Journal of Memory and Language 64 (2011) 316–326

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Output interference in recognition memory

Amy H. Criss [a,*], Kenneth J. Malmberg [b], Richard M. Shiffrin [c]

[a] Department of Psychology, Syracuse University, Syracuse, NY 13244, United States
[b] Department of Psychology, University of South Florida, Tampa, FL 33620, United States
[c] Department of Psychological & Brain Sciences and Cognitive Science Program, Indiana University, Bloomington, IN 47406, United States

ARTICLE INFO

Article history:
Received 13 July 2010
revision received 14 February 2011
Available online 22 March 2011

Keywords:
Recognition memory
Interference
Memory models
Output interference
Item-noise
Context-noise

ABSTRACT

Dennis and Humphreys (2001) proposed that interference in recognition memory arises solely from the prior contexts of the test word: Interference does not arise from memory traces of other words (from events prior to the study list or on the study list, and regardless of similarity to the test item). We evaluate this model using output interference, a decline in accuracy as a function of the words presented during test. Output interference is consistent with models that allow interference from words other than the test word, when each test produces a memory trace, and hence a source of interference. Models positing interference solely from prior contexts of the test word itself predict no effect of items presented during test, without added assumptions. We find robust output interference effects in recognition memory. The effect remains intact after a long delay, when study-test lag is held constant, when feedback is provided, and when the test is yes/no or forced choice. These results are consistent with, and support the view that interference in recognition memory is due in part to interference from words other than the current test word.

Introduction

When attempting to remember a specific event, interference is caused by irrelevant memories. This is a well-established and extensively investigated phenomenon (Anderson & Neely, 1996; Crowder, 1976; McGeoch, 1933; Melton & Von Lackum, 1941; Mensink & Raaijmakers, 1988; Murdock, 1974; Raaijmakers & Shiffrin, 1980, 1981; Shiffrin, 1970). Interference in free recall arises when memory traces contain representations of similar items and/or more than one item encountered in similar contexts. Evidence collected over many years suggests that like free recall, item recognition is also subject to interference from traces with similar item and context information (Clark & Gronlund, 1996; Gillund & Shiffrin, 1984; Humphreys, Pike, Bain, & Tehan, 1989; Murdock, 1982). However, this conclusion has been challenged by a model of recognition memory for words that assumes interference arises only from the contextual history of the test word (Bind Cue Decided Model of Episodic Memory, BCDMEM, Dennis & Humphreys, 2001). While no model denies that interference may arise from the prior contexts in which a word has been encountered (cf., Criss & Shiffrin, 2004; Shiffrin & Steyvers, 1997; Steyvers & Malmberg, 2003), BCDMEM makes the strong assertion that this is the *only* factor producing interference, and that stored traces of other words play no role. One test of this assertion can be found by examining output interference: the effect of prior testing of other words before a critical word. BCDMEM claims that neither the number of such prior test words nor their similarity to the words on the study list should affect recognition performance. We evaluate the role of output interference in recognition memory.

Interference in recognition memory

In recognition, subjects study a list of items, and then decide whether items on a test list were studied or not. Dennis and Humphreys (2001) restrict their claims to

[*] Corresponding author.
E-mail address: acriss@syr.edu (A.H. Criss).

0749-596X/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jml.2011.02.003

words, so the primary focus of this article will be recognition memory for words. Assessing whether interference from other words on the study list reduces recognition accuracy depends on the assumptions one makes about how recognition is performed. All models assume that recognition requires the representation of two types of information. *Item information* refers to a representation of the semantic, phonological, visual, etc. content of the to-be-remembered item. This information is usually generated when performing a recall task, for instance, and it is the information that one must determine was encountered on the study list when performing a recognition task. There exists ambiguity about terminology when discussing information about other words coded together with a given word; we term such information *associative context information*. We use *list-context information* to refer to the internal and external factors that comprise the situation in which learning occurs or the to-be-remembered information was presented, other than information about other words (cf, Howard & Kahana, 2002).

Interference refers to memory loss that is the result of the interaction of a retrieval cue (consisting of both item- and context information) with similar traces stored in memory. The more similar are the interfering episodic memory traces, the more difficult it is to recall or make a recognition decision about the test item. This occurs because a typical episodic memory paradigm requires discrimination of an item presented on the recent list from *other* items stored in memory (either those stored during list presentation or those stored in previous lists or prior experience) and from prior experiences of that *same* item (either in previous lists or prior experience). In recognition, item and/or list-context information may be retrieved from traces of the test item or from similar traces, or both. Retrieved item information from memory traces of other similar items (from the list or events prior to the list) produces what is referred to as *item-noise* or *item interference*. Retrieved list-context information from memory traces of other similar items from the list or events prior to the list, or from memory traces of the test item itself from events prior to the list produces what is referred to as *context-noise* or *context interference*. In both cases, the similarity of the retrieved information to the test probe is the source of interference.

Models of recognition memory

The subject of the present investigation is the importance of item interference when words are used as stimuli: Does item information, from traces of other words on the study list, from traces of other words on the test list, or from traces of other words prior to the list, produce interference? Item information from pre-experimental traces of other items probably plays at most a small role because they differ from the retrieval probe in both item and context information. The most important source of item interference should therefore come from item information in traces of other items presented on the study list and the other items presented on the test list because they share context information. Although most models of memory assume that both item and context interference play a role in recognition, BCDMEM raises the possibility that, for words, the only relevant factor is context interference. That is, context information retrieved from traces of the test word stored prior to the study list is the sole source of interference. Thus, we seek to distinguish models of recognition memory that posit both item- and context-noise from those positing only context-noise, respectively referred to as *item-noise models* (e.g., Criss & Shiffrin, 2004) and *context-noise models* (e.g., Dennis & Humphreys, 2001).

In both models, the study trial produces a memory trace consisting of a representation of both item and the context information, and the test probe also consists of both types of information. The difference lies in what traces are retrieved from memory (thereby producing interference). In context-noise models (e.g., BCDMEM), word information in the test probe is sufficient to limit retrieval only to traces of the test word (both from the list, if such a trace exists, and from events prior to the list). In item-noise models, retrieval also occurs from traces of other items from the study list, the more similar the test word and the memory trace the more interference is caused by that trace.[1]

Whatever the source of interference, we submit the item- and context-noise models to a critical test. Item-noise models predict that traces of non-target words should have a negative impact on memory performance. Context-noise models do not predict an effect of other items. Of relevance for the present investigation, such non-target word traces include those that are stored during the sequence of recognition test trials following list study.

Prior tests of the models

Like several item-noise models (e.g., McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997), BCDMEM was designed to predict list composition effects. One of the most important list composition findings is that increasing the extent of encoding of non-target traces has no effect on recognition. This is referred to as the null list-strength effect. BCDMEM naturally predicts a null list-strength effect because item information does not contribute to the recognition decision (e.g., Starns, White, & Ratcliff, 2010). Item-noise models predict the null list-strength effect on the assumption that increasing the amount of information

[1] There are two processes by which interference can take place. In some models, the primary way the recognition decision is made is due to a general sense of familiarity (e.g. Gillund & Shiffrin, 1984; Shiffrin & Steyvers, 1997): the activations of all the memory traces that are retrieved are combined, and a positive recognition decision is made if that combined activation (i.e., familiarity) is high enough to exceed a criterion. In other models, termed dual process, a recall (or recollection) process also plays a significant role: sometimes a particular memory trace is recalled and when that trace matches the probe well enough this is sufficient to produce a positive recognition decision (e.g. Atkinson & Juola, 1974; Malmberg, 2008; Malmberg, Holden, & Shiffrin, 2004; Mandler, 1980; Xu & Malmberg, 2007; see Yonelinas, 2002 for a review). In the single process familiarity models, interference is due to additional familiarity contributed by traces matching in item information, context information or both. If recall also plays a role, interference is due to competition between traces: the chances of sampling and retrieving the desired memory trace are higher if there are fewer competing similar traces (either due to item similarity, context similarity, or both).

stored about an item decreases the similarity of non-target traces (i.e., differentiation, see Criss, 2006, 2009, 2010).

Another list composition effect is the list-length effect. There are several reports of reliable list-length effects (e.g., Cary & Reder, 2003; Gronlund & Elam, 1994; Strong, 1912), but Dennis and Humphreys (2001) attribute them to confounds such as longer study-test lags and reduced attention for longer lists, more displaced rehearsals for shorter lists, and/or the lack of context reinstatement. Critical support for BCDMEM comes from findings that changes in list length have no effect on recognition: Dennis and colleagues (Dennis & Humphreys, 2001; Dennis, Lee, & Kinnell, 2008; Kinnell & Dennis, 2011; Maguire, Humphreys, Dennis, & Lee, 2010) report a null-list-length effect in a number of experiments that implement controls for these assumed confounds. In contrast, item-noise models predict interference from other items, especially from related items. For instance, category-length manipulations affect recognition (Criss & Shiffrin, 2004; cf, Dennis & Chapman, 2010; Neely & Tse, 2009; Shiffrin, Huber, & Marinelli, 1995) and varying the proportion of high-frequency versus low-frequency words on a study list affects recognition accuracy (Dorfman & Glanzer, 1988; Malmberg & Murnane, 2002).

In summary, list composition manipulations, with the exception of null list-strength effects, remain a challenge for both item-noise models and context-noise models.[2] The one finding that supports context-noise models and does not support item-noise models is the null list-length effect. Many other manipulations of list composition support item-noise but not context-noise models. The present experiments adopt a different approach to evaluate the effect of item-noise on recognition memory performance.

Output interference

Thus far, the empirical strategy primarily used to evaluate whether item-noise plays a role in recognition memory has been to manipulate the study list. A different approach is to assess the impact of testing on recognition. Output interference has a significant negative effect on recall (Dong, 1972; Roediger, 1974; Roediger & Schmidt, 1980; Smith, D'Agostino, & Reid, 1970; Tulving & Arbuckle, 1966). For example, output order is the primary determinant of recall accuracy, outshining even the effect of serial position (Dalezman, 1976). While only a few experiments have evaluated output interference in recognition, they also document output interference. Norman and Waugh (1968) and Schulman (1974) found detrimental effects as the number of items tested increased. Murdock and Anderson (1975) replicated these findings and reported longer response times with increasing output position. Further, they found performance drops with an increase in the number of alternatives in a forced choice paradigm. These finding suggest that number of items encountered during testing is negatively related to recognition performance.

Output inference is conceptually consistent with item-noise models of memory. Although output interference has not been explicitly modeled by extant item-noise models recognition memory, this is not due to an inherent limitation of the models, or a conceptual component that is missing from the models. Rather, output interference has heretofore been ignored. However, should the test events themselves be stored in memory (as they surely must be) then the current models can be employed in straightforward fashion to make predictions. Output interference is of course consistent with item-noise models because the traces of the test items before a critical test item will be retrieved and cause interference. In fact, to the extent that the list context of such traces will be even more similar to the test probe's list context than the list context in the traces of study list items, interference from test words might be stronger than interference from study list words. There is not much question that test traces will be stored in memory, not only because it is evident that memory stores all events that occur, but also because there is evidence that learning occurs during testing (e.g., Carrier & Pashler, 1992; Jacoby, Shimizu, Velanova, & Rhodes, 2005; Roediger & Karpicke, 2006a, 2006b). The situation is much different for context-noise models; these models also allow information to be stored at test, but have no inherent mechanism to allow such storage to produce interference. Thus, an exploration of output interference in recognition memory provides a critical test of item- and context-noise models of recognition memory.

Experiment 1

Early studies evaluating output interference in recognition memory are limited in number (Murdock & Anderson, 1975; Norman & Waugh, 1968; Schulman, 1974) and used methods that may be subject to the confounds that Dennis and Humphreys (2001) suggested were responsible for list-length effects (e.g., study-test lag, differential attention across condition, displaced rehearsals, and the lack of context reinstatement). The first experiment is a simple replication of earlier experiments (though the exact details differ). In the second experiment, we add controls to eliminate the potential problems just described. Item-noise models, like REM (Shiffrin & Steyvers, 1997), predict a decrease in performance across test block as new traces are added to memory or traces from the study list are updated. Context-noise models, like BCDMEM (Dennis & Humphreys, 2001), predict no effect of the number of items tested unless augmented by other mechanisms.

Method

Participants

Fifty-six members of the University of South Florida subject pool participated to fulfill course requirements.

[2] BCDMEM is also challenged to account for item distinctiveness effects. In order to predict a null-list-length effect BCDMEM must assume that there is no interference in access to the item representations in memory. Violation of this assumption would produce noise from the occurrence of other items. Thus, BCDMEM predicts that the similarity between the features that comprise the words has no impact on recognition. However, words that consist of more unusual features are better recognized than words that consist of more common features (Criss & Malmberg, 2008; Freeman, Heathcote, Chalmers, & Hockley, 2010; Malmberg, Steyvers, Stephens, & Shiffrin, 2002; Zechmeister, 1972).

Stimulus Materials

For each subject, 900 words were randomly selected from a pool of 1154 nouns with a normative frequency of at least 20 occurrences per million (Kucera & Francis, 1967).

Design and procedure

Each participant studied 6 lists of 75 words presented on a computer monitor for 1 s separated by a 100 ms ISI. After each study list, participants performed a 30-s math task in which they kept a running summation of a series of single digits. The test list consisted of 150 self-paced recognition memory trials. Participants were asked to judge whether the item was presented during the most recent study list. Successive study-test lists were separated by a self-paced break. Words were randomly assigned as a foil or target for each list for each participant. No item repeated across lists.

Results and discussion

For analysis, the data from each of the 6 test lists were divided into 15 blocks of 10 test trials. Fig. 1 plots d' as a function of the test block in the top panel, and hit rates and false alarm rates as a function of test block in the bottom panel. A one-way repeated measure ANOVA revealed a main effect of block for each dependent measure ($F(14, 770) = 6.24$, $p < .001$; $F(14, 770) = 11.29$, $p < .001$; $F(14, 770) = 1.88$, $p = .026$, respectively). Overall, HRs and FARs moved closer together as testing proceeded, to produce an overall decrease in d'. The data replicate prior findings of output interference in yes–no recognition (Murdock & Anderson, 1975; Norman & Waugh, 1968; Ratcliff & Hockley, 1980; Schulman, 1974).

It appears that participants are reducing their criterion to respond 'old' across the first few blocks. After approximately block 3, criterion seems to stabilize and the HRs decrease substantially across test block while the FARs remain fairly stable. The mechanism for setting and dynamically adjusting the criterion is, unfortunately, beyond the scope of memory models and has received little attention in the memory literature (cf., Brown & Steyvers, 2005; Brown, Steyvers, & Hemmer, 2007). We are conducting additional experiments and model simulations exploring the criterion adjustment across the first several test trials and reserve commentary until sufficient data are gathered.

In order to attribute the reduction in accuracy to interference from other test items, we should consider whether the confounds that Dennis and Humphreys (2001) hold responsible for reported list-length effects can explain the effect of test block. Displaced rehearsal is not a logical explanation for poorer performance across test position; if anything participants have more time to rehearse items from late in the test sequence not less.

Another possible explanation holds that participants devote less effort to items later in the test sequence due to boredom, lack of motivation, etc. This seems unlikely given the short duration of the testing procedure. Nevertheless, checking whether a similar pattern might be observed for the first study-test cycle helps to dismiss fatigue or lack of vigilance as the causal factor. Indeed, d' decreased ($F(14, 476) = 4.49$, $p < .0005$), HRs decreased substantially ($F(14, 700) = 4.34$, $p < .0005$), and FARs slightly increased, although not significantly so, as testing of the first list proceeded. Fatigue does not appear to explain the decrease in recognition accuracy. We include feedback on each test trial in one condition of Experiment 2 to alleviate this concern.

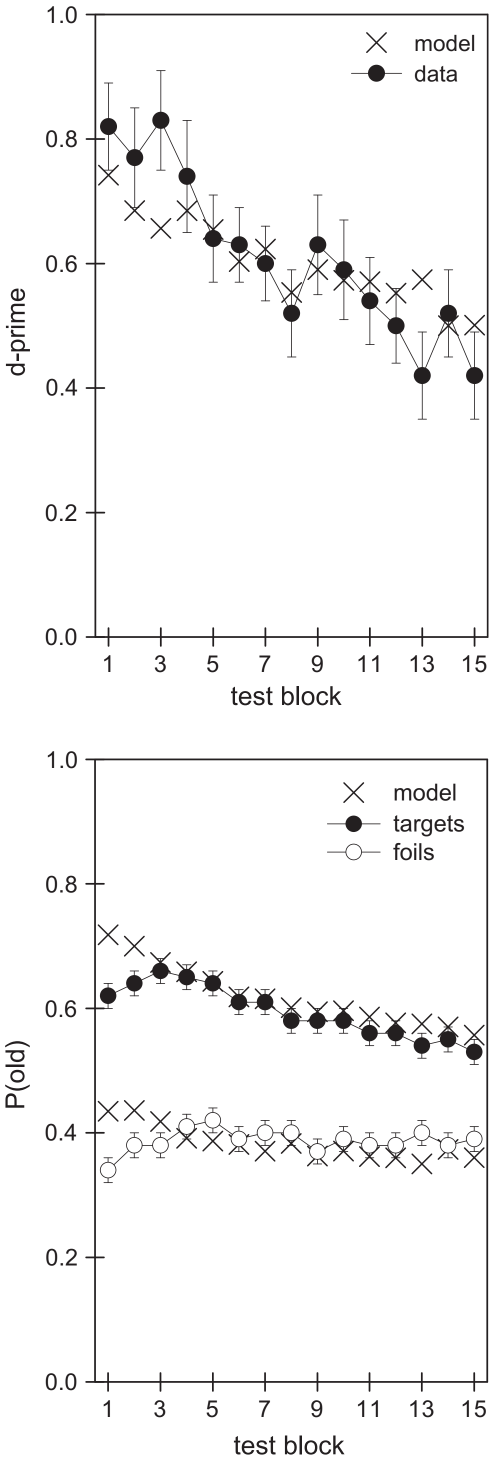

Fig. 1. The top panel shows a decline in d-prime as a function of test block and fits of the REM model. The bottom panel shows the probability of responding old (P(old)) for targets and foils as a function of test block and fits of the REM model. Error bars represent one standard error above and one below the mean.

Study-test lag differences may contribute to output interference. Test items late in the sequence have a longer retention interval than test items early in the sequence. This leaves open the possibility that context changes as a function of time or decay of item information causes the decrease in performance. We evaluate this by including a control for study-test lag in one condition of Experiment 2. To foreshadow, we find output interference effects when study-test lag is controlled.

Finally, consider context reinstatement. Dennis and Humphreys (2001) propose that a short duration between study and test may induce participants to use current context during the test rather than reinstating the study context. This differentially benefits short lists whose overall context is presumably most similar to the current context because of their relative closeness in time. To control for differences in the ability to reinstate context, Dennis and Humphreys used relatively long retention intervals (~8 min). Their logic relies on the assumption that context drifts as a function of time (e.g., Estes, 1955a, 1955b; Mensink & Raaijmakers, 1988), and thus longer retention intervals make the reinstatement of context for short and long lists equally probable. It is unclear why participants in Experiment 1 would allow or choose contexts in their test probes in a way that would make them less efficient during the course of testing, but it is nonetheless possible that the reinstated context and the prior contexts match less well as testing proceeds. We evaluate this concern by including a long delay between study and test in Experiment 2. To foreshadow, we find output interference effects with both immediate and delayed testing.

Experiment 2

In this experiment, we used forced choice testing and include conditions to control for the potential confounds described in Dennis and Humphreys (2001). In one condition accuracy feedback is provided on each trial as an incentive to fully engage throughout the test. We also included a condition where study-test lag is controlled. Last, we include an independent manipulation of the retention interval with a short (0 min) and a long delay (20 min) between study and test. The 20-min retention interval is far greater than retention intervals from Experiment 1 in which a decrease in accuracy occurred. According to Dennis and Humphreys (2001), the long delay requires participants to reinstatement the study context at test. If output interference is driven by lack of context reinstatement, then the effect should not be fully realized following a long delay. Further, if output interference is due to time rather than items, the decrease in accuracy following a long delay should be substantially larger than the decrease in accuracy across the relatively short test list.

To recap, we include a feedback condition and a study-test lag control condition to assess the effects of attention and retention interval, respectively. We also include short and long delays between study and test to evaluate the role of time and context reinstatement. Lastly, we use forced choice testing to further eliminate possible concerns about response bias changing across test block.

Method

Participants

Of the 149 members of the Syracuse University community who participated for course credit, five performed at or below chance ($P(C) < 0.25$) and their data were discarded.

Stimulus materials

The word pool consisted of 800 words between 4 and 11 letters in length with a frequency range between 9 and 13 log frequency ($M = 10.46$) in the Hyperspace Analog to Language corpus (Balota et al., 2002).

Design and procedure

The experiment was a 2 × 3 × 5 mixed design. Delay (0 or 20 min) and control condition (baseline, study-test lag, and feedback) were between-subject variables for a total of 6 different groups of participants. Test block (5 blocks each containing 15 consecutive test trials) was a within-subject factor. Each participant studied a list of 75 words presented on a computer monitor for 1 s with a 100 ms blank screen separating each trial. Participants then engaged in a 30-s math task followed by either a 0 or 20 min delay described below. After the specified delay, participants took a 75 trial 4 AFC test. A 100 ms ISI followed each test response. Unbeknownst to the participants, the test list consisted of five blocks of 15 trials. Participants engaged in study, math, and memory test sections in individual booths. Participants in the *0 min delay conditions* received the test list immediately following the math task. Participants in the *20 min delay conditions* convened in a common area in groups of 3 or 4 to complete a puzzle for 20 min, and then returned to their individual booth to receive the test list. No group successfully completed the entire puzzle, thus all groups worked on the puzzle for the full 20 min.

The details just described make up the *baseline condition*. In the *feedback condition*, participants received immediate feedback ('wrong' or 'correct' appeared on the screen during the ISI) after each test trial. In the *study-test lag condition*, the test items were presented in the same order as the study items, thereby controlling study-test lag. Each item was tested exactly 75 trials after it was studied. Words were randomly assigned to each condition for each participant with the exception of the study-test lag condition.

Results and discussion

Percent correct is plotted for each block of 15 test trials in Fig. 2. Accuracy decreased over test blocks in all conditions. This pattern holds for long and short delays, when feedback is provided at test, and when study-test lag is held constant. Accordingly, a 2 (delay) × 5 (block) × 3 (control condition) mixed ANOVA showed a main effect of block, $F(4, 552) = 66.81$, $p < .001$), a main effect of delay, $F(1, 138) = 7.10$, $p = .009$), and no effect of control condition, $F(2, 138) = 0.38$, $p = .686$). Block and delay did not interact, nor was there a 3 way interaction, (both F's < 1 and p's > .39). There was a block by control condition interaction, $F(2, 552) = 5.03$, $p = .008$). To follow-up on this

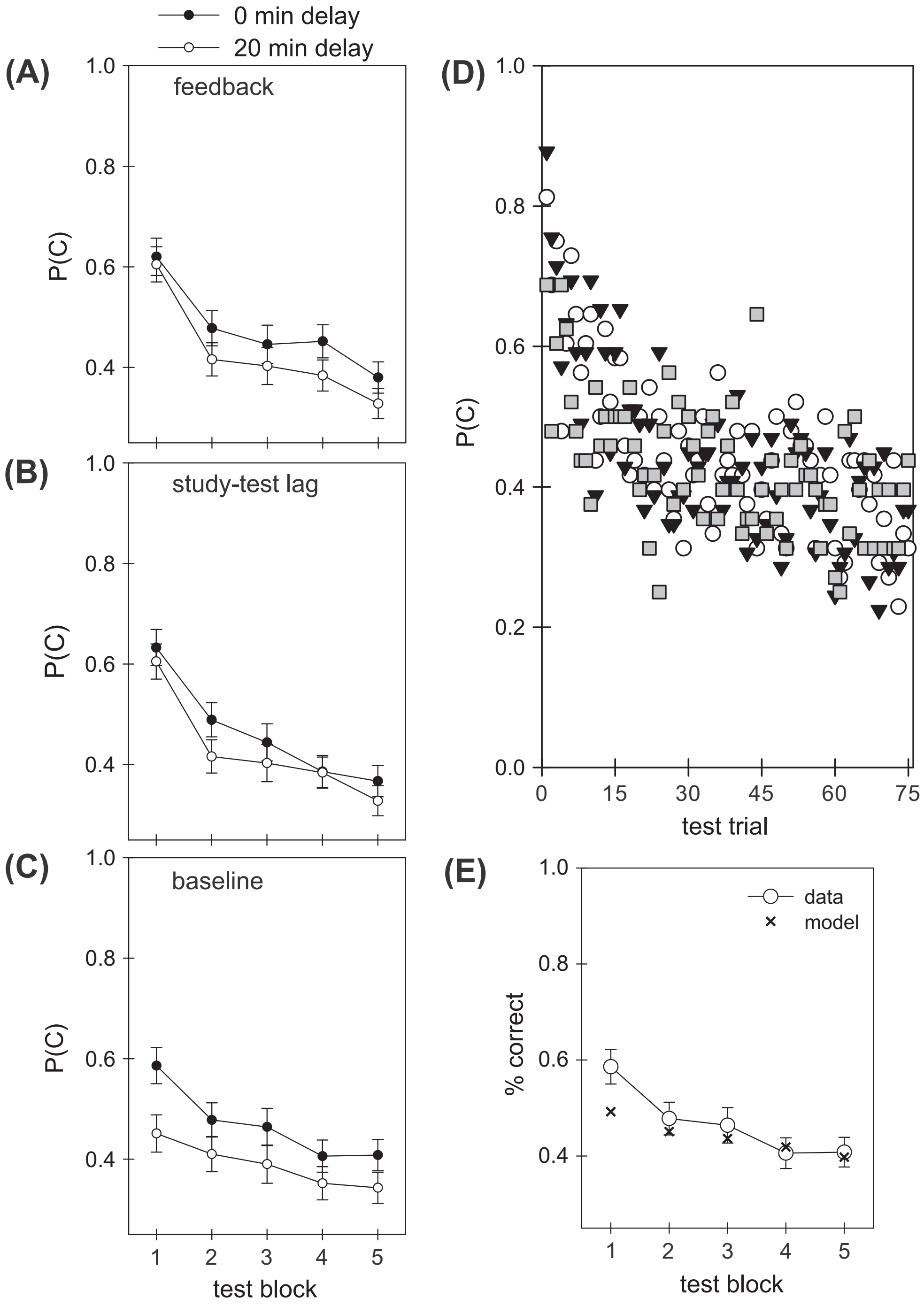

Fig. 2. Mean percent correct (P(C)) as a function of test block for the (A) feedback, (B) retention interval controlled and (C) baseline conditions. Each panel includes data from the immediate testing and delayed testing conditions. Error bars represent one standard error above and one below the mean. The scatterplot in Panel D shows negative correlation of P(C) and test trial for each condition, collapsing over delay condition. Panel E shows the REM predictions for the baseline no delay condition.

interaction, we conducted separate ANOVAs for the 3 conditions and each shows a significant decrease in accuracy across block (all F's > 10.97 and p's < .001). Fig. 2 suggests that the interaction was due to the different slope in the baseline condition compared to the others. Across all conditions the decrease in performance across block is magnified for the first block: there is a larger decrease in performance after block 1 than after subsequent blocks.

The selection of block size was somewhat arbitrary and differed from the block size used in Experiment 1. To gather converging evidence that the block size was not critical to the outcome, we computed the correlation between test trial and mean accuracy for each condition collapsing across delay and participant for simplicity. Panel D of Fig. 2 shows a strong negative correlation between test position and accuracy in the baseline ($r = -.562$, $p < .001$), the study-test lag ($r = -.720$, $p < .001$), and the feedback ($r = -.704$, $p < .001$) conditions. We find no support for the alternate explanations for output interference such as retention interval, lack of attention, or failure to reinstate context. Rather, output interference seems to be driven by the presence of additional test items, supporting an item-noise process in recognition memory.

General discussion

This article reports output interference, a decline in accuracy across test trials, in recognition memory. The effect is present in yes/no and forced choice testing procedures, indicating that the decline in accuracy is independent of response bias. The fact that the magnitude of output interference remains virtually unchanged over a 20-min retention interval suggests that it is unrelated to any variable correlated with the passage of time; nor is output interference due to a failure to reinstate the study context and use it as retrieval cue at the outset of testing. The effect remains intact when the retention interval is equated across the number of intervening study and test trials, disconfirming an explanation based on study-test lag. The decline in performance also appears unrelated to motivational factors; it is present across all test blocks, on the first list tested, and it is unaffected by feedback provided at test. In sum, output interference is a robust phenomenon in recognition memory that is observed even when a number of potentially important confounds are controlled. We turn next to the theoretical implications of output interference.

Implications for context-noise models

Context-noise models such as BCDMEM (Dennis & Humphreys, 2001) propose that all interference is the result of noisy context matches, and that the only role of item information in recognition is the activation of contexts in which that item has occurred. Thus, other items, including those on the study list, those similar to the test item, and those in the mental lexicon but not in the experiment do not contribute to the accuracy of the recognition memory decision. The data supporting this assumption come primarily from manipulations of encoding conditions, in particular null list-length effects (Dennis & Chapman, 2010; Dennis et al., 2008; Kinnell & Dennis, 2011; Maguire et al., in press). In combination with the observed output interference, null list-length effects produce a paradox: Why should the number tested items but not the number of studied items decrease recognition accuracy?

The key challenge for context-noise theory is to settle on an account of output interference that is logically consistent with its account of list-length effects. There are a number of reactions to output interference one might anticipate within the framework of context-noise theory. Although none have been implemented we will nevertheless subject them to a logical level of scrutiny. For example, one might assume that context drifts as a function of time and/or items (e.g., Howard & Kahana, 2002). At the beginning of test, context is reinstated equally well for both long and short study lists such that there is no difference in the match between reinstated and study context and hence no effect of study list length. However, during the course of test the reinstated context that is compared to the output from memory continues to change. The subject does not take the change in the reinstated context into account, and thus its match to the retrieved context decreases. The decrease in match reduces the familiarity of targets and reduces discrimination across output position. We refer to this as the *context-reinstatement hypothesis*.[3]

The context-reinstatement hypothesis seems reasonable at first glance, but has some potential drawbacks. Consider the assumption that at the beginning of test context is reinstated equally well for both long and short lists (otherwise short lists would be recognized better than long lists). Assuming equally good context reinstatement for short and long study lists but increasingly poor context reinstatement for longer test lists seems logically inconsistent. The context-reinstatement hypothesis also includes the assumption that the reinstated context is compared to the context retrieved from memory that represents all the contexts in which an item has ever been encountered; this is how context-noise models predict mirror patterned word-frequency effects (Dennis & Humphreys, 2001). How might such a set of assumptions handle the results from multi-day studies that recycle a pool words across days such as those of Nobel and Shiffrin (2001)? That experiment was comprised of 30 sessions, the same stimulus set was used each day, and the items were randomly assigned to conditions for each session.[4] The challenge for the subject is to determine whether a test item is a target or a foil in the present session, and ignore all prior presentations. After the second session, overall accuracy did not change as a function of session (there were of course differences between conditions but overall accuracy did not change). The failure to observe changes in accuracy across session days suggests that participants manage to solve this difficult context discrimination problem by ignoring the item's presentation in prior sessions. This raises a question: in more common studies with no previous word occurrences in previous lists or sessions, why would participants have trouble excluding the previous traces from life, all of which occur in remarkably different contexts than the experimental context? In other words, why should the false alarm rate be anything but negligible relative to the hit rate?

With respect to the present experiments, we note that the contextual reinstatement hypothesis has some problems accounting for our findings. The decrease in accuracy over the 20 min delay in Experiment 2 is much smaller than the decrease in accuracy over the course of testing (about 3 min). Thus, it appears that the problem does not lie in the ability to reinstate the study context. Rather, it appears that reinstated context is increasingly being subjected to interference during test. What is the source of this context-noise? We consider three hypotheses. According to the first, random fluctuations in the reinstated context

[3] We thank Simon Dennis and Jeff Starns for this suggestion.

[4] Data from each session are not reported in the published version of Nobel and Shiffrin (2001), however the data are available from RMS.

are to blame. This hypothesis is unappealing since one would expect greater random fluctuations over the 20 min retention interval than over the course of testing, but accuracy decreased more during testing than during the retention interval. The second hypothesis is that changes in the context are systematically related to testing, perhaps as the result of the retrieved context being combined with the reinstated context. Presumably, the reinstated context initially comes from the traces stored during study. If so, retrieved study context from target traces ought to refresh the reinstated context, whereas the retrieved context from pre-experimental traces ought to interfere with the reinstated context. In order to predict output interference, therefore, one would have to specify the balance between the costs of retrieving pre-experimental context and the benefits of retrieving study context. Thirdly, one might focus on associative context, the part of context consisting of other words encoded with the test word at both study and test. At study, representations of items rehearsed or otherwise studied in proximity to a given item might provide the associative context. Associative context retrieved from study list traces might contaminate the reinstated context, possibly increasingly so as testing proceeds. However, if that were the case one would expect there to be benefits for testing items in the same order in which they were studied, and such a result was not found in Experiment 2.

Implications for item-noise models

The experimental results are consistent with conceptual basis for item-noise models, that other items are one of the sources of interference in episodic memory (in many situations perhaps the primary source). At present, such models have not explicitly included encoding during test (though their predecessors did, e.g., Gillund & Shiffrin, 1984; Raaijmakers & Shiffrin, 1981). As an illustration that these models naturally account for output interference, we implement it in the REM model (Shiffrin & Steyvers, 1997). We briefly describe the model and refer readers interested in additional details to the original paper (Shiffrin & Steyvers, 1997).

In REM, an incomplete and error-prone copy of the studied item is stored in episodic memory along with the current context. Episodic memory features ($n = 20$ for each item and context information) are initialized as zeros, indicating a lack of information. Features are sampled from a geometric distribution (with parameter $g = .35$). During study each feature is stored with some probability (u^*) otherwise a value for that feature is not stored (i.e., incomplete storage). Given that a feature value is stored, the correct value is stored with some probability ($c = 0.70$). Otherwise, a random feature value selected from the geometric distribution is stored (i.e., error-prone storage). Additional storage in a given memory trace results in the storage of more features (i.e., replacing the remaining zeros) not the correction of previously stored feature values.

During retrieval, features of test item j are compared to each trace stored in memory, indexed by i, and a likelihood ratio is computed as follows,

$$\lambda_{(i,j,k)} = (1-c)^{nq_{(i,j,k)}} \prod_{v=1}^{\infty} \left[\frac{c + (1-c)g(1-g)^{v-1}}{g(1-g)^{v-1}} \right]^{nm(v,i,j,k)} \qquad (1)$$

This is the REM equation for the subjective likelihood that the item features of test stimulus j matches memory trace i for simulated subject k. The number of non-zero features that mismatch is nq and the number of non-zero features that match and have the value v is nm. Features with a value of zero do not contribute to the decision because zero indicates a lack of information. The decision about whether test stimulus j was studied or not is based on the subjective memory strength, defined as the average of the likelihood ratios (the odds). If the average is greater than some criterion (*criterion*), test stimulus j is called "studied" otherwise it is called "new."

Because the focus of this paper is item vs. context-noise, we implement a pure item-noise version of REM. This choice is not meant to suggest that context has no role in episodic memory. In fact, we have extensively discussed and modeled the role of context in REM (e.g., Criss & Shiffrin, 2004; Klein, Shiffrin, & Criss, 2007; Malmberg & Shiffrin, 2005). In a pure item-noise implementation of REM, context remains constant across the study list and perfectly isolates the study list for comparison during the memory test, thus context plays no role in any changes in performance as a function of study or test block. To implement learning during test, we make the following assumptions. If an item is judged to be studied, then the memory trace that best matches that test item, following Eq. (1), is updated. If an item is judged to be new, then a new memory trace is stored. In a single-item recognition test (e.g., Experiment 1), this means that a single memory trace is updated or stored on each trial. In a 4AFC test (e.g., Experiment 2), the model decides that the test item with the highest odds value is the target. Therefore, three traces are stored and one is updated on every trial.

The assumption that the best matching memory trace is updated produces errors in storage. Sometimes a second memory trace for a target item will be stored (e.g., on a miss trial) and sometimes an incorrect memory trace will be updated. All test items that evoke false alarms will result in storage of features in an incorrect memory trace. Even when a target is correctly identified, very infrequently the best matching trace for that target will not correspond to the trace stored during study of that target due to noise. The relative amount of updating vs. storing new traces depends on memory accuracy and response bias. The more accurate memory, the more often tests of target items will trigger updating of their study memory trace and the more often foils will result in storage of a new memory trace. The lower the bias to respond ”old”, the more often new memory traces will be stored at test. Both updating memory traces and storing new traces results in a decrease in HR across test block. However, two mechanisms have different effects on the FAR. Adding new traces increases the size of the memory set. Each additional trace stored in memory is likely to match a foil by chance, increasing the FAR. Updating memory traces results in differentiation. The more features stored in a memory trace, the less likely it is to match an unrelated item, decreasing the FAR. The predicted pattern of FAR is a balance between an increase due to storing new traces and a decrease due to updating traces (from both foil and target test trials).

We note there are no new parameters in this model; the modeling of encoding during test is directly bound to the design of the relevant experiment. All parameter values identified above were held constant at standard values except the u^* parameter ($u^* = .16$ for the experiments presented here) and the criterion for single-item recognition which were fit to the empirical data. Usually the REM model produces good fits using the optimal criterion of 1.0. However, the FAR in Experiment 1 was high enough to warrant a more liberal criterion of 0.72. In Experiment 2 there is no criterion because it is a forced choice test where the odds for each test item are compared to one another (rather than a criterion) and the item with the highest odds is chosen as the target. Model fits are presented in Fig. 1 and panel D of Fig. 2. The model fits very well overall, especially given that identical parameter values were used for both experiments, an exhaustive search of the parameter space was not conducted, and the fits are to average data, not taking into account variability across individuals. There are two discrepancies between the model predictions and the data. First, the model does not capture the criterion adjustment taking place in blocks 1–3 of Experiment 1 (Fig. 1). We could have easily implemented a criterion shift and described the data well. However, as mentioned earlier a mechanism for dynamic criterion adjustment is outside the scope of all extant memory models. Rather than implement an ad hoc mechanism to describe the current data, we prefer to fully explore the phenomenon and develop a mechanism that more fully describes the complexities of dynamic criterion setting in recognition memory. One possibility is to build up expected target and foil distributions over the course of the first several trials and use the resulting expected value of the odds as the criterion (cf., Brown, 2010; Turner, Van Zandt & Brown, submitted for publication). The second discrepancy is that the model does not predict the relative sparing of accuracy for items tested in block 1 compared to subsequent test blocks (Fig. 2). As noted earlier, there are many reports that memory is helped more by a test trial than by a study trial (e.g., Roediger & Karpicke, 2006a, 2006b), known as the testing effect. Implementing the testing effect in the model by allowing a higher u^* parameter for test than study trials would serve to increase the overall slope of output interference and provide a better fit to the data. As noted earlier, we implemented a pure item-noise model to illustrate that such models can produce output interference similar to empirical observations. Including context-noise in the model, which surely exists, is necessary to predict for the decline in performance in the 20 min. delay condition and this may also provide a better fit to the data.

There are of course many possible ways to implement learning during test other than the mechanism we implemented here. For example, adding a new memory trace containing both context and item features on every trial (e.g., Ratcliff & Hockley, 1980), updating memory traces only when the test item is recollected, storing a new memory trace and updating the trace generated during the study list, among other possibilities. The important point illustrated here is that REM framework nicely accounts for the observed data with the simple assumption of encoding item information into episodic memory during test.

Conclusion

The present findings suggest that output interference effects in recognition memory are reliable and robust. We believe the best explanation for output interference in recognition memory requires item-noise, or the additional noise that is due to adding items to episodic memory. Certainly the present demonstrations of output interference effects, in combination with other findings that item-noise models account for and context-noise models do not (e.g., Criss, 2006; Criss & Shiffrin, 2004; Malmberg & Murnane, 2002), suggest to us that additional scrutiny should be given to the null list-length effect. We have emphasized item-noise and present a model that relies solely on item-noise to account for output interference. Our singular focus on item-noise was intended to make the point that item-noise is critical. However, we are not proposing that context-noise plays no role in recognition memory. To the contrary, we believe both item-noise and context-noise are necessary to understand episodic memory (cf., Criss & Shiffrin, 2004). For example, Klein et al. (2007) presented evidence that context changes relatively little during list study but changes substantially after the presence of a test, suggesting that the contextual information in the recognition test probe is likely more similar to the context stored in the traces of the prior test words than the traces of the earlier study list words. A context change mechanism like this could help explain why the present results are similar whether testing is immediate or delayed by 20 min in the current data. Despite recent attempts to demonstrate no role for item-noise in recognition memory, the current results suggest that the relative role of item- and context-noise merits further consideration.

Acknowledgments

Correspondence regarding this article may be directed to Amy H. Criss (acriss@syr.edu) or Kenneth J. Malmberg (malmberg@usf.edu). AHC and KJM contributed equally to this manuscript. We thank Nick Fischetti, Arielle Port, and Brianna Girton for help collecting data. This research was funded by National Science Foundation grant 0951612 to AHC and Air Force Office Scientific Research grant AFOSR-BAA-2010-1 to RMS.

References

Anderson, M. C., & Neely, J. H. (1996). Interference and inhibition in memory retrieval. In E. Bjork & R. A. Bjork (Eds.), *Memory* (pp. 237–313). San Diego, CA, USA: Academic Press.

Atkinson, R. C., & Juola, J. F. (1974). Search and decision processes in recognition memory. In D. Krantz, R. Atkinson, R. Luce, & P. Suppes (Eds.). *Contemporary developments in mathematical psychology* (Vol. 1). San Francisco: W.H. Freeman.

Balota, D.A., Cortese, M.J., Hutchison, K.A., Neely, J.H., Nelson, D., Simpson, G.B., & Treiman, R. (2002). The English Lexicon Project: A web-based repository of descriptive and behavioral measures for 40,481 English words and nonwords. <http://elexicon.wustl.edu/>, Washington University.

Brown, S.D. (2010). The pervasive problem of criterion setting. Presented at the 51st Annual meeting of the Psychonomic Society.

Brown, S. D., & Steyvers, M. (2005). The dynamics of experimentally induced criterion shifts. *Journal of Experimental Psychology: Learning, Memory and Cognition, 31*(4), 587–599.

Brown, S. D., Steyvers, M., & Hemmer, P. (2007). Modeling experimentally induced strategy shifts. *Psychological Science, 18*, 40–45.

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory and Cognition, 20*, 632–642.

Cary, M., & Reder, L. M. (2003). A dual-process account of the list-length and strength-based mirror effects in recognition. *Memory and Cognition, 49*, 231–248.

Clark, S., & Gronlund, S. (1996). Global matching models of recognition memory: How the models match the data. *Psychonomic Bulletin and 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. (2009). The distribution of subjective memory strength: Foils and response bias. *Cognitive Psychology, 59*, 297–319.

Criss, A. (2010). Differentiation and response bias in episodic memory: Evidence from reaction time distributions. *Journal of Experimental Psychology: Learning, Memory and Cognition, 36*(2), 484–499.

Criss, A., & Malmberg, K. (2008). Evidence in favor of the early-phase elevated-attention hypothesis: The effects of letter frequency and object frequency. *Journal of Memory and Language, 59*(3), 331–345.

Criss, A. H., & Shiffrin, R. M. (2004). Context-noise and item-noise jointly determine recognition memory: A comment on Dennis & Humphreys (2001). *Psychological Review, 111*(3), 800–807.

Crowder, R. (1976). *Principles of learning and memory*. Oxford England: Lawrence Erlbaum.

Dalezman, J. (1976). Effects of output order on immediate, delayed, and final recall performance. *Journal of Experimental Psychology: Human Learning and Memory, 2*(5), 597–608.

Dennis, S., & Chapman, A. (2010). The Inverse List Length Effect: A Challenge for Pure Exemplar Models of Recognition Memory. *Journal of Memory and Language, 63*(3), 416–424. doi:10.1016/j.jml.2010.06.001.

Dennis, S., & Humphreys, M. S. (2001). The role of context in episodic recognition: The bind cue decide model of episodic memory. *Psychological Review, 108*, 452–478.

Dennis, Simon., Lee, Michael. D., & Kinnell, Angela. (2008). Bayesian analysis of recognition memory: The case of the list-length effect. *Journal of Memory and Language, 59*(3), 361–376.

Dong, T. (1972). Cued partial recall of categorized words. *Journal of Experimental Psychology, 93*(1), 123–129.

Dorfman, D., & Glanzer, M. (1988). List composition effects in lexical decision and recognition memory. *Journal of Memory and Language, 27*(6), 633–648.

Estes, W. (1955a). Statistical theory of distributional phenomena in learning. *Psychological Review, 62*(5), 369–377.

Estes, W. (1955b). Statistical theory of spontaneous recovery and regression. *Psychological Review, 62*(3), 145–154.

Freeman, E., Heathcote, A., Chalmers, K., & Hockley, W. (2010). Item-effects in recognition memory for words. *Journal of Memory and Language, 62*(1), 1–18.

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

Gronlund, S., & Elam, L. (1994). List-length effect: Recognition accuracy and variance of underlying distributions. *Journal of Experimental Psychology: Learning, Memory and Cognition, 20*(6), 1355–1369.

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology, 46*(3), 269–299. doi:10.1006/jmps.2001.1388.

Humphreys, M., Pike, R., Bain, J., & Tehan, G. (1989). Global matching: A comparison of the SAM, Minerva II, Matrix, and TODAM models. *Journal of Mathematical Psychology, 33*(1), 36–67.

Jacoby, L. L., Shimizu, Y., Velanova, K., & Rhodes, M. G. (2005). Age differences in depth of retrieval: Memory for foils. *Journal of Memory and Language, 52*(4), 493–504. doi:10.1016/j.jml.2005.01.007.

Kinnell, A., & Dennis, S. (2011). The list length effect in recognition memory: An analysis of potential confounds. *Memory and Cognition, 39*(2), 348–363.

Klein, K., Shiffrin, R. M., & Criss, A. H. (2007). Putting context into context. In J. S. Nairne (Ed.), *The foundations of remembering: Essays in honor of Henry L. Roediger, III* (pp. 171–190). Psychology Press.

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

Maguire, A. M., Humphreys, M. S., Dennis, S., & Lee, M. D. (2010). Global similarity accounts of embedded-category designs: Tests of the Global Matching Models. *Journal of Memory and Language, 63*(2), 131–148.

Malmberg, K. J. (2008). Recognition memory: A review of the critical findings and an integrated theory for relating them. *Cognitive Psychology, 57*, 335–384.

Malmberg, K. J., Holden, J. E., & Shiffrin, R. M. (2004). Modeling the effects of repetitions, similarity, and normative word frequency on judgments of frequency and recognition memory. *Journal of Experimental Psychology: Learning, Memory and Cognition, 30*(2), 319–331.

Malmberg, K., & 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., & Shiffrin, R. M. (2005). The "one-shot" hypothesis for context storage. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(2), 322–336. doi:10.1037/0278–7393.31.2.322.

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

Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review, 87*, 252–271.

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review, 105*(4), 734–760.

McGeoch, J. A. (1933). Studies in Retroactive Inhibition: I. The Temporal Course of the Inhibitory Effects of Interpolated Learning. *Journal of General Psychology, 9*, 24–42.

Melton, A. W., & von Lackum, W. J. (1941). Retroactive and proactive inhibition in retention: Evidence for a two-factor theory of retroactive inhibition. *American Journal of Psychology, 54*, 157–173.

Mensink, G., & Raaijmakers, J. (1988). A model for interference and forgetting. *Psychological Review, 95*(4), 434–455.

Murdock, B. (1974). *Human memory: Theory and data*. Oxford England: Lawrence Erlbaum.

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

Murdock, Bennet B., & Anderson, Rita E. (1975). Encoding, storage and retrieval of item information. In R. L. Solso (Ed.), *Theories in cognitive psychology: The Loyola symposium*. Hillsdale, N.J.: Erlbaum.

Neely, J. H., & Tse, C. (2009). Category length produces an inverted-U discriminability function in episodic recognition memory. *The Quarterly Journal of Experimental Psychology, 62*(6), 1141–1172. doi:10.1080/17470210802373035.

Nobel, P. A., & Shiffrin, R. M. (2001). Retrieval processes in recognition and cued recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(2), 384–413. doi:10.1037/0278-7393.27.2.384.

Norman, D. A., & Waugh, N. C. (1968). Stimulus and response interference in recognition memory experiments. *Journal of Experimental Psychology, 78*, SS1–S59.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1980). SAM: A theory of probabilistic search of associative memory. In G. H. Bower (Ed.). *The psychology of learning and motivation* (Vol. 14, pp. 207–262). New York: Academic Press.

Raaijmakers, J., & Shiffrin, R. (1981). Search of associative memory. *Psychological Review, 88*(2), 93–134.

Ratcliff, R., & Hockley, W. E. (1980). Repeated negatives in item recognition: Nonmonotonic lag functions. In R. S. Nickerson (Ed.), *Attention and performance VIII*. Hillsdale, NJ: Erlbaum.

Roediger, H. (1974). Inhibiting effects of recall. *Memory and Cognition, 2*(2), 261–269.

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

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

Roediger, H., & Schmidt, S. (1980). Output interference in the recall of categorized and paired-associate lists. *Journal of Experimental Psychology: Human Learning and Memory, 6*(1), 91–105.

Schulman, A. I. (1974). The declining course of recognition memory. *Memory and Cognition, 1.4*, 14–18.

Shiffrin, R. (1970). Forgetting: Trace erosion or retrieval failure? *Science, 168*(3939), 1601–1603.

Shiffrin, R., Huber, D., & 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.

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

Smith, A., D'Agostino, P., & Reid, L. (1970). Output interference in long-term memory. *Canadian Journal of Psychology, 24*(2), 85–89.

Starns, J., White, C., & Ratcliff, R. (2010). A direct test of the differentiation mechanism: REM, BCDMEM, and the strength-based mirror effect in recognition memory. *Journal of Memory and Language, 63*(1), 18–34.

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

Strong, Edward K. Jr., (1912). The effect of length of series upon recognition memory. *Psychological Review, 19*(6), 447–462.

Tulving, E., & Arbuckle, T. (1966). Input and output interference in short-term associative memory. *Journal of Experimental Psychology, 72*(1), 145–150.

Turner, B., Van Zandt, T., & Brown, S. (submitted for publication). A dynamic stimulus-driven model of signal detection.

Xu, J., & Malmberg, K. (2007). Modeling the effects of verbal- and nonverbal-pair strength on associative recognition. *Memory and Cognition., 35*(3), 526–544.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46*, 441–517.

Zechmeister, E. (1972). Orthographic distinctiveness as a variable in word recognition. *The American Journal of Psychology, 85*(3), 425–430.