i.e., low output dominance items, if our earlier prediction is correct) should also receive the greatest proportion of hits (i.e., the mirror effect, as explained by Glanzer & Adams, 1985). In pilot studies (Roediger & DeSoto, 2011), however, correct recognition rates did not differ between high output dominance items (i.e., 1-5 in the norms) and low output dominance items (i.e., 21-25) even though they differed in frequency. Considering these results, we expected that the hit rate would not change as a function of output dominance in this task even though the false alarm rates would (contrary to the findings in the word frequency effect literature).

We also predicted that metacognitive monitoring for studied words would be more accurate than for related lures, due to the deceptive nature of these items (e.g., their similarity to studied items). We expected that calibration and resolution would show the standard pattern as described by Roediger et al. (2011): that confidence and accuracy would indeed be related for hits (with higher confidence associated with greater accuracy), but that there would nevertheless be a high likelihood of false recognition, particularly for related lure items, and that the confidence-accuracy correlation would be greatly reduced for lures.

Participants

Forty-four students from Washington University in St. Louis participated for either course credit or payment.

Materials

Twelve categorized word lists were selected from the Van Overschelde et al. (2004) category norms. These were created by asking a large sample of subjects to generate as many members as possible of a given category (e.g., "a bird"). These responses were aggregated, producing lists of category items that were ordered from the most frequently mentioned category member in the first list position (e.g., "eagle" in Position 1) to the least frequently mentioned category member in the last list position (e.g., "raven" in Position 20).

The first 20 words from each of the 12 selected categorized lists were used as the stimulus set. If any words appeared twice in the stimulus set or could be categorized in two or more of the selected categories (e.g., "squash" is both a vegetable and a sport), the word was removed from any lists it appeared in and the appropriate twenty-first word was added, creating a new 20-word list. The 12 lists can be found in the Appendix.

To examine whether output dominance (i.e., list position) was related to word frequency, a Pearson correlation coefficient was calculated between each word's output dominance value (ranging from 1 to 20) and three different measures of word frequency taken from the English Lexicon Project (Balota et al., 2007): Kučera and Francis (1967), HAL (Lund & Burgess, 1996), and log HAL. No significant relationship was found between output dominance and word frequency as measured by Kučera and Francis, $r(190) = -.102$, $p = .161$. Moreover, only a weak relationship was found between output dominance and word frequency as measured by HAL, $r(234) = -.137$, $p = .036$, as well as between output dominance and log HAL, $r(234) = -.304$, $p < .001$. Thus, although there is some degree of relation between word frequency and output dominance, the

relationship is not strong. Because the relation is significant, however, we will partial out the effects of word frequency when investigating the effects of output dominance on response rates.

A recording was made of a female speaker reading the words with a Logitech desktop microphone in tandem with Apple GarageBand software installed on an Apple MacBook Pro. The speaker read the category name, paused for four seconds, then read each of the category items in random order at a rate of one word per two seconds. Digital post-processing was used to eliminate any extraneous sounds from the recording.

Experiments 1 and 2 were programmed in E-Studios E-Prime 2.0 and Experiment 3 was programmed in Adobe Flash CS 4 (Weinstein, 2011) and hosted on a secure Internet server. Subjects came into the laboratory and were tested on Dell personal computers running Microsoft Windows XP.

Design

Subjects were assigned randomly to one of two counterbalancing groups. One group was presented with the odd-numbered words from the first six lists and the even-numbered words from the last six lists. The other group was presented with the alternate words — the even-numbered words from the first six lists and the odd-numbered words from the last six lists. The experiment consisted of three phases: (1) the study phase, (2) the distractor phase, and (3) the recall test phase.

Procedure

During the study phase, each subject heard the audio recording of the female speaker reciting the category names and items of the 12 lists assigned to the subject's

counterbalancing group. For each list, the subject heard the category name (e.g., "a bird"), followed by a four-second pause, then the corresponding words from that list. Once the category name and all 10 words were presented, the procedure was repeated with the remaining lists, in random order. The experiment then continued to the distractor phase.

During the distractor phase, subjects were given five minutes to list as many of the United States presidents as possible, then order as many of the presidents as possible. This task was included to eliminate any contributions to recognition memory performance from words still in short-term memory.

During the recognition test phase, subjects were given a yes-no recognition test of 360 words, randomly presented (for each subject) one at a time: the 120 studied words from each of the 12 lists, the 120 related (alternate) lures from the same 12 lists, and 120 unrelated lures taken from 12 new categories. Thus, one-third of items at test were studied words (targets) and two-thirds were distractors; half of the distractors were related lures, and the other half of the distractors were unrelated lures. Subjects were not told about the distribution of the word types at test during the experiment.

Subjects indicated with a mouse click whether they believed each word to be old (studied) or new (nonstudied). After subjects made this recognition judgment, they reported confidence by clicking and dragging with the mouse cursor on an on-screen graphical slider that ranged from 0 (*not at all confident*) to 100 (*entirely confident*).

The entire experiment lasted about 60 minutes.

Results

We divided our analyses into three sections: (1) recognition of studied words, (2) recognition of related and unrelated lures, and (3) effects of output dominance on measures of corrected recognition (adjusted by taking into account both hit and false alarm rates). These analyses involved calculating response rates and confidence ratings and applying hierarchical multiple regression to investigate the effects of output dominance on hit and false alarm rates. Because we were interested in the effects of output dominance above and beyond the effects of word frequency, we first eliminated (i.e., partialed out) the effects of word frequency on response rates in these regression analyses. We also examined metacognitive performance for studied words and related and unrelated lures. Although we were most interested in performance for related and unrelated lures, our report also includes results for studied items in the interest of providing all the data from our test. The third analysis investigated the effects of output dominance on corrected recognition (as measured by hits minus false alarms; Macmillan & Creelman, 2004) and discrimination (as measured by *d'*).

Recognition of Studied Words

Hit rate and confidence ratings. Hit rate and confidence ratings for hits and misses were calculated for studied items as basic measures of subjects' cognitive and metacognitive performance. These means can be found in the top row of Tables 1 and 2, respectively. Subjects correctly responded "old" to studied words approximately 75% of the time (hit rate = .73).

Table 1

Response rates as a function of word type in Experiment 1.

Word Type	Hits	Misses	Correct Rejections	False Alarms	d’
Studied Words	.73 (.19)	.27 (.19)	--	--	--
Related Lures	--	--	.61 (.15)	.39 (.15)	0.89
Unrelated Lures	--	--	.90 (.12)	.10 (.12)	1.89

Note. Standard deviations presented in parentheses.

Table 2

Confidence as a function of word type and response in Experiment 1.

Word Type	Hits	Misses	Correct Rejections	False Alarms
Studied Words	82.6 (11.15)	51.1 (16.66)	--	--
Related Lures	--	--	55.8 (16.37)	62.7 (12.83)
Unrelated Lures	--	--	62.7 (22.36)	52.7 (17.02)

Note. Confidence judgments were made on a 0-100 scale. Standard deviations presented in parentheses.

When subjects correctly responded "old," they were more confident ($M = 82.6$) than when they incorrectly responded "new," ($M = 51.1$), $t(43) = 15.66$, $p < .001$.

Effects of output dominance on the hit rate. To investigate the relationship between output dominance and hit rate, we used hierarchical multiple regression to eliminate potentially confounding effects of word (i.e., printed) frequency. These results can be found in Table 3. Multiple stepwise regression was used to assess the effects of output dominance on the hit rate while controlling for printed frequency as measured by log HAL (Hyperspace Analogue to Language; Lund & Burgess, 1996). HAL scores come from an analysis of 160 million words taken from online newsgroups in the 1990s; we obtained these values from the English Lexicon Project (Balota et al., 2007). Balota et al. recommend the use of log HAL as a measure of word frequency because it successfully controls for words of outlying frequency (i.e., those that are extremely common or uncommon in the language). Printed frequency was entered into the regression on its own in the first step, followed by output dominance in the second step. Printed frequency significantly predicted the hit rate and explained a low but statistically significant proportion of variance in hit rate, meaning that subjects were more likely to correctly respond "old" to low frequency words relative to high frequency words. Thus the usual word frequency effect in hit rates was confirmed. Output dominance was not shown to predict hit rate beyond printed frequency, however; thus, the position of the categorized word in the norms (i.e., the output dominance) did not affect the tendency to correctly call a studied word "old."

Table 3

Summary of multiple regression analysis for effects of output dominance on hit rate in Experiment 1.

Variable	*B*	*SE B*	β
Step 1			
Printed frequency	-.01	.00	-.17*
Step 2			
Printed frequency	-.01	.00	-.20*
Output dominance	-.00	.00	-.10

Note. $R^2 = .03$ for Step 1 ($p < .05$), $\Delta R^2 = .01$ for Step 2 ($p > .05$). *$p < .05$.

Metacognitive measures. Depicting the relationship between confidence and accuracy is a complicated task, and different analyses can often lead to different results (Roediger et al., 2011). As such, a thorough assessment of metacognitive performance requires considering the data in multiple ways. We chose to measure subjects' absolute and relative accuracy using three methods suggested by Roediger and colleagues. To measure absolute accuracy, we computed both calibration plots and accuracy plots. To measure of relative accuracy, we calculated the Goodman-Kruskal gamma (Nelson, 1984).

Absolute accuracy for studied words: Confidence-accuracy correlations. Figure 1 depicts the across-subjects calibration plot for studied words. This plot illustrates each subject's hit rate for studied items and average confidence for studied items. The correlation between these 44 points (one for each subject), computed with a Pearson *r*, represents whether subjects that were more confident on average also tended to be more accurate on average.

Calibration was high, $r(42) = .68$, $p < .001$, which means that on average for studied words, subjects who were more confident were also more likely to be accurate. As shown by Figure 1, though, two subjects appear to have unusually low hit rates. When these subjects are removed, the correlation remains high, $r(42) = .62$, $p < .001$.

Relative accuracy for studied words: Gamma. Gamma was also calculated for each subject as a measure of memory resolution, or relative accuracy. This statistic makes use of the ordinal ranking of subjects' confidence ratings and accuracy for the items that were assigned each rating. Gamma correlations, like Pearson correlations,

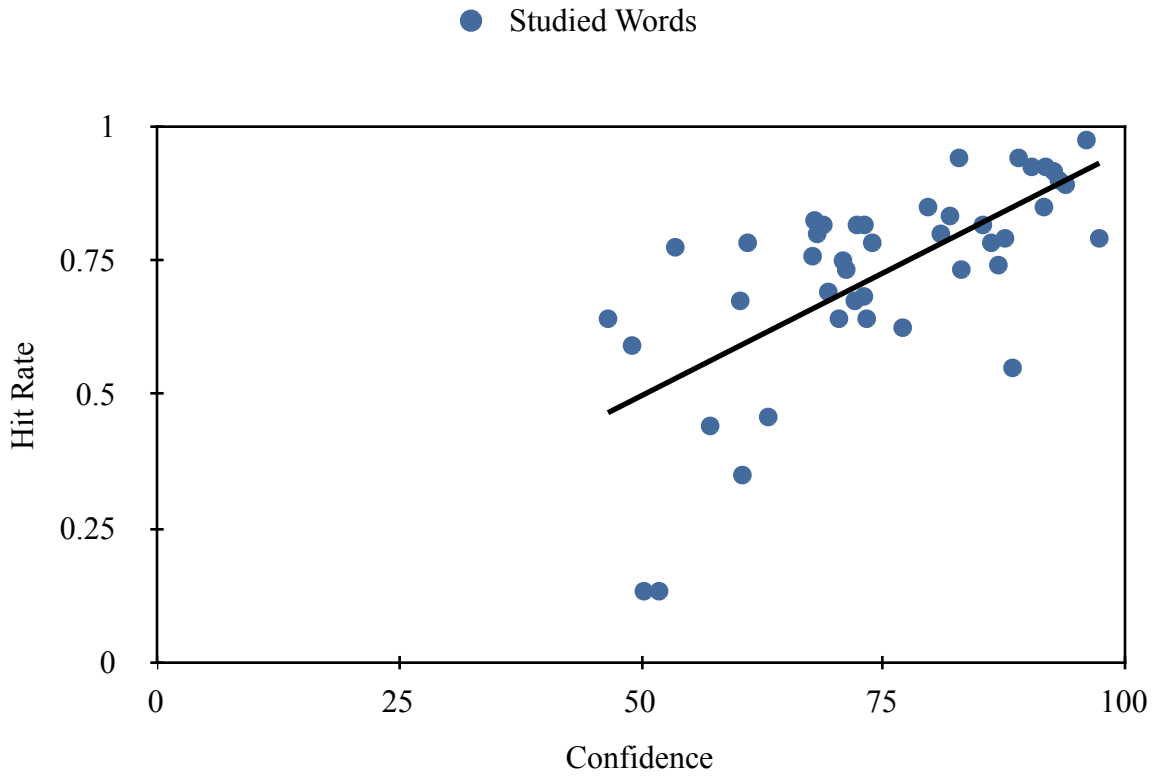


Figure 1. The across-subjects confidence-accuracy plot for studied words in Experiment 1. Each point represents a given subject's average confidence and hit rate across all studied words.

range from -1 to 1; a negative gamma score would indicate that subjects were more confident on inaccurate than accurate items, a zero gamma indicates no relationship between a subject's own confidence and accuracy, and a positive gamma indicates that subjects were more confident on accurate than inaccurate items.

The gamma correlation was high (M = .73), meaning that subjects tended to assign higher confidence ratings when correctly responding "old" (hits) compared to incorrectly responding "new" (misses) to studied words.

In sum, subjects regularly correctly recognized studied words when they were presented at test. A standard word frequency effect was found: the less common the studied word was in the lexicon, the more likely a subject was to correctly respond "old" when it was presented at test. The size of this effect (as measured by variance explained), however, was small. Metacognitive analyses showed that subjects who were more confident for studied words were also more likely to recognize these items correctly. Subjects tended to show accurate metacognitive monitoring for studied words, as well, mostly assigning higher confidence ratings to those studied words to which they correctly responded "old."

Recognition of Related and Unrelated Lures

False alarm rates and confidence ratings. False alarm rates for related and unrelated lures and confidence ratings can be found in the second and third rows of Tables 1 and 2, respectively. Subjects incorrectly responded "old" to related lures (false alarm rate or FAR = .39) more often than they incorrectly responded "old" to unrelated lures (FAR = .10), $t(86) = 9.76$, $p < .001$. Turning to confidence ratings, subjects were

also more confident in their false alarms to related lures (M = 62.7) than to unrelated lures (M = 52.7), $t(39) = 3.17$, $p = .003$. T-tests additionally confirmed that, for related lures, subjects were more confident for false alarms than for correct rejections (M = 55.8), $t(44) = 4.20$, $p < .001$, but this relationship did not exist for related lures; subjects were no more confident for correct rejections (M = 62.7) than for false alarms, $t(39) = 1.71$, $p = .095$. Five subjects were excluded from the analyses involving unrelated lures because they correctly rejected each of these items.

Effects of output dominance on the false alarm rate. Multiple regression was used to assess the effects of output dominance on false alarm rate for related lures while controlling for printed frequency (as measured by log HAL). Unrelated lures were excluded from this analysis (because they have no output dominance value). These results can be found in Table 4. Printed frequency significantly predicted false alarm rate and explained a statistically significant proportion of variance in false alarm rate, such that subjects were more likely to false alarm to items low in printed frequency. Critically, however, when output dominance was added to the regression equation, output dominance was the sole predictor of the false alarm rate. As we predicted, subjects were more likely to incorrectly respond "old" (i.e., false alarm) to related lures of higher output dominance than lower output dominance.

Metacognitive measures.

Absolute accuracy for related and unrelated lures: Confidence-accuracy correlations. Figure 2 depicts the across-subjects calibration plot for related lures and unrelated lures on the left and right panels, respectively. For related lures, calibration

Table 4

Summary of multiple regression analysis for effects of output dominance on false alarm rate in Experiment 1.

Variable	*B*	*SE B*	β
Step 1			
Printed frequency	-.01	.00	-.16*
Step 2			
Printed frequency	.00	.00	.02
Output dominance	-.01	.00	-.45*

Note. R^2 = .03 for Step 1 ($p < .05$), ΔR^2 = .18 for Step 2 ($p < .05$). *$p < .05$.

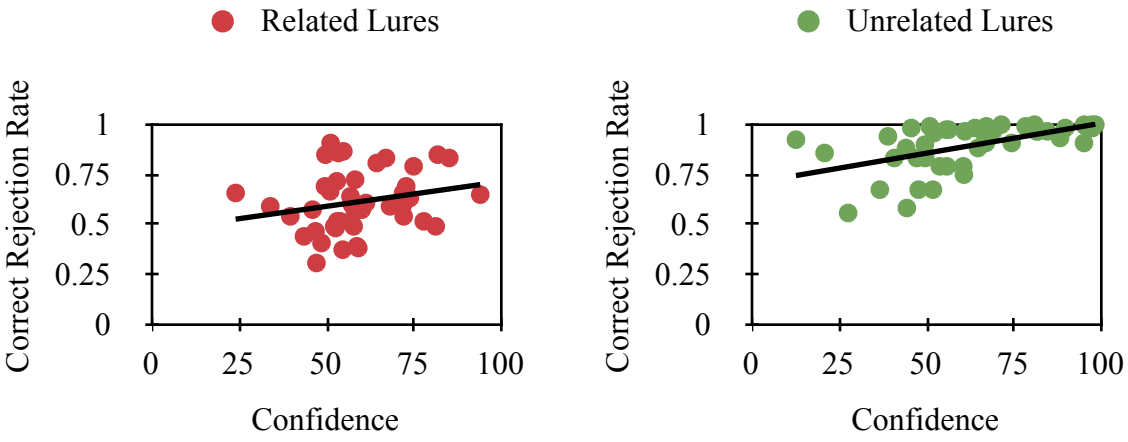


Figure 2. The across-subjects confidence-accuracy plot for related lures (left panel) and unrelated lures (right panel) in Experiment 1. Each point represents a given subject's average confidence and correct rejection rate across all related lures.

was not significantly different from zero, $r(42) = .23$, $p = .141$, which means that there was no significant relationship between confidence and accuracy for related lures across subjects. In contrast, for unrelated lures, the relationship between confidence and accuracy was strong, $r(42) = .55$, $p < .001$. Subjects who were more confident when responding to unrelated lures were also more likely to make correct rejections. Although a relationship between confidence and accuracy was not found for related lures but was found for unrelated lures, a direct comparison using the Fisher r-to-z transformation revealed that the difference between these two correlations was only marginally significant, $p = .082$.

Absolute accuracy for related and unrelated words: Accuracy plot. Figure 3 depicts the accuracy plot for related lures and unrelated lures. This curve depicts subjects' ability to discriminate between studied words and related lures, as well as studied words and unrelated lures. For this analysis, as suggested by Roediger et al. (2011) and Wixted (personal communication, October 14, 2011), we turned the 0 to 100 confidence scale into a -100 to 100 scale such that -100 confidence represented maximum confidence "new" responses and 100 confidence represented maximum confidence "old" responses. (Previously, a confidence of 100 could be associated with either an "old" or "new" response.) We then grouped confidence ratings into 20-point bands (0-19, 20-39, etc.) and observed the proportion of correct and incorrect responses falling into each band. On the right half of Figure 3, each point on the related lures line represents the proportion of hits to studied words falling in that confidence band, divided by the proportion of hits to studied words for that band plus the proportion of false alarms to

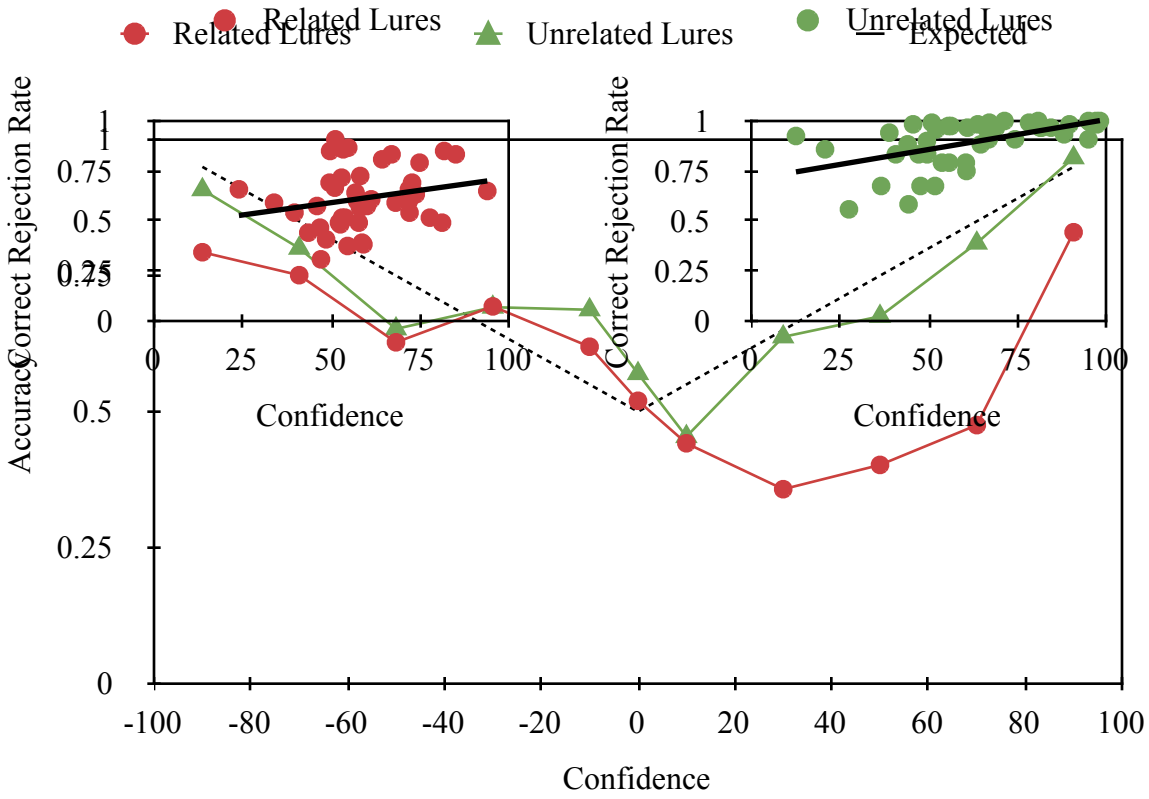


Figure 3. The accuracy plot for Experiment 1, representing discrimination between studied words and related lures and discrimination between studied words and unrelated lures. The dotted line represents expected accuracy; points above this line represent underconfidence, whereas points below this line represent overconfidence.

related lures in that band (HR ÷ [HR + $FAR_{related}$]). Each point on the unrelated lure line is calculated similarly: HR ÷ (HR + $FAR_{unrelated}$). On the left half of Figure 3, each point on the related lures line represents the proportion of correct rejections to related lures falling in that confidence band, divided by the proportion of correct rejections to related lures for that band plus the proportion of misses to studied items in that band (CR ÷ [CR + $MISS_{related}$]). For unrelated lures, the equation is: CR ÷ (CR + $MISS_{unrelated}$). Note that although this plot resembles a traditional calibration curve (e.g., Dunlosky & Metcalfe, 2008), it is not the same, partly because it puts the emphasis on accuracy rather than on individual word type.

When a perfectly calibrated subject makes a 0 confidence rating, he or she should have a 50% chance of correctly responding to the recognition prompt. Thus, at 0 confidence on the figure, expected accuracy is 50. This value climbs linearly in either direction such that expected accuracy when given a confidence rating of -100 or 100 is 100%. The dotted line on Figure 3 represents expected accuracy throughout the confidence scale. Underconfidence is represented when points deviate above the dotted line (i.e., subjects are more accurate than their confidence ratings indicate); overconfidence is shown when points deviate below the dotted line.

Figure 3 shows several important patterns. First, a general relationship between confidence and accuracy is shown, such that as confidence increases (moving from the center of the calibration plot to the left or right edges), accuracy does as well. Despite this general relationship, however, subjects still made a great number of high confidence errors when discriminating between studied words and related lures; when maximum

confidence judgments were assigned to an "old" or "new" response, these judgments were only 80% likely to be correct. Additionally, when discriminating between studied words and related lures, subjects were less accurate than their confidence judgments suggested they would have been (i.e., they were overconfident) when they assigned ratings above 20 when responding "old."

Relative accuracy for related and unrelated lures: Gamma. For related lures, the gamma correlation was negative (M = -.21), meaning that a subject was less likely to correctly reject these items as confidence increased. For unrelated lures, the gamma correlation was positive (M = .16). A t-test revealed a significant difference between these two values, $t(79) = 6.32$, $p = .003$. The related lure gamma was demonstrated to be significantly different from zero, $t(43) = 4.24$, $p < .001$, but the unrelated lure gamma was not, $t(38) = 1.90$, $p = .065$.

In sum, subjects were likely to falsely recognize related lures. When a false alarm occurred, it was more likely to be to a word of high output dominance than low output dominance. For related lures, no relationship between confidence and accuracy was found, such that subjects who were more confident for related lures were no more likely to be accurate for them. Further, subjects were more likely to assign higher confidence ratings when incorrectly responding to related lures than when correctly responding to them. Within-subjects resolution was also poor; a subject was more likely to false alarm to a related lure as his or her confidence increased.

Performance for unrelated lures, in contrast, was generally very strong. Subjects made very few false alarms to these items and showed accuracy commensurate with their confidence ratings.

Effects of Output Dominance on Corrected Recognition

Hits Minus False Alarms. One measure of corrected recognition obtained by subtracting the false alarm rate from the hit rate for each item. Stepwise regression was used to assess the effects of output dominance on corrected recognition while controlling for printed frequency (as measured by log HAL). This regression was run across the 240 studied words that served as either studied words or related lures, depending on counterbalancing group. Thus, unrelated lures were excluded from this analysis because they had no hit rate or output dominance value. These results can be found in Table 5. Printed frequency significantly predicted corrected recognition and explained a statistically significant proportion of variance in corrected recognition. Critically, however, output dominance did predict corrected recognition above and beyond printed frequency. These results, taken together, mean that subjects were more accurate for less common words in the lexicon or words of lower output dominance than more common words in the lexicon or higher output dominance words.

Discrimination. As a second measure of corrected recognition, we calculated *d'* for each item. Multiple regression was used to assess the effects of output dominance on *d'* while controlling for printed frequency (as measured by log HAL). Unrelated lures were excluded from this analysis because they had no output dominance value. These results can be found in Table 6. As with hits minus false alarms, printed frequency did

Table 5

Summary of multiple regression analysis for effects of output dominance on corrected recognition as measured by hits minus false alarms in Experiment 1.

Variable	*B*	*SE B*	β
Step 1			
Printed frequency	-.01	.00	-.27*
Step 2			
Printed frequency	-.01	.00	-.16*
Output dominance	.01	.00	.37*

Note. R^2 = .07 for Step 1 ($p < .05$), ΔR^2 = .13 for Step 2 ($p < .05$). *$p < .05$.

Table 6

Summary of multiple regression analysis for effects of output dominance on corrected recognition as measured by discrimination (d') in Experiment 1.

Variable	*B*	*SE B*	β
Step 1			
Printed frequency	-.05	.01	-.25*
Step 2			
Printed frequency	-.02	.01	-.12
Output dominance	.02	.00	.43*

Note. R^2 = .06 for Step 1 ($p < .05$), ΔR^2 = .17 for Step 2 ($p < .05$). *$p < .05$.

significantly predict *d'* as well as a significant proportion of variance in *d'*. However, when output dominance was added in the second step, it served as the only significant predictor of *d'*. These results indicate that subjects showed better discrimination for items of low output dominance than items of high output dominance.

Discussion

The main finding of Experiment 1 was that subjects were more likely to false alarm to high output dominance categorized list items than low output dominance items in a

regular fashion across the first 20 items in the category norms. Experiment 1 also revealed that subjects had relatively poor metacognitive awareness when responding to nonstudied categorized list items from studied categories, which was demonstrated by the observation that related lures were the only class of items in which confidence for false alarms exceeded confidence for correct rejections. Related lures also showed poor calibration (especially for high confidence ratings) and negative resolution.

As hypothesized, these results corroborate previous accounts of how errors arise in the categorized list procedure. When subjects study interpolated items from different categorized lists, they may implicitly generate other items from the same category that were never studied (Dewhurst & Farrand, 2004). An item's likelihood of being implicitly generated is closely linked to its output dominance, or position, in the category norms. At test, it is hypothesized that subjects are more likely to false alarm to items of higher output dominance, since these items were more likely to be generated at encoding. These false alarms are likely to be the ones that occur with high confidence because they

represent source monitoring errors and are also likely associated with remember responses in similar remember/know experiments (e.g., Dewhurst, 2001). In addition, other false alarms may result from spreading activation, or the summing of small degrees of match between other studied items and the item at test (Arndt & Hirshman, 1998). These spreading activation false alarms most likely occur with lower confidence or know responses. As such, we recognize that different qualitative bases of responding might have contributed to the errors that occurred in Experiment 1. Because we collected confidence ratings instead of remember/know judgments, however, we cannot differentiate between these bases of responding in the current experiment.

As mentioned in the Introduction, associative strength does not seem to be driving these false alarms, given the lack of associative strength between studied items and the critical lures in categorized lists (Nelson et al., 1998; Nunes & DeSoto, raw data). Dewhurst et al. (2009) further supported this conclusion by reporting associative strength in their experiments; mean associative strength between lures and studied items in DRM lists taken from Stadler et al. (1999) was moderate (M = .17), but it was low (M = .03) between lures and studied items in categorized list materials taken from Van Overschelde et al. (2004). Smith et al. were led by the data from a free association experiment to report similarly:

> "Critical words from associatively structured lists were given as responses in the free association task at a very high rate… critical words from categorically structured lists, however… evoked critical items at a rate of less than one-tenth that found for the associatively structured list words." (p. 339)

The fact that associative strength is low for categorical materials weakens Dewhurst's (2001) theory that high output dominance category members are generated at encoding, however, and it is unclear that this theoretical mechanism explains how high confidence false memories occur. Partly for this reason, it still seems appropriate to consider alternative explanations for our findings. A fuzzy-trace theory account, as discussed earlier, might also account for these data. Fuzzy-trace theory posits that processing verbatim traces increases true and decreases false memories, whereas processing gist traces increases both true and false memories. It is possible that the higher output dominance category items are afforded additional gist processing due to their overall similarity with the category prototype, whereas lower output dominance category items are targeted by additional verbatim processing, possibly because of their distinctiveness. Put another way, a subject trying to memorize a list of birds might be more likely to process individual item information for low output dominance item but focus more on a high dominance item's relation to the bird category or its prototype. This initial account would explain why a greater number of false memories arise for high output dominance category members than low output dominance category members.

Nevertheless, one important conceptual issue here is whether the experimental results reported are manifestations of two other well-known effects in recognition memory: the word frequency effect and the mirror effect. The word frequency effect is the finding that recognition performance is better for low frequency words than high frequency words (Balota & Neely, 1980). The mirror effect is the related observation that the classes of items for which "new" responses are most accurate are also the class of

items for which "old" responses are most accurate (Glanzer & Adams, 1985). For example, in recognition, studied low frequency words are distinctive and may "pop out" when they appear at test, improving the hit rate. Similarly, when such distinctive words appear at test but were not studied, the lack of familiarity or episodic context also improves the correct rejection rate (i.e., subjects feel they would have remembered the word had it been presented). Thus, both "old" and "new" responses are more accurate for low frequency words as compared to high frequency words.

In Experiment 1, we used hierarchical regression to factor out the effects of word frequency and determined that output dominance affects false recognition above and beyond word frequency. In addition to this finding, two additional lines of evidence suggest that the conclusions reported in Experiment 1 are not explained entirely by word frequency: (1) the mirror effect is not present in the reported recognition data; and (2) other empirical studies have controlled for word frequency while varying output dominance and shown similar effects to the ones reported.

A first reason output dominance appears to have a different effect than word frequency on false recognition is the lack of a mirror effect in the dataset. The mirror effect is that in recognition memory, low frequency words should enjoy a greater number of hits and correct rejections than high frequency words. Our data show that the correct rejection rate (i.e., the inverse of false alarm rate) is higher for low output dominance items than high output dominance items. Critically, however, the hit rate was the same over the entire range of output dominance, contradicting the mirror effect pattern. Although the mirror effect is a common finding in recognition memory, Greene (2007)

argued that the mirror effect is not a law of memory, and our findings support that claim. They show that the relationship between response rate for hits, false alarms, and output dominance cannot be attributed to word frequency, suggesting some other mechanism is at play.

Last, other empirical research suggests that the effects of output dominance are separable from the effects of word frequency. In the paper by Dewhurst (2001), printed frequency (i.e., word frequency as measured by Kučera and Francis, 1967) was kept constant while output dominance of lure items varied. Nevertheless, Dewhurst still found results consistent with those reported here — that more false alarm remember responses were made for high output dominance items than low output dominance items. These results led Dewhurst to conclude, "In the present study, high and low [output dominance] items were matched for printed frequency. The present findings therefore show that instance frequency exerts an effect in recognition memory over and above that of printed frequency" (p. 159).

Although word frequency does not seem to describe entirely the Experiment 1 data, it is important to consider the possibility that some of these results may be due to differential processing allotted to distinctive, low output dominance category members. Across four experiments, Jacoby, Craik, and Begg (1979) observed that distinctiveness of an item led to superior memory for it on a later test. The results of Experiment 1 do not show improved recognition for studied items of low output dominance relative to high output dominance, however. Explaining this finding (i.e., the lack of the mirror effect) will be an important pursuit for future research; one tentative hypothesis could be that the

powerful influence of output dominance offsets the effects of distinctiveness of the list items.

Experiment 2

Subjects are thought to commit false alarms in the categorized list procedure because they encode related (nonstudied) category members and thus incorrectly judge at test that these items were studied. Are there any experimental manipulations that can reduce or otherwise affect the likelihood that subjects make these source monitoring errors? One explanation of why subjects are so likely to accept related category members as "old" is offered by fuzzy-trace theory (Brainerd & Reyna, 2005; Reyna & Brainerd, 1995). When subjects study categorized lists, they process both the gist as well as the verbatim traces of the materials. For instance, when subjects study the "birds" list, they extract and process both the individual items in the list as well as a more general sense of "birdness." At test, subjects engage in differing amounts of gist and verbatim processing to make recognition decisions.

In Experiment 1, subjects were highly likely to accept high output dominance members that were not studied as "old." One possible explanation for this, consistent with fuzzy-trace theory, is that the composition of the test induced subjects to rely on gist processing over verbatim processing to make recognition decisions. Because half of the lures in Experiment 1 were from unrelated categories, subjects may have made recognition decisions based on gist for words from studied categories (i.e., 67% of all test items were related to the 12 categories and only 33% were unrelated). Thus, subjects

might have been making recognition judgments using the heuristic "was this item related to the ones I saw?"

If recognition judgments were made in this way, subjects might have compared each item on the test to the gist-based representation extracted during encoding, and higher output dominance category members might have been better matches to this representation than lower output dominance category members. Because fuzzy-trace theory predicts that processing gist information increases both correct and false recognition, subjects might have been more likely to incorrectly respond "old" to higher output dominance category members than lower output dominance members — a result shown for false alarms in our data.

To test this possibility, we conducted a second experiment with the same materials as in Experiment 1, but with all unrelated lure words omitted during the test. Thus, subjects were given the task of discriminating between 120 studied words (10 each from 12 categories) and 120 lure words from the same categories. By omitting the unrelated lures, we expected that subjects might be less likely to use gist processing (i.e., a global similarity heuristic) and that they would be encouraged to rely more heavily on verbatim or controlled processing to make recognition decisions. Changing conditions at test should not affect the implicit generation of related category members at encoding (according to Dewhurst's theory), however, assuming this process occurs.

Thus, if the format of the test list in Experiment 1 lead subjects to make recognition decisions based on gist rather than verbatim information, eliminating unrelated distractors from the list should emphasize verbatim processing to a greater

degree, eliminating or at least attenuating the relationship between output dominance and false alarm rate. If implicit generation of category members occurs during encoding, however, subjects should still false alarm to high output dominance category members. This is because subjects will have source memory errors for the generated items at test regardless of the composition of the test list.

Although Experiment 2 is not a direct test of these two hypotheses, as they are not mutually exclusive, it should provide an opportunity to investigate the nature of the processing that occurs during the categorized list procedure. Additionally, because Experiment 2 is very similar to Experiment 1 — it employs the same pool of subjects, the same stimuli, and the same procedure — it enables us to do a cross-experiment comparison (albeit with some caution).

Participants

Twenty-eight students from Washington University in St. Louis participated for either course credit or payment.

Materials

Experiment 2 used the same materials used in Experiment 1.

Procedure

The study, distractor, and recognition test phases in Experiment 2 were identical to the phases in Experiment 1, except for one change: the 120 unrelated words taken from new categories were omitted from the final test. Thus, the recognition test consisted of 240 words: the 120 studied words and the 120 related lures from the studied lists. As a result, half of the test items were targets and half were distractors. Subjects were not told

of the composition of test items. We expected that omission of unrelated lures would reduce false alarm rates relative to Experiment 1 but would not affect the negative relationship between output dominance and false alarm rate.

Results

As with Experiment 1, the results of Experiment 2 are presented in three sections: (1) recognition of studied words, (2) recognition of related lures, and (3) effects of output dominance on corrected recognition.

Recognition of Studied Words

Hit rate and confidence ratings. Hit rate and confidence ratings can be found in the top row of Tables 7 and 8, respectively. Subjects correctly responded "old" to studied words approximately three-quarters of the time (hit rate = .70). When subjects correctly responded "old," they were more confident (M = 84.2) than when they incorrectly responded "new," (M = 48.5), $t(25) = 18.65$, $p < .001$.

Effects of output dominance on the hit rate. A multiple hierarchical regression analysis was used to assess the effects of output dominance on the hit rate while controlling for printed frequency as measured by log HAL. The results can be found in Table 9. Printed frequency did not explain a significant amount of variance in hit rate. When output dominance was added to the regression equation, however, it did significantly predict hit rate such that subjects were more likely to correctly respond "old" to items of higher output dominance than lower output dominance.

Metacognitive measures.

Table 7

Response rates as a function of word type in Experiment 2.

Word Type	Hits	Misses	Correct Rejections	False Alarms	d’
Studied Words	.70 (.12)	.30 (.12)	--	--	--
Related Lures	--	--	.72 (.13)	.28 (.13)	1.10

Note. Standard deviations presented in parentheses.

Table 8

Confidence as a function of word type and response in Experiment 2.

Word Type	Hits	Misses	Correct Rejections	False Alarms
Studied Words	84.2 (9.47)	48.5 (12.87)	--	--
Related Lures	--	--	59.9 (13.62)	55.1 (12.77)

Note. Confidence judgments were made on a 0-100 scale. Standard deviations presented in parentheses.

Table 9

Summary of multiple regression analysis for effects of output dominance on hit rate in Experiment 2.

Variable	*B*	*SE B*	β
Step 1			
Printed frequency	.00	.00	.04
Step 2			
Printed frequency	.00	.00	-.01
Output dominance	.00	.00	-.18*

Note. R^2 = .00 for Step 1 ($p > .05$), ΔR^2 = .03 for Step 2 ($p < .05$). *$p < .05$.

Absolute accuracy for studied words: Confidence-accuracy correlations. Figure 4 depicts the across-subjects calibration plot for studied words. The correlation was high, $r(26) = .71$, $p < .001$, which means that on average for studied words, subjects who were more confident were also more likely to be accurate.

Relative accuracy for studied words: Gamma. The gamma correlation was high ($M = .79$), meaning that subjects tended to assign higher confidence ratings when correctly responding "old" (hits) compared to incorrectly responding "new" (misses) to studied words.

In sum, subjects were highly likely to correctly respond "old" when presented with a studied item on the recognition test. Items of higher output dominance were more likely to be correctly responded to than items of lower output dominance — a finding inconsistent with the word frequency effect. Subjects who were more confident in responding to studied items were also more likely to be accurate, as shown by both calibration and resolution measures.

Recognition of Related Lures

False alarm rates and confidence ratings. False alarm rates for related lures and confidence ratings can be found in the second and third rows of Tables 7 and 8, respectively. Subjects incorrectly responded "old" to related lures almost two-fifths of the time (FAR = .39). They were similarly confident for false alarms ($M = 55.1$) and correct rejections ($M = 59.9$), $t(25) = 1.69$, $p = .103$.

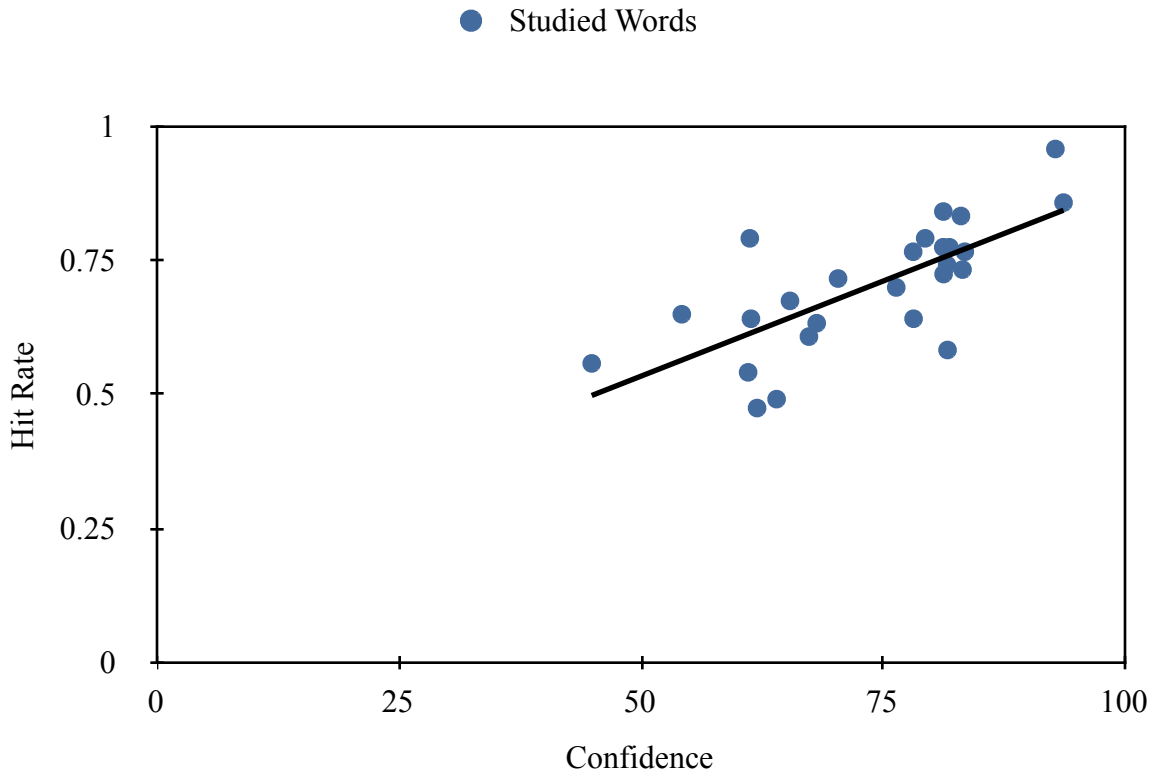


Figure 4. The across-subjects confidence-accuracy plot for studied words in Experiment 2. Each point represents a given subject's average confidence and hit rate across all studied words.

Table 10

Summary of multiple regression analysis for effects of output dominance on false alarm rate in Experiment 2.

Variable	*B*	*SE B*	β
Step 1			
Printed frequency	.01	.00	.11
Step 2			
Printed frequency	.00	.00	-.03
Output dominance	-.01	.01	-.45*

Note. R^2 = .01 for Step 1 ($p > .05$), ΔR^2 = .19 for Step 2 ($p < .05$). *$p < .05$.

Effects of output dominance on the false alarm rate. Multiple regression was used to assess the effects of output dominance on the false alarm rate for related lures while controlling for printed frequency (as measured by log HAL). These results can be found in Table 10. Printed frequency neither predicted false alarm rate nor explained a statistically significant proportion of variance in the false alarm rate. When output dominance was added to the regression equation, however, output dominance significantly predicted the false alarm rate.

Subjects were more likely to incorrectly respond "old" (i.e., false alarm) to related lures of higher output dominance than lower output dominance.

Metacognitive measures.

Absolute accuracy for related lures: Confidence-accuracy correlations. Figure 5 depicts the across-subjects calibration plot for related lures. For these items, the correlation was not significantly different from zero, $r(26) = .28$, $p = .169$, which means that there was not a significant relationship between confidence and accuracy for related lures across subjects.

Absolute accuracy for related lures: Accuracy plot. Figure 6 depicts the accuracy plot for related lures, which depicts subjects' ability to discriminate between studied words and related lures at test. Responses of "new" or "old" that were made with higher accuracy were indeed more likely to be correct, but some errors were still made when a confidence rating of 100 was assigned to a recognition decision. In particular, subjects were less accurate than their confidence ratings would have predicted when they responded "old" with mid-range (20-80) confidence.

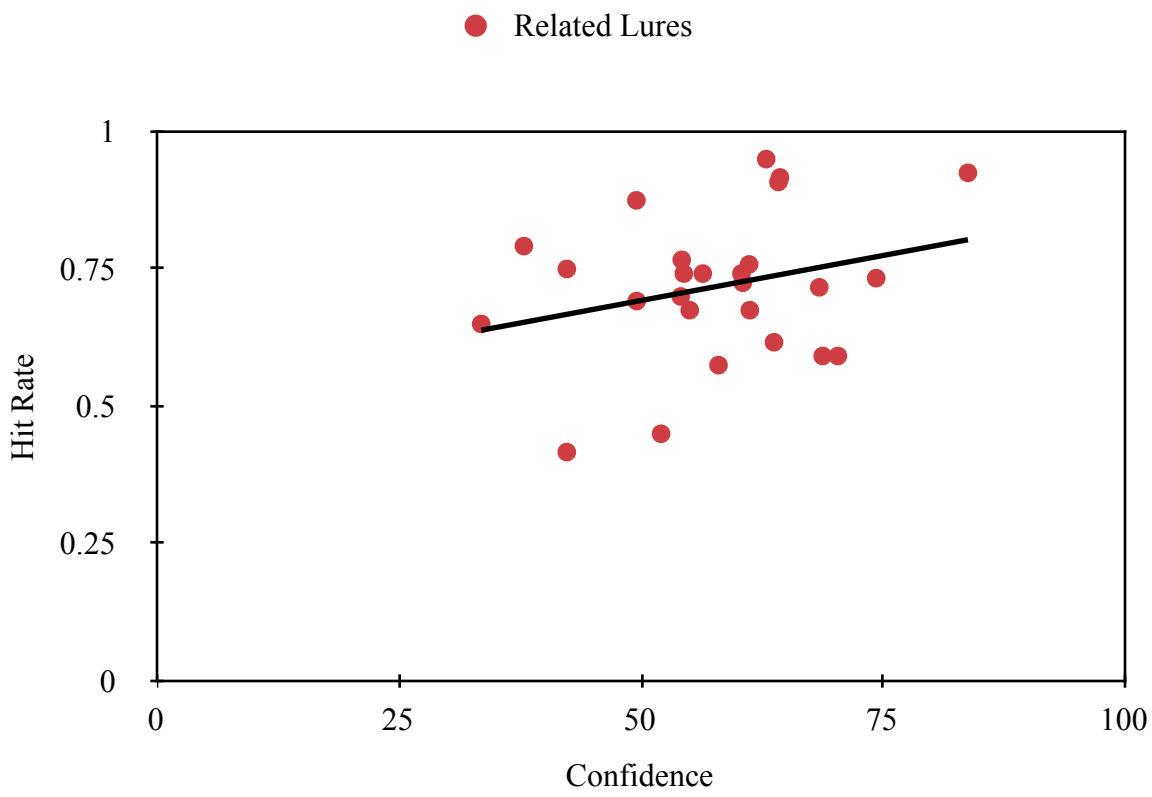


Figure 5. The across-subjects confidence-accuracy plot for related lures in Experiment 2. Each point represents a given subject's average confidence and correct rejection rate across all related lures.

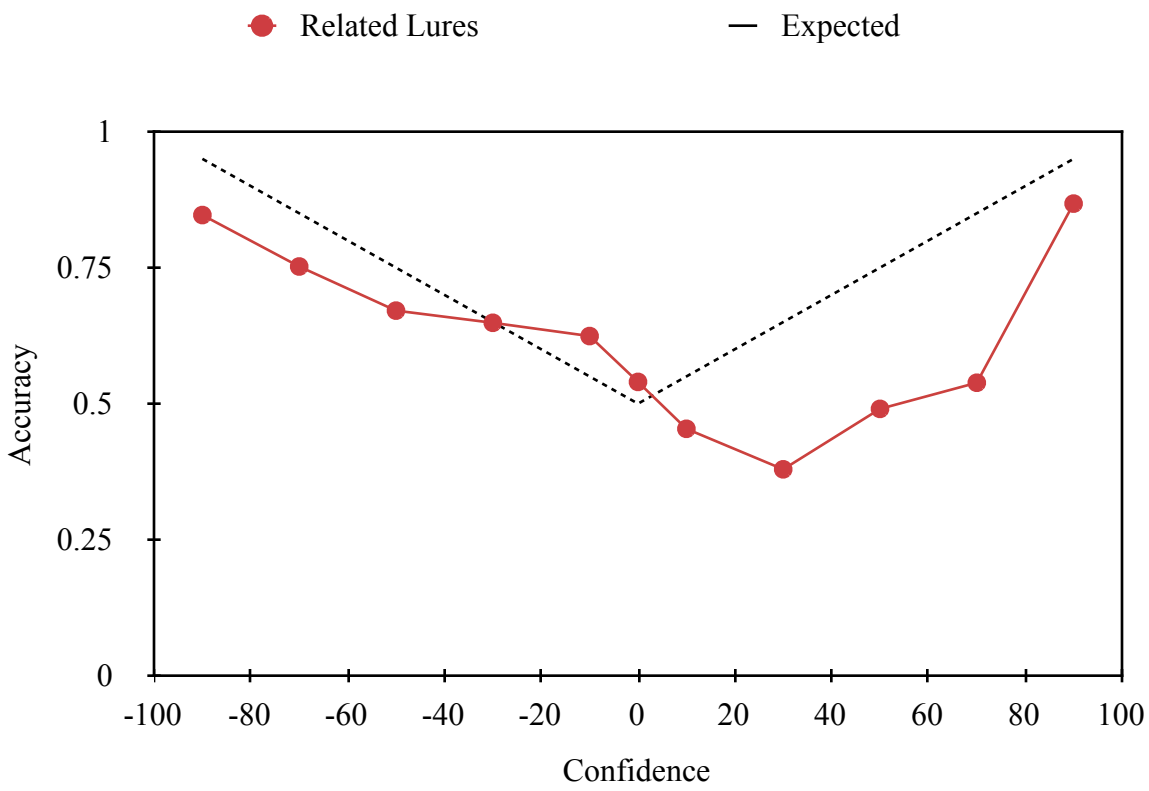


Figure 6. The accuracy plot for Experiment 2, depicting subjects' ability to discriminate between studied words and related lures. The dotted line represents expected accuracy; points above this line represent underconfidence, whereas points below this line represent overconfidence.

Relative accuracy for related lures: Gamma. The gamma correlation was negative (*M* = -.13), meaning that a given subject was less likely to assign higher confidence ratings when correctly rejecting responding "new" (as compared to incorrectly responding "old").

In sum, subjects were likely to false alarm to related lures and did so more often for higher output dominance category members than for lower output dominance members. Subjects who were more confident when responding to related lures were no more or less accurate for that item type, and were generally less accurate than they believed they were when responding "old."

Effects of Output Dominance on Corrected Recognition

Because the corrected recognition and discrimination results were so similar in Experiment 1, we only report the hits minus false alarms data here. Stepwise regression was used to assess the effects of output dominance on corrected recognition while controlling for printed frequency (as measured by log HAL). The results can be found in Table 11. Printed frequency neither significantly predicted corrected recognition nor explained a statistically significant proportion of variance in corrected recognition. When output dominance was added to the regression equation, however, only output dominance was related to corrected recognition such that subjects were more accurate overall for words of lower output dominance than higher output dominance.

Comparison of Experiments 1 and 2

Because Experiments 1 and 2 used subjects from the same population and tested over a very similar set of materials, we believe it useful to present a brief comparison of

Table 11

Summary of multiple regression analysis for effects of output dominance on corrected recognition as measured by hits minus false alarms in Experiment 2.

Variable	*B*	*SE B*	β
Step 1			
Printed frequency	.00	.00	-.06
Step 2			
Printed frequency	.00	.00	.01
Output dominance	.00	.00	.25*

Note. R^2 = .00 for Step 1 ($p > .05$), ΔR^2 = .06 for Step 2 ($p < .05$). *$p < .05$.

Experiments 1 and 2. A comparison of response rates, confidence ratings, confidence-accuracy correlations, and gamma correlations can be found in Table 12. As this table shows, hit rate in Experiment 2 (M = .73) did not differ from hit rate in Experiment 1 (M = .70), $t(68) = .59$, $p = .555$. In Experiment 1, no relationship was found between output dominance and hit rate. Experiment 2, however, an effect was found such that subjects were more likely to correctly respond "old" to items of higher output dominance. This finding may imply that subjects invoked a greater degree of recollective processing when they encountered high output dominance items. Experiments 1 and 2 also showed similar absolute accuracy (as demonstrated by similar confidence-accuracy plots and accuracy plots) and similar relative accuracy (as demonstrated by gamma correlations). Moreover, a comparison of Experiments 1 and 2 does show that eliminating unrelated lures at test did reduce the false alarm rate — subjects were much more likely to false alarm to related lures in Experiment 1 (M = .39) compared to Experiment 2 (M = .28), $t(68) = 2.90$, $p = .005$. The regression analyses between Experiments 1 and 2, however, were similar — output dominance was a significant predictor of false alarm rate regardless of whether unrelated lures were present on the recognition test. Subjects exhibited similar absolute accuracy and relative accuracy for related lures in both experiments.

Further, when looking at corrected recognition, regression showed that output dominance predicted recognition performance in both Experiments 1 and 2. Overall, performance was best for lower output dominance items than higher output dominance items. In sum, removing unrelated distractors from test reduced the overall false alarm rate and marginally improved metacognitive monitoring. Subjects were likely to false

Table 12

Comparison of Experiment 1 and Experiment 2 response rates and confidence ratings.

False alarm response proportions and confidence ratings are calculated for related lures.

	Experiment 1	Experiment 2	Difference
Response Proportions			
Hits	.73	.70	.03
False Alarms	.39	.28	.11
Confidence			
Hits	82.6	84.3	-1.7
Misses	50.1	48.5	1.7
False Alarms	62.7	59.9	2.8
Correct Rejections	55.8	55.1	0.7
Confidence-Accuracy Correlations			
Hits	.68	.71	-.03
False Alarms	.23	.28	-.06
Gamma Correlations			
Hits	.73	.79	-.06
False Alarms	-.21	-.13	.07

alarm to high output dominance related lures regardless of whether unrelated lures were present at test, however.

Discussion

The main finding of Experiment 2 was that subjects were more likely to falsely recognize (i.e., false alarm to) and to correctly recognize (i.e., hit) higher output dominance items. The comparison of Experiments 1 and 2 demonstrates that removing unrelated lure words from the test in the categorized list procedure does not affect the relationship between output dominance and false alarms. Removing unrelated distractors reduced the false alarm rate by 39% relative to the original rate. Although there were fewer false alarms overall in Experiment 2 relative to Experiment 1, the fact that a negative relationship continued to exist between output dominance and false alarms indicates that subjects still generated associates of the studied categories at encoding. As high confidence false recognition is thought by Dewhurst (2001) to stem from source monitoring errors resulting from nonstudied associates being generated at encoding, the persistence of this negative relationship in both Experiments 1 and 2 suggests this process is still occurring. These findings are consistent with the research of Dewhurst et al. (2009), which concluded that false alarms to categorized list words are a function of associations made at encoding and not at retrieval; in other words, the composition of the test list does not much affect the relationship between output dominance and the false alarm rate.

Although removing unrelated lures from test did not disrupt the linear relationship between false alarm rate and output dominance, it did reduce the overall false alarm rate.

One explanation for this reduction is that when these items are eliminated, subjects are motivated to invoke recollective processing to a greater degree than they were in Experiment 1, because all items appearing on test are similar (i.e., came from studied categories and had the same superordinates). This controlled type of processing (recollection) is less error-prone. Similarly, a fuzzy-trace explanation might argue that removing unrelated distractors reduces the degree to which subjects process gist at test, also reducing false alarm rate for nonstudied categorized items. Either of these explanations would still be compatible with Dewhurst's implicit generation theory; in fact, it is possible that Experiment 2 also serves as a first step toward distinguishing these two possible influences (i.e., gist or activation and source monitoring) on false recognition.

Taken together, the results of Experiments 1 and 2 are consistent with the previously established account of false alarms in the categorized list procedure as set forth by Dewhurst (2001). Subjects studied lists of category items and likely generate other items of those studied categories. At test, subjects likely experienced source monitoring errors, regardless of the composition of the test list. These errors lead subjects to falsely recognize nonstudied category members items, with errors being more likely for high output dominance lures. We hypothesized that removing unrelated distractors from test would make subjects less likely to use a global similarity heuristic (or gist processing). Although this may have occurred, the findings of Experiment 2 suggest that eliminating unrelated distractors did not make subjects less likely to confirm high output dominance lures as studied.

Experiment 3

Experiment 3 targeted processes occurring at encoding and at test. This experiment was inspired by Benjamin (2001), who investigated the effects of study repetition on recognition memory. In the first experiment of his article, Benjamin presented young and older adults with 10 DRM lists. Five of the DRM lists were presented once; the other five lists were presented to subjects three consecutive times. Benjamin found that presenting a list three times increased feelings of familiarity both for the items that were repeated as well as their corresponding lures for all subjects, but he also found that younger adults exhibited an ability to counter false familiarity for nonstudied associates through an increased reliance on recollection. Thus, as a result of repetition, both young and older adults showed improved correct recognition, but older adults suffered from additional false alarms (because of their weaker ability to counter false familiarity). For young adults, the tendency to false alarm decreased with repetition of the list.

The patterns described by Benjamin are also supported in the literature on associative recognition (e.g., Buchler, Faunce, Light, Gottfredson, & Reder, 2011; Light, Patterson, Chung, & Healy, 2004) and feature and conjunction errors (e.g., Jones, Jacoby, & Gellis, 2001; Reinitz, Verfaellie, & Milberg, 1996). In associative recognition experiments, subjects study pairs of words (e.g., "sailor - anchor") and are later given a recognition test on several different classes of items: (1) intact pairs of words in the same configuration as they were studied ("sailor - anchor" or "village - park"), (2) rearranged pairs ("sailor - park"), (3) new-old or old-new pairs ("anchor - grease"), and (4) new-new

pairs ("bacon - grease"). The subsequent recognition test can vary; sometimes, subjects respond only "old" or "new" based on whether a presented pair was an exact copy of a studied pair, but more recent techniques (e.g., as seen in Buchler et al., 2011) have subjects choose one of multiple responses ("old-old," "old-new," etc.).

Associative recognition procedures are often used in concert with repetition to dissociate the effects of familiarity and recollection on recognition memory. When a given word pair is repeated, both familiarity and recollection are improved for that pair. When a rearranged pair is presented at test, however, familiarity of the two words is high. Subjects must use controlled recollective processes to discriminate between studied pairs and lure pairs that contain one or two words that were actually studied during the experiment. Light and colleagues (2004) presented young and older adults with unrelated word pairs, half presented once and half presented four times. They found, similar to Benjamin (2001), that both groups exhibited an improved hit rate as a result of repetition. Older adults also showed an increased false alarm rate for rearranged pairs, however, whereas young adults showed either no change or a decrease in false alarm rate. The explanation provided by Light et al. was similar to that of Benjamin's: familiarity was increased for both groups of subjects as a result of repetition, but young adults were able to draw on recollective processes to avoid making false alarms to rearranged pairs. Older adults, who have been shown to have a decreased ability to use recollection, were less able to draw on this type of processing and thus made more false alarms to pairs that were familiar but not in their intact form.

Experiment 3 was designed to determine if repeated experience with categorized lists would change the relationship between output dominance and false alarm rates relative to Experiments 1 and 2. This third experiment used the basic method of Experiment 1 (120 studied words, 240 lures) but included three experimental conditions: study once and test, as in Experiment 1; study twice and then test, similar to Benjamin (2001), and study once, test with feedback, then study and test second time — a new condition featuring three presentations of the material. Note that all subjects received one final test, but the subjects in the third condition received a test twice. Both study and test phases were expected to increase familiarity for studied words, because tests were also predicted to increase familiarity for lures (because subjects saw them during the test). The beneficial effects of testing with feedback were also predicted to increase recollection for studied words, however. Studying twice relative to once was predicted to lead the young adults to reduce their false alarm rates (due to greater reliance on recollection and less on familiarity); including feedback after the first test was also predicted to reduce false alarm rates even further for the same reasons.

Participants

Sixty students from Washington University in St. Louis participated for either course credit or payment.

Design

This study used a 2 (counterbalancing group) x 3 (experimental condition) between-subjects design. One group of subjects was presented with the odd-numbered words from the first six lists and the even-numbered words from the last six lists. The

other group was presented with the alternate words — the even-numbered words from the first six lists and the odd-numbered words from the last six lists. After the study phase, subjects from both counterbalancing groups were then assigned randomly to one of three experimental conditions. Subjects in the ST condition (S representing study and T representing test) studied the 120 word list and took a recognition memory test (360 items) and made confidence ratings, as in Experiment 1. Subjects in the SST condition, however, studied and then immediately restudied the same materials, then took the recognition memory test and made confidence ratings. Subjects in the STST condition studied the 120 words, took a recognition memory test, made confidence ratings and, critically, received feedback (a screen displaying "correct" or "incorrect") after each rating. These subjects then restudied the materials and took a final test (without feedback) in the same manner as those in the ST and SST conditions. An illustration of these different experimental conditions can be found in Figure 7. Note that the final criterial test occurred immediately after the last list presentation across the three conditions, albeit after the only list presentation in the ST condition and after the second presentation for the SST and STST conditions.

Thus, ST was a replication of Experiment 1, SST applied the method used by Benjamin (2001), and STST added a test with feedback between the study repetitions used in SST. Note that although Benjamin presented each DRM list three times in a row, this experiment presented the 12 categories once in their entirety before repeating them (i.e., in two separate blocks). Spacing the material in this fashion was thought to provide subjects with greater temporal discrimination, making it easier to invoke more controlled

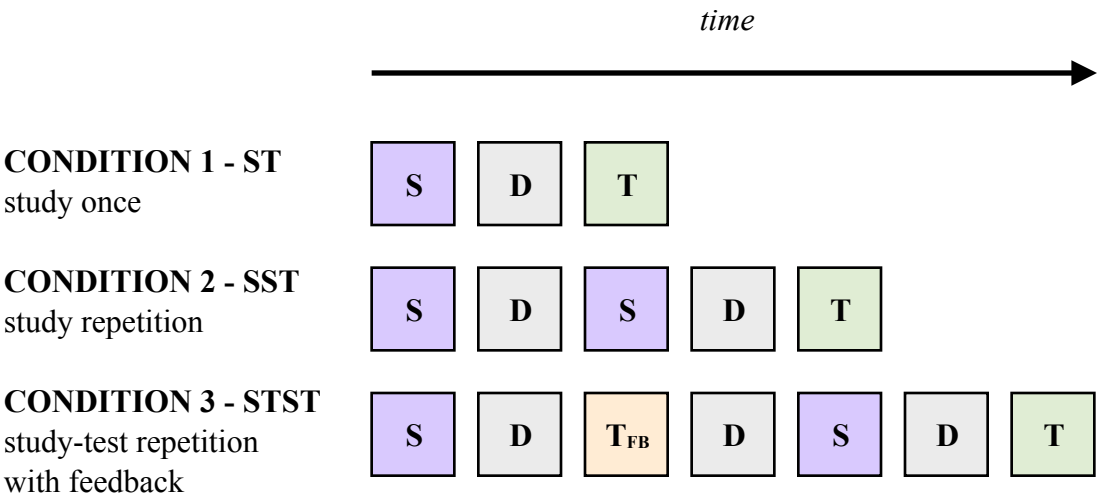


Figure 7. A pictorial representation of the three experimental conditions in Experiment 3.

S = study phase, D = distractor phase, T_{FB} = test with feedback phase, T = test phase.

processing on the final test. This alteration also allowed insertion of the intermediate test unique to the STST condition.

The entire experiment took 45 minutes (for subjects in the ST condition) to 90 minutes (for those in the STST condition).

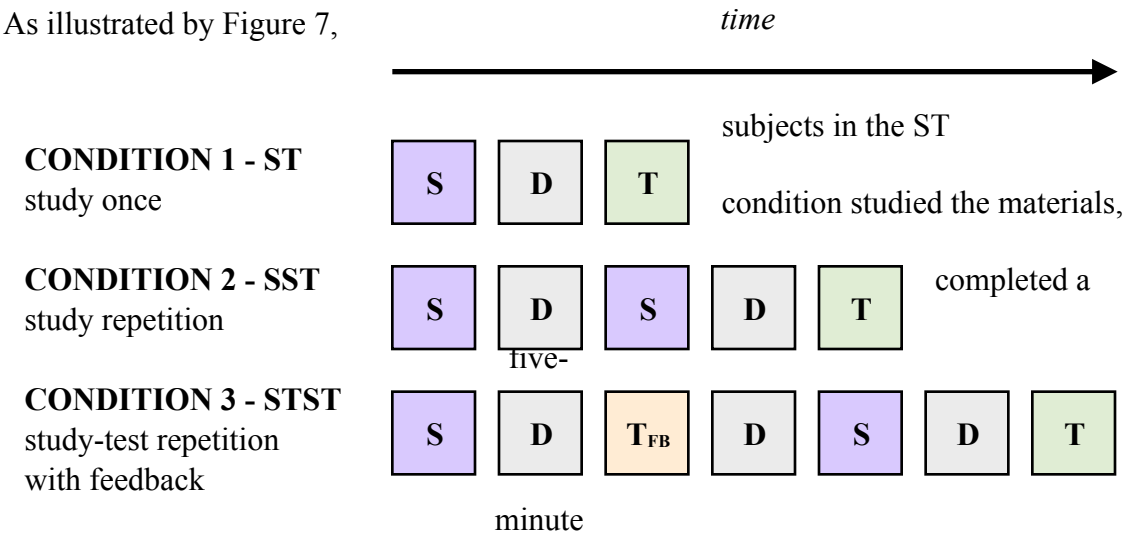
Procedure

During each study phase, subjects were presented with 120 items in the same manner as in the prior experiments. The category name was presented over headphones, followed by a four second pause; then, each category item was presented at a rate of one word per two seconds. This procedure was repeated for each of the 11 remaining lists of the stimulus set.

During each distractor phase, subjects were randomly asked to list as many United States presidents as possible, as many United States cities as possible, or as many of the 50 United States as possible. In the SST and STST conditions, where subjects completed more than one distractor task, no subject was assigned the same distractor task twice.

During each test phase, subjects made yes-no (i.e., old or new) recognition judgments on the 360 items, randomly presented, as in Experiment 1. Following each recognition judgment, subjects rated their confidence on a 0 (*not at all confident*) to 100 (*entirely confident*) sliding scale. The experiment then continued on to the next word. During the intermediate test in the STST condition only, an on-screen feedback message was displayed for 1.5 seconds after each confidence rating indicating whether the recognition judgment was correct or incorrect.

As illustrated by Figure 7, subjects in the ST condition studied the materials, completed a five-minute distractor task, then took the recognition test. Subjects in the SST condition studied all the materials once, completed a five-minute distractor task, restudied the materials again, in a different random order, completed a second distractor task, and then took the recognition test. Subjects in the STST condition studied, completed a five-minute distractor task, took an intermediate recognition test with feedback (as described above), completed a second distractor task, studied the materials again, completed a third distractor task, and then took the final recognition test.



Results

In our analysis of Experiment 3, we focused primarily on investigating performance for studied items and related lures across the three conditions. We do not emphasize responses to unrelated lures because performance for this item type was high; only rarely did subjects false alarm to these items. As the ST condition in Experiment 3 was an exact replication of Experiment 1, we first compared the results of both experiments to ensure there were no notable performance differences between groups of subjects. As expected, this was the case (FAR_{E3} = .37 vs. FAR_{E1} = .39, etc.)

Recognition of Studied Words

Hit rate and confidence ratings. Hit rate and confidence ratings can be found in Tables 13 and 14, respectively. A one-way ANOVA failed to find an effect of condition on the hit rate, $F(2, 57) = 1.47$, $p = .238$. Two more one-way ANOVAs failed to find an effect of condition on confidence for hits, $F(2, 57) = .13$, $p = .882$, or an effect of condition on confidence for misses, $F(2, 56) = 2.42$, $p = .098$. However, this last

Table 13

Response rates as a function of word type in Experiment 3.

	Item Type	Hit	Miss	Correct Rejection	False Alarm	d'
Condition 1 (ST)	Studied Words	.75 (.17)	.25 (.17)	--	--	--
	Related Lures	--	--	.63 (.22)	.37 (.22)	1.00
	Unrelated Lures	--	--	.91 (.07)	.09 (.07)	2.01
Condition 2 (SST)	Studied Words	.82 (.14)	.18 (.14)	--	--	--
	Related Lures	--	--	.70 (.16)	.30 (.16)	1.44
	Unrelated Lures	--	--	.94 (.07)	.06 (.07)	2.47
Condition 3 (STST)	Studied Words	.71 (.27)	.29 (.27)	--	--	--
	Related Lures	--	--	.86 (.12)	.14 (.12)	1.63
	Unrelated Lures	--	--	.99 (.02)	.01 (.02)	2.88

Note. Standard deviations presented in parentheses.

Table 14

Confidence as a function of word type in Experiment 3.

	Item Type	Hit	Miss	Correct Rejection	False Alarm
Condition 1 (ST)	Studied Words	84.5 (11.75)	55.3 (14.31)	--	--
	Related Lures	--	--	64.3 (14.35)	61.4 (13.16)
	Unrelated Lures	--	--	72.7 (16.66)	43.0 (18.98)
Condition 2 (SST)	Studied Words	86.2 (9.20)	45.3 (21.17)	--	--
	Related Lures	--	--	63.9 (18.14)	51.1 (14.02)
	Unrelated Lures	--	--	78.0 (18.36)	29.1 (22.36)
Condition 3 (STST)	Studied Words	85.3 (11.76)	44.3 (15.85)	--	--
	Related Lures	--	--	71.3 (14.29)	54.1 (15.05)
	Unrelated Lures	--	--	86.2 (15.26)	51.6 (30.95)

Note. Confidence judgments were made on a 0-100 scale. Standard deviations presented in parentheses.

relationship did appear to approach significance, however, and was likely driven by a marginally significant difference between the ST and STST conditions, $p = .053$.

Effects of output dominance on the hit rate. Multiple stepwise regression was used to assess the effects of output dominance on the hit rate while controlling for printed frequency as measured by log HAL. The regression results can be found in Table 15. In the ST condition, printed frequency did not significantly predict hit rate, but when output dominance was added to the regression equation, it predicted a significant amount of variance such that subjects were more likely to correctly respond "old" to high output dominance items. Neither printed frequency nor output dominance explained any variance in hit rate in either the SST or STST conditions, however.

Metacognitive measures.

Absolute accuracy for studied words: Confidence-accuracy correlations. Table 16 depicts the across-subjects calibration correlations for studied words. The relationship between confidence and accuracy for studied words was positive for subjects in the ST condition, $r(20) = .51$, $p = .022$, the SST condition, $r(20) = .66$, $p = .001$, and the STST condition, $r(20) = .84$, $p < .001$.

Relative accuracy for studied words: Gamma. Gamma correlations were good for studied words in both the ST ($M = .66$), SST ($M = .74$) and STST ($M = .75$) conditions. A one-way ANOVA failed to detect any differences among these three groups, $F(2, 56) = .64$, $p = .529$.

In sum, the ST, SST, and STST conditions did not show any differences in hit rate, hit confidence, or the relationship between output dominance and hit rate. However, the

Table 15

Summary of multiple regression analysis for effects of output dominance on hit rate in Experiment 3.

Variable	*B*	*SE B*	β
Condition 1 (ST)			
Printed frequency	-.01	.00	-.11
Output dominance	.00	.00	-.16*
Condition 2 (SST)			
Printed frequency	.00	.00	.02
Output dominance	.00	.00	.04
Condition 3 (STST)			
Printed frequency	.00	.00	.00
Output dominance	.00	.00	.01

Note. First step (printed frequency only) not shown. In the ST condition, $R^2 = .00$ for Step 1 ($p > .05$), $\Delta R^2 = .03$ for Step 2 ($p < .05$). All other effects are nonsignificant ($p > .05$). *$p < .05$.

Table 16

Across-subjects confidence-accuracy correlations as a function of word type in Experiment 3.

	Condition 1 (ST)	Condition 2 (SST)	Condition 3 (STST)
Studied Items	.51	.66	.84
Related Lures	.20	.75	.43

link between confidence and accuracy across subjects strengthened with additional study or test periods. Resolution (i.e., gamma) was similar among the three conditions.

Recognition of Related Lures

False alarm rates and confidence ratings. False alarm rates for related lures and confidence ratings can be found in Tables 13 and 14, respectively. A one-way ANOVA detected an effect of condition on false alarm rate, $F(2, 57) = 9.73$, $p < .001$; Fisher LSD tests confirmed that false alarms were lower in the STST condition than the other two conditions.

Turning to confidence ratings, a one-way ANOVA examining the effect of condition on confidence for false alarms only approached significance, $F(2, 57)$, $p = .078$, suggesting the possibility that subjects were more confident in false alarms in the ST condition relative to the SST and STST conditions. A second one-way ANOVA to analyze the effect of condition on confidence for correct rejections failed to detect any differences, $F(2, 57) = 1.42$, $p = .249$.

Effects of output dominance on the false alarm rate. Multiple regression was used to assess the effects of output dominance on false alarm rate for related lures while controlling for printed frequency (as measured by log HAL). The results can be found in Table 17. In the ST and SST conditions, printed frequency neither significantly predicted false alarm rate nor explained a statistically significant proportion of variance in false alarm rate. When output dominance was added to the regression equations in the ST and SST conditions, though, it explained a significant amount of variance. In these conditions, subjects were more likely to incorrectly respond "old" to higher output

Table 17

Summary of multiple regression analysis for effects of output dominance on false alarm rate in Experiment 3.

Variable	*B*	*SE B*	β
Condition 1 (ST)			
Printed frequency	.00	.00	.07
Output dominance	-.01	.00	-.32*
Condition 2 (SST)			
Printed frequency	.00	.00	-.01
Output dominance	-.01	.00	-.40*
Condition 3 (STST)			
Printed frequency	.01	.00	.17*
Output dominance	.00	.00	-.15*

Note. First step (printed frequency only) not shown. In the ST condition, $R^2 = .03$ for Step 1 ($p < .05$), $\Delta R^2 = .10$ for Step 2 ($p < .05$). In the SST condition, $R^2 = .01$ for Step 1 ($p > .05$), $\Delta R^2 = .14$ for Step 2 ($p < .05$). In the STST condition, $R^2 = .05$ for Step 1 ($p < .05$), $\Delta R^2 = .06$ for Step 2 ($p < .05$). All other effects are nonsignificant ($p > .05$). *$p < .05$.

dominance items than lower output dominance items. In contrast, in the STST condition, both printed frequency and output dominance played a role in explaining false alarm rate; subjects were more likely to false alarm to items of higher frequency or higher output dominance than items of lower frequency or lower output dominance.

Metacognitive measures.

Absolute accuracy for related lures: Confidence-accuracy correlations. Table 16 depicts the across-subjects calibration plot for related lures. The relationship between confidence and accuracy for related lures was not significant in the ST condition, $r(20) = .20$, $p = .411$, but was significant in the SST condition, $r(20) = .75$, $p < .001$. The correlation was marginally significant in the STST condition, $r(20) = .43$, $p = .06$.

Absolute accuracy for related and unrelated lures: Accuracy plots. Figures 8 and 9 depict the accuracy plots for related and unrelated lures, respectively. These curves show that when recognition judgments (whether "old" or "new") were made with higher confidence, they were also more likely to be accurate. Differences emerge as a result of experimental condition, however. Notably, subjects in the ST condition made the most incorrect high confidence recognition decisions. Although accuracy corresponded with confidence most strongly in the STST condition, all three groups were less accurate than predicted by their confidence ratings when responding "old."

Relative accuracy for related lures: Gamma. Gamma correlations for related lures appeared to improve from the ST ($M = .07$) to the SST ($M = .24$) to the STST ($M = .46$) conditions. A one-way ANOVA detected significant statistical differences among these

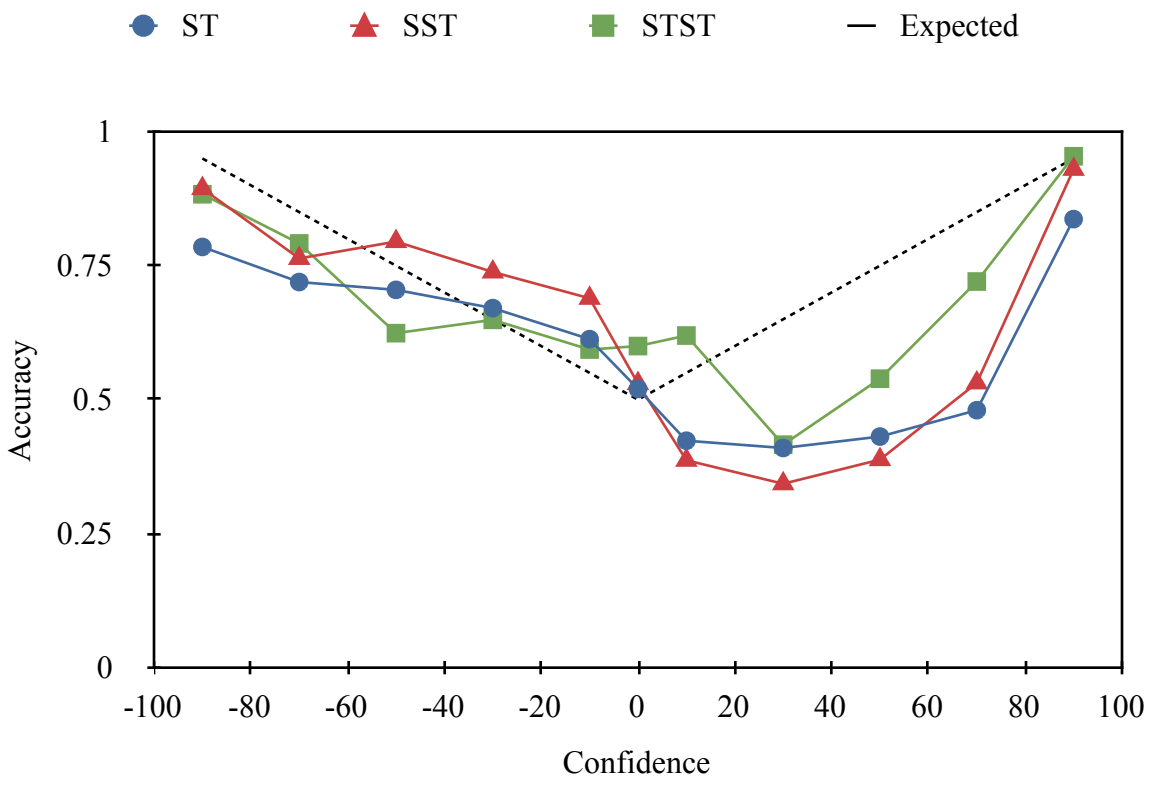


Figure 8. The accuracy plot for Experiment 3 depicting subjects' ability to discriminate between studied words and related lures. The dotted line represents expected accuracy; points above this line represent underconfidence, whereas points below this line represent overconfidence.

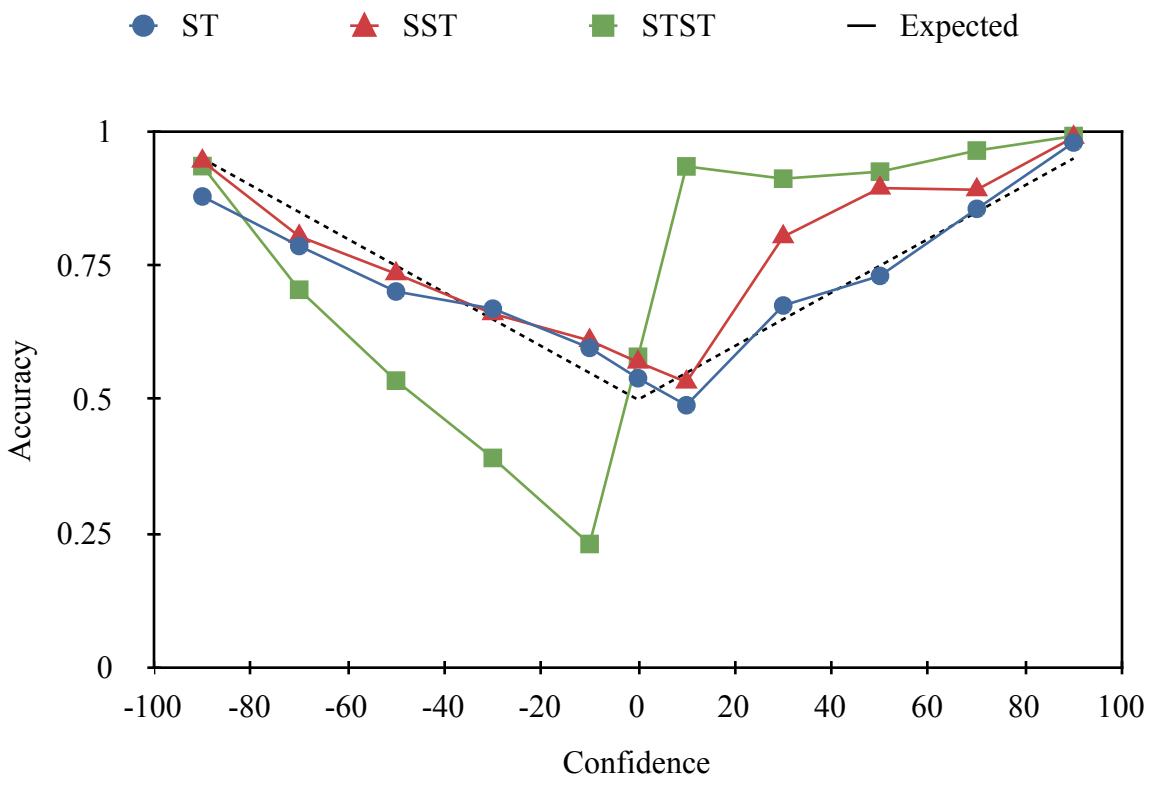


Figure 9. The accuracy plot for Experiment 3 depicting subjects' ability to discriminate between studied words and unrelated lures. The dotted line represents expected accuracy; points above this line represent underconfidence, whereas points below this line represent overconfidence.

three groups, $F(2, 57) = 7.28$, $p = .002$. Fisher LSD post hoc comparisons indicated that all comparisons were significant.

In sum, subjects in the STST condition incorrectly responded "old" less than subjects in the ST and SST conditions to related lures, probably because of feedback given after the first test. Output dominance was a predictor of false alarm rate in all three conditions, however. An extra study or an extra study and test strengthened the relationship between confidence and accuracy such that subjects in the SST and STST conditions who were more accurate were also more confident. Subjects in the STST condition also showed fewest false alarms and demonstrated the best memory resolution.

Effects of Output Dominance on Corrected Recognition

The effects of output dominance on corrected recognition (hits minus false alarms) can be found in Table 18. In the ST and SST conditions, printed frequency neither predicted corrected recognition nor explained a statistically significant proportion of variance in corrected recognition. When output dominance was added to the regression equation, however, it explained a significant proportion of variance in both conditions such that items lower in output dominance were more likely to be correctly recognized. In the STST condition, though, neither printed frequency nor output dominance had any effect on corrected recognition.

Discussion

Experiment 3 resulted in several key findings. First, even when subjects were provided with repeated exposure to categorized lists at study and at initial test, false alarms were still more likely for category members of higher output dominance. As

Table 18

Summary of multiple regression analysis for effects of output dominance on corrected recognition in Experiment 3.

Variable	*B*	*SE B*	β
Condition 1 (ST)			
Printed frequency	-.01	.01	-.12
Output dominance	.00	.00	.14*
Condition 2 (SST)			
Printed frequency	.00	.01	.02
Output dominance	.01	.00	.36*
Condition 3 (STST)			
Printed frequency	-.01	.00	-.12
Output dominance	.00	.00	.11

Note. First step (printed frequency only) not shown. In the ST condition, $R^2 = .03$ for Step 1 ($p < .05$), $\Delta R^2 = .02$ for Step 2 ($p < .05$). In the SST condition, $R^2 = .01$ for Step 1 ($p > .05$), $\Delta R^2 = .12$ for Step 2 ($p < .05$). All other effects are nonsignificant ($p > .05$). *$p < .05$.

reported in Experiments 1 and 2, these patterns were not found for hits (except in the Experiment 3 ST condition) — a failure to find the mirror effect of recognition memory in these data. Experiment 3 also showed that providing subjects with an additional study and test opportunity (as in the STST condition) statistically significantly lowered the false alarm rate for related lures and improved metacognition for these difficult items.

Thus, the analysis of the response rates and mean confidence ratings in Experiment 3 suggests that the study and test repetition in the STST condition led to markedly improved performance in that condition relative to the ST condition, but that only repeating the study phase (as in the SST condition) had no effect. This is an interesting pattern of data and may provide a tie-in to literature on the testing effect (see Roediger, Putnam, & Smith, 2011, for a review, but note that most testing research is done with recall rather than recognition); repeated study (SST) has little effect on subsequent performance, but interpolating a test with feedback (STST) substantially reduces high confidence false alarms, improving recognition memory. Importantly, however, neither of the experimental manipulations reduced the occurrence of false memories to high output dominance members, suggesting that the processes causing these types of memory errors in the ST condition persisted with repeated study (SST) and repeated study and test (STST). On the other hand, overall metacognition, as shown primarily by accuracy plots and gamma correlations, was poorest in the ST condition but best in the STST condition.

General Discussion

Table 19 presents a summary of the results of Experiments 1, 2, and 3. Across these three experiments, subjects studied sets of words belonging to different semantic categories. Experiment 1 demonstrated that after studying these materials, subjects were highly likely to falsely recognize unstudied category members from studied categories. More specifically, the likelihood of false recognition varied linearly with output dominance such that category members that were more frequent in the norms (i.e., high output dominance category members) were more likely to be falsely recognized than category members that were less frequent. Output dominance was shown to affect false recognition above and beyond lexical frequency. These patterns were not observed for hits, however, a result inconsistent with the mirror effect in recognition memory (as well as the word frequency effect more generally).

The findings in Experiment 1 are consistent with the work of Dewhurst (2001), who manipulated output dominance in a similar yet less gradated fashion and found that subjects were more likely to false alarm to high output dominance members than low output dominance members. Dewhurst theorized that this effect was due to errors of both spreading activation and source monitoring that occur during learning (encoding) and that the reason subjects were more likely to falsely recognize high output dominance members was because they were more likely to implicitly generate these members during encoding relative to lower dominance members.Experiment 2 was designed to assess another possible explanation of the negative relation between output dominance and false alarm rate: that the composition of the test list (with half related and half unrelated lures)

Table 19

Summary of Experiments 1, 2, and 3.

Variable	E1	E2	E3 (SST)	E3 (STST)
Memory for Studied Words				
Hit Rate	.73	.70	.82	.71
Hit Confidence	82.6	84.2	86.2	85.3
Regression β	-.10	-.18*	.04	.00
Confidence-Accuracy Correlation	.68	.71	.66	.84
Gamma Correlation	.73	.79	.74	.75
Memory for Related Lures				
False Alarm Rate	.39	.29	.30	.14
False Alarm Confidence	62.7	55.1	51.1	54.1
Regression β	-.45*	-.45*	-.40*	-.15*
Confidence-Accuracy Correlation	.23	.28	.75	.43
Gamma Correlation	-.16	-.13	.24	.46
Corrected Recognition β	.37*	.25*	.36*	.11

Note. E1 = Experiment 1; note that this was identical to the ST condition in Experiment 3. E2 = Experiment 2. E3 = Experiment 3. Regression β = beta weight for relationship between output dominance and response rate, *$p < .05$. Corrected recognition means hits minus false alarms.

directed subjects toward using output dominance and category membership as diagnostic cues at test, which led to a greater degree of error-prone processing. To investigate this alternative explanation, Experiment 2 omitted the distractors that were unrelated to the studied categories that were included in the final recognition test in Experiment 1. This modification had a mixed effect: it reduced the overall false alarm rate for lures as compared to Experiment 1, but failed to affect the relationship between output dominance and false alarm rate. This observation was also consistent with the work of Dewhurst et al. (2009), who investigated whether categorized list errors arise as a result of processes that occur during encoding versus processes that occur during retrieval. The observation in Experiment 2 supports the proposition that the associations made during encoding fuel memory errors, as manipulating the structure of the test list did little to affect high confidence false recognition for high output dominance category members.

Experiment 3 was designed to evaluate methods for improving subjects' metacognitive monitoring in the categorized list procedure. This experiment featured three different conditions: one an exact replication of Experiment 1 (ST), a second which presented the study material to subjects twice before final test (SST), and a third in which subjects studied the material, took a test in which they were given feedback, studied the material a second time, and then took a final test (STST). Experiment 3 uncovered several novel patterns of data. First, providing subjects with an additional repetition of the study materials did little to reduce false alarm rate; only an interpolated test with feedback, when combined with a study repetition, reduced false alarms and improved metacognition. Even though false alarm rate was decreased in this condition, however,

subjects were still more likely to false alarm to higher output dominance category members than lower output dominance members.

These experiments can be integrated into prior research by offering a possible explanation of how recognition memory operates for categorized list materials. When subjects study categorized list words, they implicitly generate associates of those words. More typical or frequent members of the category are more likely to be generated than less frequent category members. Meanwhile, nonstudied category members receive small amounts of categorical spreading activation as a result of the small degrees of overlap between the studied words and representations of these other words in memory. At test, subjects commit false alarms as a result of these two different processes. Previous research suggests that recollection-based errors are produced by source memory errors for items implicitly generated at test and familiarity-based errors are produced by summation of small degrees of match between memory representation and test word when this summation passes a certain threshold (Dewhurst, 2001). Because implicit generation of category members occurs at encoding and not at retrieval, as suggested by Dewhurst et al. (2009), manipulating the test list (as in Experiment 2) did not result in a reduction of false alarms for high dominance category members found in Experiment 1.

Providing subjects with an additional opportunity to study the categorized list words enhanced the memory traces for the studied items. On the other hand, repeated study did not reduce — and may have even increased — the likelihood that nonstudied category members were generated at encoding. As a result, although memory performance improves overall when a second study opportunity is provided, a high

number of memory errors (false alarms) still occur for high output dominance category members. Providing subjects with an intermediate test and then a second study opportunity mitigates these effects by providing subjects with an opportunity to edit the contents of memory for persistent errors; namely, eliminating any source monitoring errors that arose between first study and intermediate test.

These results can also be partially explained by models of episodic memory such as the subjective likelihood model (SLiM; McClelland & Chappell, 1998) or the retrieving effectively from memory model (REM; Shiffrin & Steyvers, 1997; although see Criss & McClelland, 2006, for a distinction). These models address the important effects of *differentiation* — that is, repetition or increased study time for study material — on recognition memory. According to these models, subjects respond "old" when there is a large degree of match between an encoded memory representation and word at test (similar to the Arndt & Hirshman [1998] theory discussed earlier). Every time a word is restudied, it is added to the preexisting memory trace, which becomes richer and more feature-filled as a result. This memory trace enrichment makes it more likely that subjects will respond "old" to words that were actually studied, since these representations feature the highest degree of match with a given studied word at test. As memory traces become enriched through restudy or repetition, the likelihood that they will match to a lure word is also decreased.

Differentiation, as explained by SLiM and REM, is used in this context to explain the mirror effect, which is traditionally a troubling finding for recognition memory theorists. These theories dovetail nicely, however, with the more associative-based

account of Dewhurst (2001) and colleagues to explain the Experiment 3 findings. When categorized lists are restudied, the individual traces for each word become enriched, leading to improved differentiation at final test. However, increasing the number of times a categorized list is presented increases the possibility that nonstudied category members may be implicitly generated during encoding. This means that despite the improved differentiation provided by repeated list presentation, the possibility of source memory errors still exists. When subjects make source memory errors on a test with feedback, they have an opportunity to edit memory for these items, possibly invoking a greater degree of controlled processing to do so. Subjects may continue to generate category associates during an intermediate test or during the second presentation of the material (in the STST condition), however, leading to the lower rate of false alarms to high output dominance members demonstrated in this condition.

This set of experiments also highlights the ways categorized lists can be used to study false memories. As shown with these materials, critical items need not be the most common, or even the five most common, members of the category to evoke reliable false alarms — rather, items that are studied and tested from categorized lists can be taken from any subset of the ordered, vertically-associated categories. These experiments demonstrated one powerful way this can be done: by presenting alternate items and testing across all of them to obtain hit, miss, false alarm, and correct rejection rates for each item as well as each output dominance position. Calculating response rates and confidence ratings for different output dominance positions allows investigation of the

relationships between output dominance and the variable of interest with correlation coefficients or regression.

This paper used this more nuanced methodology to examine the linear relationships between output dominance and response rates. The findings reported here further establish and refine the account provided by Dewhurst et al. (2009) and suggest two explanations as to why errors may arise in the categorized list procedure. This account provides information about the processing that takes place during the encoding and retrieval of items that are organized categorically. In particular, studying some items from the same semantic category appears to affect memory for related category members. These effects seem to be related to the output dominance or typicality of a given category member and the ease with which that category member comes to mind, and these findings differentiate processing that occurs for categorized list material from associative lists (e.g., DRM lists) and even for frequent versus infrequent words in the lexicon.

Conclusion

In summary, the categorized list procedure presents both theoretical and methodological advantages over other popular procedures for investigating false memories. Theoretically, categorized lists provide a closer look at the processes that come into play in category learning, prototypicality, and instance generation; methodologically, categorized lists make it easier to vary lure "strength" and have more than one critical item per list (as compared to the DRM procedure). The flexibility of the categorized list procedure makes it an ideal vehicle for the further investigation of many

of the ideas touched on in this thesis: effects of word frequency, implicit generation effects during encoding, differentiation, and the testing effect.

The three experiments reported here used the categorized list procedure to show high levels of false recognition for high output dominance category members in a pattern above and beyond the effects of word frequency. These effects were partly attributed to source monitoring errors that arose as a result of implicit generation of nonstudied associates to studied lists at encoding. These findings were supported by the observation that manipulating the test list (the retrieval phase) did little to eliminate the relatively high number of false alarms for high output dominance category members; this pattern was so robust that it could not be attenuated by providing subjects with additional study and test periods. The persistence of these false alarms to high output dominance category members, and the processes that cause them, will be a necessary topic for future research. The initial findings reported here suggest that additional study and test repetitions provide a useful means for modulating recognition memory errors.

References

Arndt, J., & Hirshman, H. (1998). True and false recognition in MINERVA2: Explanations from a global matching perspective. *Journal of Memory and Language*, *39*, 371-391.

Balota, D. A., & Neely, J. H. (1980). Test-expectancy and word-frequency effects in recall and recognition. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 576-587.

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

Barsalou, L. W. (1987). The instability of graded structure: implications for the nature of concepts. In U. Neisser (Ed.), *Concepts and conceptual development: ecological and intellectual factors in categorization*. Cambridge, UK: Cambridge University Press.

Battig, W. F., & Montague, W. E. (1969). Category norms for verbal items in 56 categories: A replication and extension of the Connecticut norms. *Journal of Experimental Psychology*, *80*, 1-46.

Benjamin, A. S. (2001). On the dual effects of repetition on false recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 941-947.

Brainerd, C. J., & Reyna, V. F. (2005). *The science of false memory*. New York: Oxford University Press.

Brainerd, C. J., Reyna, V. F., & Zember, E. (2011). Theoretical and forensic implications of developmental studies of the DRM illusion. *Memory & Cognition*, *39*, 365-380.

Brewer, W. F., & Sampaio, C. (2006). Processes leading to confidence and accuracy in sentence recognition: A metamemory approach. *Memory*, *14*, 540-552.

Brewer, W. F., Sampaio, C., & Barlow, M. R. (2005). Confidence and accuracy in the recall of deceptive and nondeceptive sentences. *Journal of Memory and Language*, *52*, 618-627.

Buchler, N. G., Faunce, P., Light, L. L., Gottfredson, N., & Reder, L. M. (2011). Effects of repetition on associative recognition in young and older adults: Item and associative strengthening. *Psychology and Aging*, *26*, 111-126.

Clark, S. E., & Gronlund, S. D. (1996). Global matching models of recognition memory: How the models match the data. *Psychonomic Bulletin & Review*, *3*, 37-60.

Criss, A., McClelland, J. L. (2006). Differentiating the differentiation models: A comparison of the retrieving effectively from memory model (REM) and the subjective likelihood model (SLiM). *Journal of Memory and Language*, *55*, 447-460.

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology*, *58*, 17-22.

Dewhurst, S. A. (2001). Category repetition and false recognition: Effects of instance frequency and category size. *Journal of Memory & Language*, *44*, 153-167.

Dewhurst, S. A., & Anderson, S. J. (1999). Effects of exact and category repetition in true and false recognition memory. *Memory & Cognition*, *27*, 665-673.

Dewhurst, S. A., Bould, E., Knott, L. M., & Thorley, C. (2009). The roles of encoding and retrieval processes in associative and categorical memory illusions. *Journal of Memory & Language*, *60*, 154-164.

Dewhurst, S. A., & Farrand, P. (2004). Investigating the phenomenological characteristics of false recognition for categorised words. *European Journal of Cognitive Psychology*, *16*, 403-416.

Dry, M. J., & Storms, G. (2010). Features of graded category structure: Generalizing the family resemblance and polymorphous concept models. *Acta Psychologica*, *133*, 244-255.

Dunlosky, J., & Metcalfe, J. (2008). Law and eyewitness accuracy. In *Metacognition* (pp. 171-234). Thousand Oaks, California: SAGE Publications.

Gallo, D. A. (2006). *Associative illusions of memory: False memory research in DRM and related tasks*. New York: Psychology Press.

Gallo, D. A. (2010). False memories and fantastic beliefs: 15 years of the DRM illusion. *Memory & Cognition*, *37*, 831-846.

Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory & Cognition*, *13*, 8-20.

Greene, R. L. (2007). Foxes, hedgehogs, and mirror effects: The role of general principles in memory research. In J. Nairne (Ed.), *The foundations of remembering: Essays in honor of Henry L. Roediger III*. London: Psychology Press.

Huron, C., Servais, C., & Danion, J. M. (2001). Lorazepam and diazepam impair true, but not false, recognition in healthy volunteers. *Psychopharmacology*, *155*, 204-209.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory & Language*, *30*, 513-541.

Jacoby, L. L., Craik, F. I. M., & Begg, I. (1979). Effects of decision difficulty on recognition and recall. *Journal of Verbal Learning and Verbal Behavior*, *18*, 585-600.

Jacoby, L. L., & Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General*, *110*, 306-340.

Jacoby, L. L., Woloshyn, V., & Kelley, C. (1989). Becoming famous without being recognized: Unconscious influences of memory produced by dividing attention. *Journal of Experimental Psychology: General*, *118*, 115-125.

Johnson, M. K. (2006). Memory and reality. *American Psychologist*, *61*, 760-771.

Jones, T. C., Jacoby, L. L., & Gellis, L. A. (2001). Cross-modal feature and conjunction errors in recognition memory. *Journal of Memory and Language*, *44*, 131-152.

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

Light, L. L., Patterson, M. M., Chung, C., & Healy, M. R. (2004). Effects of repetition and response deadline on associative recognition in young and older adults. *Memory & Cognition*, *32*, 1182-1193.

Lindsay, D.S. (2008). Source monitoring. In Roediger, H. L. (Ed.). *Cognitive psychology of memory. Vol. 2 of Learning and memory: A comprehensive reference* (J. Byrne, Ed.). Oxford: Elsevier.

Loftus, E. F. (1975). Eyewitness testimony: The influence of the wording of a question. *Bulletin of the Psychonomic Society*, *81*, 86-88.

Loftus, E. F., Miller, D. G., & Burns, H. J. (1978). Semantic integration of verbal information into a visual memory. *Journal of Experimental Psychology: Human Learning and Memory*, *4*, 19-31.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods*, *28*, 203-208.

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide*. London: Psychology Press.

Marsh, E. J., Eslick, A. N., & Fazio, L. K. (2008). False memories. In Roediger, H. L. (Ed.). *Cognitive psychology of memory. Vol. 2 of Learning and memory: A comprehensive reference* (J. Byrne, Ed.). Oxford: Elsevier.

Marx, M. H., & Henderson, B. B. (1996). A fuzzy trace analysis of categorical inferences and instantial associations as a function of retention interval. *Cognitive Development*, *11*, 551-569.

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review*, *105*, 724-760.

McDermott, K. B. (2006). Paradoxical effects of testing: Repeated retrieval attempts enhance the likelihood of later accurate and false recall. *Memory & Cognition*, *34*, 1273-1284.

Meade, M. L., & Roediger, H. L. (2006). The effect of forced recall on illusory recollection in younger and older adults. *The American Journal of Psychology*, *119*, 433-462.

Meade, M. L., & Roediger, H. L. (2009). Age differences in collaborative memory: The role of retrieval manipulations. *Memory & Cognition*, *37*, 962-975.

Mickes, L., Hwe, V., Wais, P. E., & Wixted, J. T. (2011). Strong memories are hard to scale. *Journal of Experimental Psychology: General*, *140*, 239-257.

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

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychology Bulletin*, *95*, 109-133.

Norman, K. A., Schacter, D. L. (1997). False recognition in younger and older adults: Exploring the characteristics of illusory memories. *Memory & Cognition*, *25*, 838-848.

Nunes, L. D., & DeSoto, K. A. (2010). [Association norms for words not in Nelson, McEvoy, & Schreiber (1998)]. Unpublished raw data.

Park, L., Shobe, K. K., & Kihlstrom, J. F. (2005). Associative and categorical relations in the associative memory illusion. *Psychological Science*, *16*, 792-797.

Pusen, C., Erickson, J. R., Hue, C., & Vyas, A. P. (1988). Priming from category members on retrieval of other category members: Positive and negative effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 627-640.

Rajaram, S. (1996). Perceptual effects on remembering: Recollective processes in picture recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 365-377.

Read, J. D. (1996). From a passing thought to a false memory in 2 minutes: Confusing real and illusory events. *Psychonomic Bulletin & Review*, *3*, 105-111.

Reinitz, M. T., Verfaellie, M., & Milberg, W. P. (1996). Memory conjunction errors in normal and amnesic subjects. *Journal of Memory and Language*, *2*, 286-299.

Reyna, V. F., & Brainerd, C. J. (1995). Fuzzy-trace theory: An interim synthesis. *Learning and Individual Differences*, *7*, 1-75.

Roediger, H. L. (1996). Memory illusions. *Journal of Memory and Language*, *35*, 76-100.

Roediger, H. L., & DeSoto, K. A. (in preparation). Complexities in the relation between confidence and accuracy in recognition memory.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 803-814.

Roediger, H. L., Putnam, A. L., & Smith, M. A. (2011). Ten benefits of testing and their applications to educational practice. In J. Mestre & B. Ross (Eds.), *Psychology of learning and motivation: Cognition in education*. (pp. 1-36). Oxford: Elsevier.

Roediger, H. L., Watson, J. M., McDermott, K. B., & Gallo, D. A. (2001). Factors that determine false recall: A multiple regression analysis. *Psychonomic Bulletin & Review*, *2001*, 385-407.

Roediger, H. L., Wixted, J. H., & DeSoto, K. A. (in press). The curious complexity between confidence and accuracy in reports from memory. In L. Nadel & W. Sinnott-Armstrong (Eds.), *Memory and law*. New York: Oxford University Press.

Rosch, E. H. (1973). Natural categories. *Cognitive Psychology*, *4*, 328-350.

Rosch, E. H. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, 104*,* 192-233.

Rosch, E. H., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7*,* 573-605.

Rotello, C. M., Macmillan, N. A., & Van Tassel, G. (2000). Recall-to-reject in recognition: Evidence from ROC curves. *Journal of Memory and Language*, *43*, 67-88.

Sampaio, C., & Brewer, W. F. (2009). The role of unconscious memory errors in judgments of confidence for sentence recognition. *Memory & Cognition*, *37*, 158-163.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM - retrieving effectively from memory. *Psychonomic Bulletin & Review*, *4*, 145-166.

Smith, S. M., Gerkens, D. R., Pierce, B. H., & Choi, H. (2002). The roles of associative responses at study and semantically guided recollection at test in false memory: The Kirkpatrick and Deese hypotheses. *Journal of Memory and Language*, *47*, 436-447.

Smith, S. M., Tindell, D. R., Pierce, B. H., Gilliland, T. R., & Gerkens, D. R. (2001). The use of source memory to identify one's own episodic confusion errors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 362-374.

Smith, S. M., Ward, T. B., Tindell, D. R., Sifonis, C. M., & Wilkenfeld, M. J. (2000). Category structure and created memories. *Memory & Cognition*, *28*, 386-395.

Stadler, M. A., Roediger, H. L., & McDermott, K. B. (1999). Norms for word lists that create false memories. *Memory & Cognition*, *27*, 494-500.

Thompson-Schill, S. L., Kurtz, K. J., & Gabrieli, J. D. E. (1998). Effects of semantic and associative relatedness on automatic priming.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychologist*, *26*, 1-12.

Weinstein, Y. (in press). *Flash programming for the social & behavioral sciences: A simple guide to sophisticated online surveys and experiments.* Thousand Oaks, CA: SAGE.

Wixted. J. T. (2011). Strong memories are not scalable. Talk given at the 51st Annual Meeting of the Psychonomic Society, November, 2011.

Van Overschelde, J. P., Rawson, K. A., & Dunlosky, J. (2004). Category norms: An updated an expanded version of the Battig and Montague (1969) norms. *Journal of Memory and Language*, *50*, 289-335.

Appendix

Categorized lists:

a bird

1. eagle
2. robin
3. bluejay
4. cardinal
5. hawk
6. bluebird
7. crow
8. hummingbird
9. parrot
10. sparrow
11. pigeon
12. seagull
13. dove
14. parakeet
15. falcon
16. canary
17. owl
18. ostrich
19. penguin
20. raven

a fish

1. salmon
2. trout
3. goldfish
4. catfish
5. tuna
6. shark
7. flounder
8. swordfish
9. herring
10. carp
11. cod
12. angelfish
13. dolphin
14. blowfish
15. guppy
16. halibut
17. marlin
18. minnow
19. piranha
20. snapper

an insect

1. fly
2. ant
3. spider
4. bee
5. mosquito
6. beetle
7. ladybug
8. grasshopper
9. butterfly
10. wasp
11. roach
12. moth
13. gnat
14. caterpillar
15. centipede
16. cricket
17. worm
18. mantis
19. dragonfly
20. flea

a vegetable

1. carrot
2. lettuce
3. broccoli
4. cucumber
5. pea
6. corn
7. potato
8. celery
9. onion
10. spinach
11. squash
12. bean
13. cauliflower
14. cabbage
15. radish
16. asparagus
17. pepper
18. beet
19. turnip
20. zucchini

a musical instrument

1. drum
2. guitar
3. flute
4. piano
5. trumpet
6. clarinet
7. violin
8. saxophone
9. trombone
10. tuba
11. cello
12. oboe
13. viola
14. harp
15. keyboard
16. piccolo
17. banjo
18. harmonica
19. cymbal
20. tambourine

an article of clothing

1. shirt
2. ants
3. sock
4. underwear
5. shoe
6. hat
7. shorts
8. jacket
9. sweater
10. skirt
11. jeans
12. coat
13. dress
14. glove
15. sweatshirt
16. scarf
17. blouse
18. tie
19. belt
20. undershirt

a weather phenomenon

1. tornado
2. hurricane
3. rain
4. snow
5. hail
6. flood
7. lightning
8. blizzard
9. earthquake
10. sleet
11. monsoon
12. thunder
13. tsunami
14. wind
15. storm
16. typhoon
17. drought
18. cloud
19. sunshine
20. drizzle

a sport

1. football
2. basketball
3. soccer
4. baseball
5. tennis
6. hockey
7. swimming
8. golf
9. volleyball
10. lacrosse
11. track
12. rugby
13. softball
14. skiing
15. cheerleading
16. running
17. gymnastics
18. polo
19. raquetball
20. wrestling

an occupation or

profession

1. doctor
2. teacher
3. lawyer
4. nurse
5. professor
6. accountant
7. psychologist
8. dentist
9. engineer
10. secretary
11. manager
12. cook
13. firefighter
14. policeman
15. athlete
16. banker
17. carpenter
18. janitor
19. scientist
20. student

a fruit

1. apple
2. orange
3. banana
4. grape
5. pear
6. peach
7. strawberry
8. kiwi
9. pineapple
10. watermelon
11. tomato
12. plum
13. grapefruit
14. mango
15. cherry
16. lemon
17. blueberry
18. cantaloupe
19. raspberry
20. lime

a part of the human

body

1. leg
2. arm
3. finger
4. head
5. toe
6. eye
7. hand
8. nose
9. ear
10. foot

11. mouth
12. stomach
13. heart
14. knee
15. neck
16. brain
17. hair
18. elbow
19. shoulder
20. chest

a four legged animal

1. dog
2. cat
3. horse
4. lion
5. bear
6. tiger
7. cow
8. elephant
9. deer
10. mouse
11. pig
12. rat
13. giraffe
14. squirrel
15. rabbit
16. goat
17. zebra
18. moose
19. sheep
20. cheetah

Items are listed in order of highest to lowest output dominance. Specifically, the item in Position 1 has the highest output dominance and the item in Position 20 has the lowest.

Unrelated items:

- adjective
- aluminum
- amethyst
- anaconda
- aspen
- axe
- barge
- battleship
- bazooka
- blender
- brass
- butter
- cabin
- cabinet
- cave
- cedar
- chapel
- cobra
- coffee
- conjunction
- cousin
- curry
- daughter
- day
- decade
- denim
- diamond
- dogwood
- essay
- father
- ferry
- fleece
- flyer
- futon
- garnet
- gin
- governor
- grass
- grenade
- igloo
- iris
- island
- kayak
- ketchup
- kilometer
- ladle
- lawnmower
- lead
- letter
- level
- lilac
- liquor
- magazine
- mansion
- mayor
- micrometer
- mile
- milk
- millimeter
- minute
- monastery
- nail
- nanosecond

- nickel
- niece
- noun
- nylon
- oil
- opal
- ottoman
- palm
- pamphlet
- petunia
- pick
- pitchfork
- plow
- preposition
- president
- pronoun
- python
- raft
- rattlesnake
- recliner
- rifle
- river
- rock
- rose
- rum
- sanctuary
- sander
- sapphire
- screwdriver
- senator
- shovel
- soda
- sofa
- son
- spruce
- stove
- sugar
- sword
- synagogue
- temple
- tent
- tongs
- townhouse
- treasurer
- velvet
- vinegar
- violet
- viper
- vodka
- week
- whisk
- whiskey
- wine
- wool
- wrench
- yard
- zinc