Vanishing the mirror effect: The influence of prior history & list composition on recognition memory

Melody Dye (meldye@indiana.edu), Michael N. Jones (mnjones@indiana.edu), & Richard Shiffrin (shiffrin@indiana.edu)
Department of Psychological & Brain Sciences, Indiana University, Bloomington

Abstract

In the study of recognition memory, a mirror effect is commonly observed for word frequency, with low frequency items yielding both a higher hit rate and lower false alarm rate than high frequency items. The finding that LF items consistently outperform HF items in recognition was once termed the "frequency paradox", as LF items are less well represented in memory. However, recognition is known to be influenced both by 'context noise'—the prior contexts in which an item has appeared—and 'item noise'—interference from other items present within the list context. In a typical recognition list, HF items will suffer more interference than LF items. To illustrate this principle, we deliberately manipulated both the contexts in which critical items had been encountered prior to study, and the confusability of targets and distractors. Our results suggest that when noise sources are balanced, the mirror effect disappears.

Keywords: recognition memory; context noise; item noise; prior history; semantic similarity; orthographic similarity; list length; word frequency; mirror effect; differentiation

Introduction

In a typical episodic memory experiment, subjects are introduced to a new item or list of items within the experimental context, and memory is then tested for that set. In an old-new recognition task, for example, subjects study a list of words, and then are asked to discriminate words seen at study (***targets***) from non-studied words (***foils***). What is potentially challenging about the task is that subjects must identify just those items seen at study from all other words encountered in everyday life. In other words, they must discriminate between pre-experimental familiarity with the test items and familiarity that is specific to the task context. Performance at test is assessed by the *d'* sensitivity index, common in signal detection, which computes the distance between the means of the hit-rate distribution (the probability of correctly identifying a target) and the false alarm-rate distribution (the probability of misidentifying a lure), normalized by the common standard deviation.

The study of recognition memory has been dominated by ***global matching models***, which are variants on signal detection models. These capture the idea that recognition of a particular item depends not solely on the properties of the item itself, in isolation, but also on other items present in memory (for a review, see Clark & Gronlund, 1996). When a particular item is tested, the available cues—such as item and context—form a joint probe of memory, which is accessed in parallel. This global search yields a numerical value, which prompts an 'old'-response if it exceeds some criterion. The returned value is variously understood as the global familiarity of the test item, the match between the test item and the contents of memory, and the activation strength of memory for that test item. How the value is calculated also depends on the process specified by the model, ranging from the sum of retrieval strengths (Gillund & Shiffrin, 1984) to the match between vectors (Murdock, 1982).

A general assumption is that the distribution of familiarity values will have a higher mean for studied items than for unstudied lures. However, interference at retrieval can arise from two sources: item noise (McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997) and context noise (Dennis & Humphreys, 2010). ***Item noise*** refers to the probability of a chance match between an item at test and memory traces for other studied items. ***Context noise*** refers to the probability of a match between the experimental context and other contexts in which the tested item has occurred.

To study how noise arises in recognition, designs typically manipulate one or more variables of interest, such as the number of items on the list (list length), the number of repetitions or exposure duration of a particular item at study (item strength), and the number of repetitions of the list (list strength). The properties of the items may also be systematically manipulated: For instance, a list might be comprised of an equal proportion of randomly selected high (HF) and low frequency (LF) items (mixed list), or alternately, contain only items selected from one frequency band (pure list).

The study presented here was designed to investigate the extent to which item and context noise affect recognition processes, by systematically manipulating both the prior contexts in which critical items had been encountered (Kinsbourne & George, 1974; Estes & Maddox, 1997), and the similarity of items within the list (Hintzman, 1988; Shiffrin, Huber, & Marinelli, 1995). There are a number of reasons to believe that these manipulations to item and context noise should differentially affect items as a function of their frequency, which we shall now review.

Word Frequency Effects in Recognition

In studies of recognition memory, a ***mirror effect*** is commonly observed in subject performance with regards to item frequency: Compared to HF items, LF items are better discriminated, yielding a higher hit rate (HR) and lower false alarm rate (FAR) (Glanzer & Adam, 1985; Glanzer et al., 1993). A similar effect is observed in forced-choice recognition paradigms that include (in addition to the usual old-new pairs) old-old and new-new pairs, for which there is no 'correct' answer. Subjects in these studies preferentially choose LF words over HF words for target pairs, and HF words over LF words for foil pairs (Glanzer & Bowles, 1976).

In assessing word frequency effects (WFE), there are a few wrinkles to consider: For one, the mirror effect is not always perfectly symmetric; the performance gap is typically smaller for hits than false alarms, and there may be differences in criterion as well as sensitivity (Hintzman, Caulton, & Curran, 1994). For another, recognition performance does not vary monotonically with word frequency. Instead, LF words only appear to benefit when subjects have some familiarity with them (Schulman, 1976; Zechmeister, Curt, & Sebastian, 1978). Further, when frequency is considered as a continuous variable,

HR follows a U-shape, with the greatest decrements observable in the mid-frequency band (Hemmer & Criss, 2013).

Clearly, differences in performance on high and low frequency items cannot be reduced to their differential repetition in prior history. Instead, there appear to be multiple, interacting factors at play in producing differences between frequency bands. Some of the key factors include: 1) how well a particular item is differentiated from other items in the lexicon given prior learning history; 2) how discriminable that item is from other items on the present list, given the specific list composition; and 3) the degree to which that item will be associated with the present task context, which should be inversely related to the number of distinct contexts in which it has previously appeared.

Differentiation over Learning

How do these dimensions differ for high and low frequency items in a standard recognition experiment? A common theoretical assumption is that greater experience with an item over learning leaves it better ***differentiated*** in memory—the idea being that repeated exposure acts to increase similarity between the studied item and its memory trace, while decreasing the similarity between its trace and all others (Criss, 2006). This view of repetition falls naturally out of discriminative learning models (Rescorla, 1972; Ramscar et al., 2010), in which cue weights are tuned to produce ever more efficient responding. It is also common to the study of categorization, where it is well known that similarity relations among items change in systematic ways as a function of learning (Nosofsky, 1986).

Models of recognition memory formalize this notion in slightly different ways. In the Retrieving Effectively from Memory (REM) model, each time an item is encountered within a given context, its episodic memory trace is updated, accruing more complete and accurate feature information (Shiffrin & Steyvers, 1997). Thus, more encoding opportunities lead to a higher probability of self-match and a lower probability of matching an unrelated item, leading to 'differentiation' of the trace. Likewise, in the subjective likelihood model (SLiM), initial experience with an item yields a noisy and underspecified representation of its features, which is refined over learning (McClelland & Chappell, 1998). A logical inference from these models is that HF items—by virtue of having been experienced more often, and in more contexts—should be better differentiated from one another in long-term memory than are LF items.

If HF items are better learned, why do they not routinely outperform LF items in recognition, as they do in other memory paradigms, like lexical decision and recall? This is known as the ***frequency paradox*** (Gregg, 1976).

Context Noise

To address this question, it helps to consider how memory for a word depends on the contexts in which it has been previously encountered. Events stored in memory are comprised of both information that was central to processing (the item itself) and information that was available in the peripheral environment (the broader context). The contextual information that is encoded may include aspects of the temporal or physical context in which an item is presented, the emotional state of the learner, and so on (Murnane, Phelps, & Malmberg, 1999; Smith, Glenberg, & Bjork, 1978).

Because contextual information is stored alongside item information, memory for an item is facilitated when there is a high degree of match between its encoding and retrieval contexts, a principle known as ***encoding specificity*** (Tulving & Thomson, 1973). However, similarity between contexts can also produce interference in tasks, like recognition, that require discrimination among encoding contexts. In making an accurate recognition judgment, one of the key challenges is in distinguishing between familiarity with the item from the study list and familiarity from previous experiences in everyday life. The more prior contexts in which an item has occurred, and the more confusable those contexts with the study list, the harder the problem.

One way to demonstrate this is by incidentally exposing subjects to critical targets and lures in a ***familiarization*** phase prior to study, which shares many contextual features with the recognition task (e.g., the location, time of day, etc). Recognition for pre-exposed items is reliably impaired (Kinsbourne & George, 1974; Tulving & Kroll, 1995). Another method is to select list items that vary in their ***contextual diversity*** (CD)—i.e., the number of different pre-experimental contexts in which they have appeared. When CD varies, items with higher diversity scores are less well recognized overall, with a lower HR and higher FAR (Jones, Johns, & Reccia, 2012; Steyvers & Malmberg, 2003).

These findings establish context noise as an important source of interference at retrieval. Importantly, context noise is also a key dimension on which HF and LF items differ. Not only are HF words experienced more often than LF words, they are experienced in a more variable set of verbal contexts (Adelman et al., 2006; Jones, Johns, & Reccia, 2012). Given their high frequency of occurrence in text and speech, they are also more likely to have been experienced more recently (Scarborough, Cortese, & Scarborough 1977; Anderson & Schooler, 1991). Relative to LF items, the contexts in which HF items are experienced will thus be more confusable with the study list, significantly increasing the difficulty of the recognition task for those items.

Item Noise

Another clue to the "frequency paradox" concerns how memory for a single item depends on the composition of the surrounding list. As von Restorff (1933) demonstrated in a classic experiment, distinct items fare well on tests of recognition. For example, in a 10 item-list comprised of 9 nonsense syllables and 1 number, the number is recalled with far greater accuracy than the syllables. However, the extent to which a particular item benefits from its ***distinctiveness***—i.e., its dissimilarity from other items—depends crucially on how dissimilar the rest of the items on the list are from each other.

To illustrate this idea, von Restorff placed the lone number on a list with several equally unrelated items, including "a syllable, a color patch, a single letter, a word, a photograph, a symbol, an actual button, a punctuation mark, and the name of a chemical compound" (as reported by Hunt, 1995, p. 109). Unsurprisingly, once all the items were similarly distinct, no advantage for the lone number was found. A benefit only obtained when the other items were clustered in similarity space relative to the critical item. That is, "similarity must establish a context in which difference functions" (Hunt, 1995).

The ***distinctiveness hypothesis*** proposes that memory for a given item should vary inversely with its featural overlap with other items at study (Hunt & Mitchell, 1982). In line with this, when subjects are asked to remember a list of statements that are either congruent or incongruent with their expectations, incongruent-facts tend to be advantaged in recall—but only so long as they comprise a minority of the list (Hastie & Kumar, 1979). Parallel results have been reported for word recall, where it has been found that orthographically or semantically

unusual items only benefit when presented with common ones (Hunt & Eliott, 1980; see also Zechmeister, 1972). This is consistent with a surprisal-based account of the von Restorff effect (Green, 1956).

In recognition, it is clear that distinctiveness matters both at ***encoding*** and at ***retrieval***. Researchers as far back as Postman (1951) have observed that performance on tests of recognition memory varies inversely with the similarity of items at study and at test, and thus, with the choice of distractor (Anisfield & Knapp, 1968; Bahrick, Clark, & Bahrick, 1967). For example, in face recognition, distinctive faces are better recognized than common faces when lures are selected at random, but recognized more poorly when the similarity of lures to targets is controlled (Davidenko & Ramscar, 2005).

Importantly, distinctiveness is a feature on which low and high frequency items are bound to vary. LF words are, on average, more ***orthographically distinctive*** than HF words—comprised of more rare letters, and more uncommon combinations of letters (Estes & Maddox, 2002; Malmberg et al. 2002)—and belong to much sparser orthographic neighborhoods, with both fewer and rarer neighbors (Landauer & Streeter, 1973).

In a random selection of words, LF items will also be more ***semantically distinctive*** than their HF counterparts. This is guaranteed by the distributional properties of the lexicon—specifically the fact that LF words are drawn from a much larger sample than their HF counterparts (using a 1 per million word cutoff, 80% of all words can be classified as low frequency; van Heuven et al. 2014). As a result, LF items will be less semantically similar to one another, on average, than HF items. A variety of measures of semantic richness attest to this: LF words have fewer closer associates (Deese, 1960; Balota et al., 2004), fewer close semantic neighbors (Pexman et al., 2008), and more sparse network connectivity (Steyvers & Tenenbaum, 2005).

While HF items are better differentiated in memory, they are also drawn from a much more tightly clustered similarity space, both in terms of their surface and semantic features. When presented in a mixed lists of randomly selected items, they should thus be less distinctive at encoding and more confusable at test.[1]

Study

In standard recognition experiments, there is a significant imbalance between frequency bands. When item selection is random, LF words should tend to be more orthographically and semantically distinctive than HF words, suggesting that they demand more attentional resources at encoding, and are less confusable with frequency-matched distractors at test. At the same time, their occurrence as a list item is less confusable with other, previous occurrences: LF items have been experienced in fewer, less diverse contexts, and are less likely to have been experienced recently.

In this study, our goal is to bring the sources of noise for high and low frequency items more in balance. To accomplish this re-balancing act, we manipulated two key variables: (1) recency of exposure ('context noise') and (2) inter-item similarity ('item noise'). When these noise sources are equalized, HF items, which are better represented in memory, should outperform LF items.

Design In the ***familiarization phase*** of the study, subjects completed a simple reading comprehension test in which they were incidentally exposed to a set of critical words. Following a short delay, subjects returned to complete a list recognition task in which they studied a list of words, and at test, were asked to distinguish between studied items (targets) and novel items (foils).

Context noise was manipulated by inserting previously encountered critical words at study and at test. To assess how recent exposure affected recognition, the study counterbalanced both whether a given word was encountered in reading, and whether it occurred as a target or foil. ***Item noise*** was manipulated by selecting control words for the recognition task from dense semantic categories (**Figure 1**). To assess for frequency effects, both critical and control words were evenly divided between high and low frequency bands.

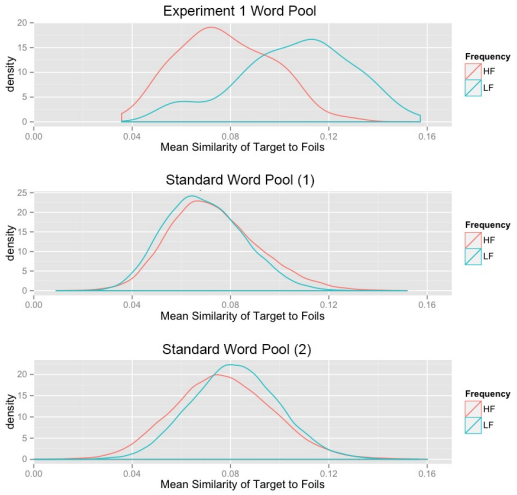

Figure 1: The average semantic similarity of targets to distractors in lists randomly generated from the Exp. 1 control items, as compared to standard episodic word pools (see Dye et al., 2017 for methodology). Drawing items from semantic categories disproportionately increases similarity for LF items.

Participants 54 undergraduate students at Indiana University participated in the experiment for course credit. All were native American English speakers with normal or corrected-to-normal vision. 3 subjects were excluded from the analysis for performing at chance on the reading comprehension portion of the experiment.

Materials Two word lists were constructed (see **Appendix**), each of which comprised 40 critical words: 20 HF (165 occurrences/million) and 20 LF (1 occurrence/million), frequency matched across lists, using counts drawn from the Corpus of Contemporary American English (COCA: Davies, 2010).

In addition, an inventory of 240 control words was created, drawn from sixteen semantic categories (such as 'music' and 'time'). Half of these semantic categories were comprised of HF items, and half LF items. These control items were included to assess how item noise affects recognition. Introducing

[1] This conclusion fits well with the finding that word frequency effects in recognition are closely related to list composition. Systematically varying the frequency of targets and foils in pure list conditions neatly illustrates this point. While a list of LF targets is similarly well-discriminated when paired with a set of HF or LF foils, foil-frequency dramatically affects discrimination of HF targets. When paired with LF foils, the error rate is close to zero; when paired with other HF foils, the error rate far exceeds that of LF targets (Underwood & Freund, 1970). In line with this, raising the proportion of HF items on a list increases the magnitude of the WFE (Dorfman & Glanzer, 1988; Malmberg & Murnane, 2002).

semantic categories should disproportionately amplify item noise for LF items, by increasing the semantic and orthographic confusability of targets and distractors (**Figure 1**).

To create reading materials for the comprehension task, short passages were excerpted from the collected works of the notable Columbian author, Gabriel Garcia Marquez. Specifically, for each word on each of the lists, a passage containing that word was identified and paired with a true statement that synthesized the sentence in which the word had occurred. Affirming the statement as true relied on correct comprehension of the word. Additionally, 20 control passages, which contained no critical items, were selected and paired with a false statement. Each critical word appeared in only one of all possible paragraphs, and only once in that paragraph.

To gauge how pre-exposure affected recognition accuracy and response time, four counterbalanced conditions were created, such that across subjects, each critical item was presented as both a ***target*** and as a ***foil***, and was either ***pre-exposed*** (encountered once previously in reading) or ***novel*** (occurring for the first time in the recognition task).

Study lists comprised 40 critical items and 120 control, and test lists comprised all 160 targets and an additional 160 foils, with the same 1:3 distribution between critical and control items. Here again, controls were evenly split between high and low frequency items, drawn from the same part of the frequency distribution as the critical words.

Procedure In the first stage of the experiment, subjects completed a self-paced reading comprehension task in which they read a series of short passages and, following each paragraph, were presented with a short statement and asked to determine whether it was *true* or *false*. Subjects then moved to a different experiment room to complete a 20-minute distractor task, in which they solved a series of tangram puzzles. They then returned to the original room to complete the list recognition task.

At study, 160 words were presented on a computer monitor for 1s each, separated by a 100 ms ISI. At test, subjects were presented with a new set of items, and asked to judge whether a given item had been presented at study. Testing consisted in 320 self-paced recognition trials, with up to 5s to respond. Order of presentation for passages and for list items was randomized.

Results Looking first to the control items, which were drawn from tightly clustered semantic categories, but did not vary in their exposure history: Welch two-sample t-tests confirmed that —consistent with the typical finding—LF targets had a significantly higher HR than HF targets, both by items [$t(212.63)=-4.12$, $p<.0001$] and by subjects [$t(99.34)=-3.20$, $p=0.002$]. However, the FAR for LF and HF foils was not significantly different ($p>.5$), and the speed of correct rejections was slower for LF foils, both by items [$t(234.24)=-5.85$, $p<.0001$] and (marginally) by subjects [$t(97.82)=-1.70$, $p<.0.092$].

Performance on control items thus shows a marked departure from the standard mirror effect: The typical FAR advantage for LF items disappears, and LF foils are more slowly rejected than HF foils (**Figure 2**). The trends captured here are robust over the course of testing (see **Figure 6** for contrast). This finding is consistent with the notion that the introduction of semantic categories differentially increases item noise for low frequency items, diminishing the typical LF advantage. However, LF control items still outperformed HF control items overall—the increase in FAR was balanced by the sustained HR advantage.

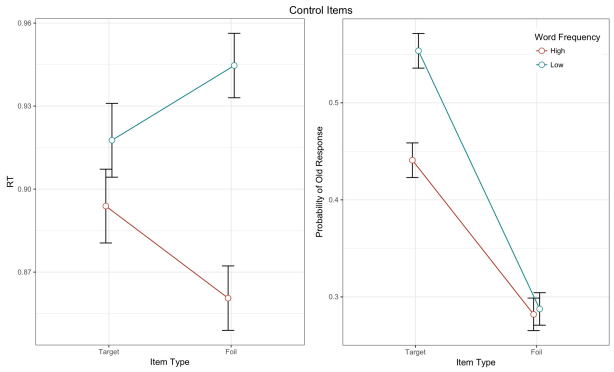

Figure 2: Control item performance for correct RT (right panel) and *p*(old) (left panel), shown by frequency and trial type. Error bars are SEM.

An identical pattern is observable for the critical items with no prior exposure (**Figures 3, 4**). However, for these items, the mirror effect disappears completely following exposure at reading, and overall performance for HF and LF items draws even (**Figure 3**). This is because while *p*(old) increases overall, the LF FAR increases sharply, far outstripping that of the HF foils (**Figure 4**).

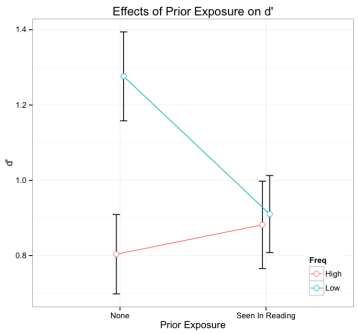

Figure 3. The effect of prior reading exposure on critical items, as measured by *d*' (using a 1/2N correction).

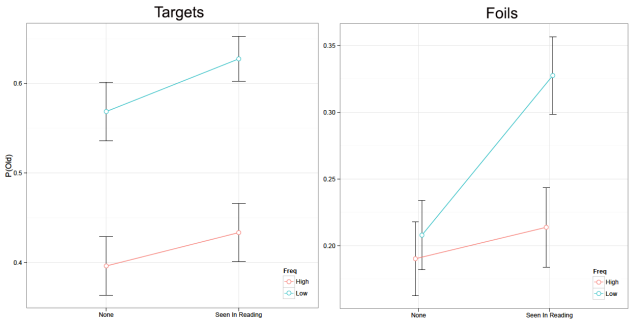

Figure 4: The effect of prior reading exposure on *p*(old), graphed by frequency and trial type. Error bars are SEM.

Performance on critical items (**Figures 4, 5**) can be broken down as follows: For ***targets***, there was a main effect of item frequency on accuracy [$F(1,50)=28.42$, $p<0.001$], and a main effect of exposure condition both on accuracy [$F(1,50)=3.35$, $p=.073$] and correct RT [$F(1,50)=14.36$, $p<.0005$]. Subjects were more likely to affirm LF targets overall, and to more quickly and (marginally more) accurately affirm targets that had previously been seen in reading.

For ***foils***, the picture was somewhat more complicated, but no less consistent. For response time, there was a main effect of

item frequency on correct RT [$F(1,50)=5.87$, $p<0.02$], but no effect of prior context. For accuracy, there were main effects of item frequency [$F(1,50)=3.98$, $p=0.052$] and prior context [$F(1,50)=14.82$, $p<.001$], modulated by a significant interaction between frequency and context [$F(1,50)=4.66$, $p<.05$]. Post hoc analyses (Tukey HSD) indicated that previous exposure significantly increased the FAR for LF items ($p<0.001$) but not HF items ($p>.5$), and that the FAR for pre-exposed LF items was significantly higher than for pre-exposed HF items ($p<0.005$).

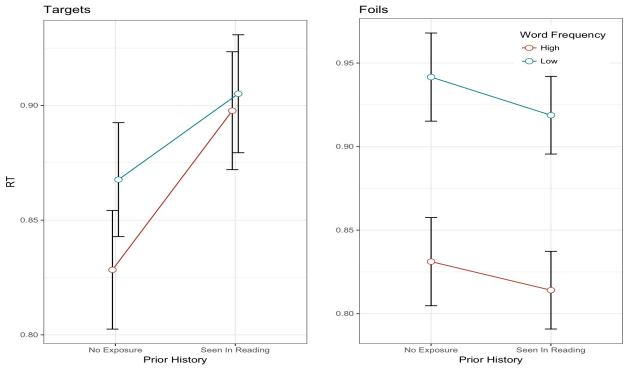

Figure 5: The effect of prior reading exposure on response latency, graphed by frequency and trial type. Error bars are SEM.

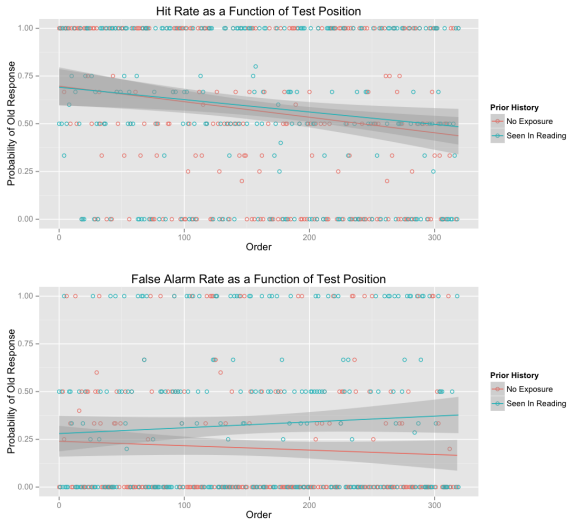

Figure 6: *P*(old) to critical LF items as a function of test position for hits (top) and false alarms (bottom). Trend lines are generated by the glm smoothing method in ggplot2.

To summarize: For HF items, the primary effect of prior exposure was to increase the speed and accuracy of hits. This effect was also observable for LF targets, and the time to execute a hit was similar for HF and LF targets. However, with LF items, the recency manipulation led to an overall bias in *p*(old), such that both the HR and the FAR were significantly higher than that of HF items. The dramatic increase in FAR, as a result of item and context noise, is mirrored by the finding that correct rejections were significantly slower for LF foils across both exposure conditions.

In this study, the magnitude of the performance drop for LF items is, in part, a function of testing (Annis, Malmberg, Criss, & Shiffrin, 2013). At the beginning of testing, no effect of prior history was apparent: the HR for exposed and unexposed critical items was identical, as was the FAR. However, while the HR for LF items uniformly declined over trials, the pattern of false alarms diverged depending on prior exposure (**Figure 6**): Whereas for unexposed items, the FAR showed a steady downward trend, for previously encountered items, precisely the opposite was true. This suggests that the ability to discriminate prior context decreased with continued testing. By contrast: For HF items, while a similar decline in HR is observable over testing, the FAR remains constant, and exposure condition does not appear to interact with these trends.

General Discussion

This paper investigates the sources and robustness of the mirror effect for normative word frequency, finding that under the right set of experimental conditions, it disappears. In particular, when noise sources for high and low frequency items are balanced, LF items prove to be more confusable than better-learned HF items.

Word Frequency Effects

The aim of the present study was to examine how item and context noise interact with word frequency effects. Item noise was manipulated by selecting control items from a small set of semantically cohesive categories, such as 'music' and 'cooking' (Shiffrin, Huber, & Marinelli, 1995). Context noise was manipulated by incidentally exposing subjects to critical items prior to the recognition task (Kinsbourne & George, 1974; Tulving & Kroll, 1995). Both noise sources have been found to impair recognition in a similar fashion: While these manipulations lead to an overall increase in the probability of responding 'old', the increase in hits is slower than the concomitant increase in false alarms, leading to a general decline in discriminability. For instance, when categories of items are present within a recognition list, hits and false alarms increase monotonically with the number of items within each semantic category, such that discriminability decreases as a function of category size (Hintzman, 1988). Similarly, when items are incidentally exposed prior to study, confusability increases as a function of the number of prior exposures (Criss & Shiffrin, 2004; Chalmers & Humphreys, 1998), and as the delay between the familiarization and recognition phases decreases (Maddox & Estes, 1997).

While previous research has tended to focus on how noise affects items from within a single frequency band, our experiment assessed how items were differentially affected as a function of their frequency. A close analysis of the mirror effect for recognition suggests that it derives from the distinctiveness of LF items relative to their HF counterparts. Specifically—in a random selection of items, LF targets will be more distinctive at study, and more distinctive at test compared to foils; in addition, the contexts in which they have previously occurred will be less confusable with the present study context.

In our study, these advantages are systematically mitigated. Introducing verbal categories entails that items will be sampled from a dense semantic space, rather than randomly from the lexicon at large. This selects for LF items that are more similar to each other than HF items, rendering them more confusable at test. Likewise, pre-exposing critical items guarantees that all such items, regardless of frequency, will have recently been experienced in a highly similar, confusable context. If the usual LF FAR advantage is mediated, at least in part, by the greater distinctiveness of randomly selected LF targets relative to potential lures, and by the greater distinctiveness of their prior contexts of occurrence, then these manipulations should diminish or reverse that advantage.

Our findings comport well with this proposal. The item noise manipulation disappeared the LF FAR advantage both for control items and for critical items with no prior exposures: LF foils attracted a similar number of false alarms as HF foils and were rejected significantly more slowly. (A similar result has been reported when orthographic similarity among items is controlled, and lures are orthographically matched to targets; Hall, 1979; Malmberg, Holden, & Shiffrin, 2004).

The context noise manipulation amplified this effect, fully reversing the FAR advantage in favor of pre-exposed HF items, a trend that intensified over the course of testing. This occurred because while pre-exposure dramatically increased the LF FAR, it had a negligible effect on HF items. These manipulations thus vanished one half of the standard mirror effect, equalizing the overall discriminability of HF and LF items.

Nevertheless, LF items maintained a strong HR advantage. There are a number of possible theoretical explanations for this result: LF items may have garnered more attentional resources at study (Malmberg & Nelson, 2003), been more easily associable with the present task context (Hirshman, Whelley, & Palij 1989), or simply been a better match to their own memory traces at retrieval. All these explanations are potentially consistent with the results of the present study, but beyond its scope to establish; further experimental work is needed to distinguish among these accounts.

Modeling Accounts

Empirical results like those presented here can provide important constraints on representational assumptions in modeling (Criss & Shiffrin, 2004). For example, to account for the standard mirror effect, the REM model assumes that LF items possess more rare features than HF items, features that are more diagnostic (Steyvers & Shiffrin, 1997). This implies 1) that LF targets will be a better match to their own memory traces than HF targets (resulting in a higher HR), and 2) that LF foils will be less likely to share features in common with targets than HF foils, resulting in less spurious matches (resulting in a lower FAR). REM thus correctly predicts that when targets are matched with highly similar foils, the FAR for LF items should substantially increase, diminishing or reversing the standard mirror effect. REM can also be modified to account for the finding that increasing the proportion of HF words on a list decreases the FAR for LF items, by assuming that the distinctiveness of the LF items at study leads to better encoding (Malmberg & Murnane, 2002).

Likewise, virtually all models of recognition memory incorporate the idea that an item's prior contexts of occurrence are a critical source of interference (Dennis & Humphreys, 2000). A common assumption is that both item and context information are stored at encoding and that similarity between the study context and prior experiences gives rise to interference at retrieval. What varies is how: In some models, the item and context on the current trial form a joint probe of memory (Gillund & Shiffrin, 1984). In others, the context cue first acts to restrict the subset of activated memory traces to those that match the current context, prior to comparing the item cue to the resulting set (Shiffrin & Steyvers, 1997).

In future work, it may be profitable to use item representations derived directly from the items themselves, by quantifying the lexical and semantic characteristics of a given list or word pool (Dye et al., 2017). Models can then be constructed and tested against the true properties of the stimulus set, permitting cleaner adjudication between competing accounts.

Acknowledgments

This research was funded by an NSF graduate fellowship to MD. Many thanks to Michael Ramscar, Brendan Johns, Gregory Cox, and Rui Cao for insightful comments and discussion.

References

Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science*, 17(9), 814–823.

Anderson, J.R. & Bower, G.H. (1972). Recognition and retrieval processes in free recall. *Psychological Review*, 79(2), 97-123.

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the Environment in Memory. *Psychological Science*, 2(6), 396–408.

Anisfeld, M., & Knapp, M. (1968). Association, synonymity, and directionality in false recognition. *Journal of Experimental Psychology*, 77(2), 171–179.

Annis, J., Malmberg, K. J., Criss, A. H., & Shiffrin, R. M. (2013). Sources of interference in recognition testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39(5), 1365–1376.

Bahrick, H. P., Clark, S., & Bahrick, P. (1967). Generalization gradients as indicants of learning and retention of a recognition task. *Journal of Experimental Psychology*, 75(4), 464–471.

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. (2004). Visual Word Recognition of Single-Syllable Words. *Journal of Experimental Psychology*, 133(2), 283–316.

Chalmers, K. A., & Humphreys, M. S. (1998). Role of generalized and episode specific memories in the word frequency effect in recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24(3), 610–632.

Clark, S. E., & Gronlund, S. D. (1996). Global matching models of recognition memory: How the models match the data. *Psychonomic Bulletin & Review*, 3(1), 37–60.

Criss, A. H. (2006). The consequences of differentiation in episodic memory: Similarity and the strength based mirror effect. *Journal of Memory and Language*, 55(4), 461–478.

Criss, A. H., & Shiffrin, R. M. (2004). Context Noise and Item Noise Jointly Determine Recognition Memory: A Comment on Dennis and Humphreys (2001). *Psychological Review*, 111(3), 800–807.

Davidenko, N. & Ramscar, M. (2005) Distinctiveness effects in face memory vanish with well-controlled distractors. *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, Mahwah, NJ.

Deese, J. (1960). Frequency Of Usage And Number Of Words In Free Recall: The Role Of Association. *Psychological Reports*, 7(2), 337–344.

Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review*, 108(2), 452–478.

Dorfman, D., & Glanzer, M. (1988). List composition effects in lexical decisions and recognition memory. *Journal of Memory and Language*, 27, 633–648.

Dye, M., Ramscar, M., & Jones, M. (2017). Representing the richness of linguistic structure in models of episodic memory. *Proceedings of the 39th Annual Conference of the Cognitive Science Society*.

Estes, W. K., & Maddox, W. T. (2002). On the processes underlying stimulus-familiarity effects in recognition of words and nonwords. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(6), 1003–1018.

Eugene, Z. B. (1972). Orthographic Distinctiveness as a Variable in Word Recognition. *The American Journal of Psychology*, 85(3), 425–430.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, 91(1), 1–67.

Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory & Cognition*, 12, 8–20.

Glanzer, M. & Bowles, N. (1976). Analysis of the word-frequency effect in recognition memory. *Journal of Experimental Psychology: Human Learning and Memory*, 2(1), 21-31.

Glanzer, M., Adams, J. K., Iverson, G. J., & Kim, K. (1993). The regularities of recognition memory. *Psychological Review*, 100, 546–567.

Hastie, R., & Kumar, P. A. (1979). Person memory: Personality traits as organizing principles in memory for behaviors. *Journal of Personality and Social Psychology*, 37(1), 25–38.

Hemmer, P., & Criss, A. H. (2013). The shape of things to come: Evaluating word frequency as a continuous variable in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39(6), 1947–1952.

Hintzman, D. L. (1988). Judgments of Frequency and Recognition Memory in a Multiple-Trace Memory Model. *Psychological Review*, 95(4), 528–551.

Hintzman, D. L., Caulton, D. A., & Curran, T. (1994). Retrieval constraints and the mirror effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20, 275–289.

Hirshman, E., Whelley, M. M., & Palij, M. (1989). An investigation of paradoxical memory effects. *Journal of Memory and Language*, 28(5), 594-609.

Hunt, R. R., & Mitchell, D. B. (1982). Independent effects of semantic and nonsemantic distinctiveness. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 8(1), 81–87.

Hunt, R. R., & Eliot, J. M. (1980). The Role of Nonsemantic Information in Memory: Orthographic Distinctiveness Effects on Retention. *Journal of Experimental Psychology: General*, 109(1), 49–74.

Hunt, R. R. (1995). The subtlety of distinctiveness: What von Restorff really did. *Psychonomic Bulletin & Review*, 2(1), 105–112.

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. Psychological Bulletin.

Jones, M. N., Dye, M., & Johns, B. T. (2017). Context as an organizing principle of the lexicon. In B. Ross (Ed.), *The Psychology of Learning and Motivation*.

Jones, M. N., Johns, B. T., & Recchia, G. (2012). The role of semantic diversity in lexical organization. *Canadian Journal of Experimental Psychology*, 66, 115–124.

Kinsbourne, M., & George, J. (1974). The mechanism of the word-frequency effect on recognition memory. *Journal of Verbal Learning and Verbal Behavior*, 13(1), 63–69.

Landauer, T. K., & Streeter, L. A. (1973). Structural differences between common and rare words: Failure of equivalence assumptions for theories of word recognition. *Journal of Verbal Learning and Verbal Behavior*, 12(2), 119–131.

Maddox, W. T., & Estes, W. K. (1997). Direct and indirect stimulus-frequency effects in recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23(3), 539–559.

Malmberg, K. J., & Murnane, K. (2002). List composition and the word-frequency effect for recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(4), 616–630.

Malmberg, K.J. & Nelson, T.O. (2003). The word frequency effect for recognition memory and the elevated-attention hypothesis. *Memory & Cognition,* 31, 35-43.

Malmberg, K. J., Steyvers, M., Stephens, J. D., & Shiffrin, R. M. (2002). Feature frequency effects in recognition memory. *Memory & Cognition*, 30(4), 607–613.

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.

Murdock, B.B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review*, 89(6), 609-626.

Murnane, K., Phelps, M. P., & Malmberg, K. (1999). Context-dependent recognition memory: the ICE theory. *Journal of Experimental Psychology: General*, 128(4), 403–415.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115(1), 39–61.

Pexman, P. M., Hargreaves, I. S., Siakaluk, P. D., Bodner, G. E., & Pope, J. (2008). There are many ways to be rich: Effects of three measures of semantic richness on visual word recognition. *Psychonomic Bulletin & Review*, 15(1), 161–167.

Postman, L. (1951). The generalization gradient in recognition memory. *Journal of Experimental Psychology*, 42(4), 231–235.

Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. (2010). The Effects of Feature-Label-Order and their implications for symbolic learning. *Cognitive Science*, 34(6), 909-957.

Rescorla, R. A. (1988). Pavlovian conditioning. It's not what you think it is. *The American Psychologist*, 43(3), 151–160.

Scarborough, D. L., Cortese, C., & Scarborough, H. S. (1977). Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human Perception and Performance*, 3(1), 1–17.

Schulman, A. I. (1967). Word length and rarity in recognition memory. *Psychonomic Science,* 9, 47–52.

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

Shiffrin, R. M., Huber, D. E., & Marinelli, K. (1995). Effects of category length and strength on familiarity in recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(2), 267–287.

Smith, S. M., Glenberg, A., & Bjork, R. A. (1978). Environmental context and human memory. *Memory & Cognition*, 6(4), 342–353.

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

Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. *Cognitive Science*, 29(1), 41–78.

Tulving, E., & Kroll, N. (1995). Novelty assessment in the brain and long-term memory encoding. *Psychonomic Bulletin & Review,* 2(3), 387–390.

Underwood, B. J., & Freund, J. S. (1968). Errors in recognition learning and retention. *Journal of Experimental Psychology*, 78(1), 55–63.

Zaki, S. R., & Nosofsky, R. M. (2001). Exemplar accounts of blending and distinctiveness effects in perceptual old new recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27, 1022–1041.

Zechmeister, E. B., Curt, C., & Sebastian, J. A. (1978). Errors in a recognition memory task are a U-shaped function of word frequency. *Bulletin of the Psychonomic Society*, 11(6), 371–373.

von Restorff, H. (1933). Über die Wirkung von Bereichsbildungen im Spurenfeld. *Psychologische Forschung*, 18, 299-342.