Journal of Experimental Psychology: Learning, Memory, and Cognition

Sources of Interference in Recognition Testing

Jeffrey Annis, Kenneth J. Malmberg, Amy H. Criss, and Richard M. Shiffrin

Online First Publication, April 8, 2013. doi: 10.1037/a0032188

Journal of Experimental Psychology:
Learning, Memory, and Cognition

0278-7393/13/$12.00 DOI: 10.1037/a0032188

Sources of Interference in Recognition Testing

Jeffrey Annis and Kenneth J. Malmberg
University of South Florida

Amy H. Criss
Syracuse University

Richard M. Shiffrin
Indiana University

Recognition memory accuracy is harmed by prior testing (a.k.a., output interference [OI]; Tulving & Arbuckle, 1966). In several experiments, we interpolated various tasks between recognition test trials. The stimuli and the tasks were more similar (lexical decision [LD] of words and nonwords) or less similar (gender identification of male and female faces) to the stimuli and task used in recognition testing. Not only did the similarity between the interpolated and recognition tasks not affect recognition accuracy but performance of the interpolated task caused no interference in subsequent recognition testing. Only the addition of recognition trials caused OI. When we presented a block of LD trials or gender identification trials before the recognition test, a decrease in accuracy was observed in the subsequent recognition tests. These results suggest a distinction between temporal context and task context, such that recognition memory performance is determined by the salience of the context cues, and the use of temporal context cues is associated with OI.

Keywords: recognition, output interference, context

Output interference (OI; Tulving & Arbuckle, 1966), the increase in forgetting of events caused by the testing of memory for other events, has been reported for paired-associate recall (Tulving & Arbuckle, 1966; Wickens, 1970; Wickens, Born, & Allen, 1963) and probed recall experiments (e.g., Raaijmakers & Shiffrin, 1981; Roediger & Schmidt, 1980; Smith, 1971). For example, Smith's (1971) participants studied items from semantically distinct categories. When sequentially prompted with each category cue to recall as many items from a category as possible, performance decreased with each successive category tested.

OI has also been observed in recognition memory tasks, in which memory is tested with items that were recently encountered and items that were not (Criss, Malmberg, & Shiffrin, 2011; Criss & Shiffrin, 2004; Malmberg, Criss, Gangwani, & Shiffrin, 2012; Murdock & Anderson, 1975; Norman & Waugh, 1968; Schulman, 1974;Smith, 1971; Smith, D'Agostino, & Reid, 1970). OI is a robust form of forgetting, but OI in recognition memory testing is especially intriguing because it is one part of a newly discovered paradox. Increasing the number of words studied, controlling for a variety of potentially confounding variables, has only a small effect on recognition accuracy (Cary & Reder, 2003; Dennis & Humphreys, 2001), whereas increasing the number of items tested causes relatively large amounts of forgetting. Why does testing recognition memory cause substantial interference, but storing new traces during study cause little interference?

Interference Theory and Recognition Memory

Memory models assume that information upon which a recognition decision is made comes from memory traces consisting of *item information* and *context information*. Item information is the representation of the properties of the to-be-remembered items. For example, item information might consist of the meaning, orthographic, or phonological properties of a word or the visual or semantic content of a photo. Context information is often described as all those properties that are associated with, but not inherent to, the item itself (for a discussion, see Malmberg & Shiffrin, 2005, or McGeoch, 1942). For example, context information might include environmental variables such as the type of room in which the item was observed, the temporal setting, the orienting task, or internal subject variables such as mood (e.g., Murnane, Phelps, & Malmberg, 1999). Accordingly, when a random set of items is studied in succession, item information is uncorrelated between memory traces, whereas context information is correlated.

The distinction between item and context information is sometimes ambiguous. For example, an item (say A) may be rehearsed and coded with other items (say B and C). It is not clear whether to call information about B and C context, appropriate if such information is stored in the same memory trace as A, or item information, appropriate if such information is stored in separate memory traces. Most likely such information should best be described as both item and context. For instance, in a recent model the current item, C, is associated with other items with which it

Jeffrey Annis and Kenneth J. Malmberg, Department of Psychology, University of South Florida; Amy H. Criss, Department of Psychology, Syracuse University; Richard M. Shiffrin, Department of Psychology, Indiana University.

This work was supported by National Science Foundation Grant 0951612 to Amy H. Criss.

Correspondence concerning this article should be addressed to Jeffrey Annis, Department of Psychology, PCD 4118G, University of South Florida, Tampa, FL 33620. E-mail: jannis@mail.usf.edu

was rehearsed (A, B) by virtue of all three items along with a representation of the temporal context comprising a single episodic trace, but the same information will likely be included in several traces stored over the course of studying a list of words (Lehman & Malmberg, 2009, in press; also Howard & Kahana, 2002; Raaijmakers & Shiffrin, 1981). Thus, there may be some overlap between information contained in the traces representing context and information representing content, but because of the ambiguity associated with classifying such information, the following discussion will exclude consideration of co-rehearsed items or context defined by item-to-item associations, unless clearly stated otherwise. We use item information to refer to properties of individual items and context information to refer to properties shared across multiple items.

When memory is probed, the similarity of the contents of the target trace and other nontarget traces is a primary factor determining retrieval success. As the number of similar traces in memory increases, the greater the noise present in the retrieval process. The noise can arise from the similarity of the items that were studied or the number of items stored in similar contexts. When item or context information interferes with recognition memory, the sources of interference are referred to as *item noise* and *context noise*, respectively, and different memory models make specific assumptions about the roles of item and context noise in recognition (e.g., Dennis & Humphreys, 2001; Shiffrin & Steyvers, 1997; see Criss & Shiffrin, 2004).

Context noise models, such as BCDMEM (Dennis & Humphreys, 2001), assume that interference occurs only because of the similarity between the current context and other contexts in which the item occurred. When an item is presented for a recognition test, a representation of the item activates all of the contexts in which that item has been encountered. The composite representation of the past contexts is compared to a representation of the test context, and the more similar the composite representation of the past contexts in which an item has been encountered is to the test context, the more likely it will be positively endorsed. Since the match between the test item and other item information stored in memory does not inform the recognition decision, BCDMEM predicts that recognition accuracy will be unaffected by the number of items studied or tested, or the similarity of the items studied or tested. All forgetting is due to contextual confusion according to BCDMEM. Thus, BCDMEM predicts no effect of adding unrelated items to the study list (e.g., a null-list length effect) or the test list and therefore does not predict OI a priori (Criss et al., 2011; Malmberg et al., 2012). For the same reasons, BCDMEM would have to resort to post hoc explanations to account for why other factors related to the nature of the study or test items affects recognition accuracy (Criss & Malmberg, 2008; Criss & Shiffrin, 2004; Malmberg & Murnane, 2002; Malmberg, Steyvers, Stephens, & Shiffrin, 2002; see one example in Dennis & Chapman, 2010).

Item noise models, like REM (Criss & Shiffrin, 2004; Shiffrin & Steyvers, 1997; also Anderson & Bower, 1973; McClelland & Chappell, 1998; etc.), are really models that assume both sources of noise but for clarity have focused on the item noise component. They assume that noise derives from the match between the item information in the retrieval cue and all stored traces. Such traces include the study-list trace of the test item (if one exists), traces of other items than the test item (from the study list or otherwise), and traces of the test item that were stored but not on the basis of list-study. Thus, both context noise and item noise are both responsible for interference and forgetting. REM assumes that episodic traces consist of noisy and error prone representations of to-be-remembered items and the contexts in which they occur. When an item is presented for recognition testing, a context cue is used to identify a set of traces likely to have been stored in the test context; the more similar the context cue is to the context stored in a given trace, the more likely the trace was stored in the context in question (Lehman & Malmberg, 2009; Malmberg & Shiffrin, 2005; Shiffrin & Steyvers, 1997, 1998). This activated set of traces is then compared to the item information in the retrieval cue. The more similar the item cue is to the contents of the activated set, the more familiar the item seems. If the familiarity of the test item exceeds a subjective decision criterion, the item is endorsed. Targets are more likely than foils to be endorsed because they are more likely to be in the activated set due to matching context features and therefore at least one trace in the activated set will tend to match the item cue. Since the familiarity of each test item is based not only on the match of the item cue to its own memory trace but also to the match of the remainder of the activated set, the average match or familiarity value decreases as the number of items in the activated set increases. Hence, REM predicts a small and negatively accelerated decrease in recognition accuracy as the number of studied items increases (Criss & McClelland, 2006). In addition, REM predicts that recognition accuracy will at times be affected by the structure of the study list and factors related to the representation of the items themselves, such as orthographic distinctiveness or semantic similarity.

A pure item-noise version of REM was applied to OI (Criss et al., 2011) by assuming that items from each test trial were stored in memory. Specifically, when a test item is judged to be new, a new trace is stored, and when a test item is judged to be old, the best matching trace in the activated set is updated. The context at study and test was assumed to be sufficiently similar such that context information was the same for those traces stored during the original study list and those stored or updated during the test list (and thus context-noise did not contribute to the predicted OI). Subsequent test items activate both the traces stored during study and those stored during test. Because test items do not repeat, the additional information stored during the test only adds noise to the signal for subsequent test trials, decreasing accuracy. Further, the decrease in accuracy is greater when the new traces stored in memory contain item information that is similar to the items tested on subsequent test trials. Accordingly, if the traces stored as the result of recognition testing are relatively dissimilar to subsequent test items then recognition performance should be harmed less. For instance, in Malmberg et al. (2012) subjects studied words from two different categories. When items from both categories were randomly intermixed at test, a typical pattern of OI was observed, but when the test was blocked by category, a release from interference was observed at the category switch. This model provided a good fit to the Criss et al. (2011) data and was consistent with the Malmberg et al. (2012) data. However, an alternative model where the test trials contain context information that differs from the study list is plausible. We explore this possibility by manipulating the task relevant components of context while holding the temporal components of context constant (Experiments 1 and 2) and while varying the temporal aspects of context (Experiment 3).

Experiment 1

In Experiment 1, we extend prior findings and provide constraints for theory by testing the assumption that item noise is the sole cause of OI in recognition testing. We do this by mixing recognition memory test trials, in the form of two-alternative forced-choice (2AFC), with stimuli of different types. If OI is due to item-noise alone, then the interpolated tasks containing items that are similar to the test items should result in more OI than the tasks containing items that are less similar to the items used in recognition testing.

In the control condition, a blank screen was presented on a computer monitor for 1 s between each recognition test. In the lexical decision (LD) condition, an LD task was interpolated trial-by-trial with the recognition tests. In the gender identification (gender ID) condition, a photo of a novel face was presented, and the subject determined the gender of the face. The item information in the LD task is quite similar (e.g., word or word-like letter strings) to the 2AFC, but the task context is quite different. If item information is the dominant factor contributing to OI, then we should observe more OI in the LD condition than in the control condition. In contrast, the stimuli in the gender ID trials and 2AFC trials are quite different and furthermore, faces and words do not interfere in a recognition task (Criss, 2004). Therefore, we predicted that the gender ID trials would produce less OI than the control condition. To preview, we find that LD trials did not affect recognition accuracy. This result suggests that task context plays an important role in storage and at test and, in our view, causes recognition probes (containing recognition task context) to be dissimilar from LD test traces (containing LD context).

Method

Participants. One hundred thirty-four undergraduate students at the University of South Florida participated in exchange for course credit.

Design and materials. Participants received one study-test block of each intertrial task condition: lexical decision (LD), gender identification, or a blank interval. Each study list was composed of 160 nouns from the Kucera and Francis (1983) word pool with normative frequencies between 20 and 50 occurrences per million. The test list was composed of 160 words from the study list and 160 foils. One of three intertrial tasks was assigned to each test list. Therefore, the intertrial task was manipulated within subject and between lists. The LD trials contained 80 words drawn randomly and anew from the Kucera and Francis (1983) word pool and 79 nonwords. The words in the LD task were different from those used in the study/test trials. For each of the 159 gender ID trials, either a male or female face was presented with approximately equal probability. Faces were drawn randomly from the black and white Facial Recognition Technology (FERET) database (Phillips, Wechsler, Huang, Rauss, 1998).

Procedure. The participants were informed of the nature of the memory experiment and given instructions. Immediately following the instructions, participants were presented with a study list in which words were presented for 1.0 s each. There was a 0.1-s interstimulus interval (ISI) following each studied word. After all of the words from the study list were presented, a math task was performed that involved adding successive integers from 1 to 9, each presented for 3 s over a span of 30 s. Following the math task, the test list was presented.

For each 2AFC recognition test trial, participants were presented with a target and foil that appeared side by side, with the target location randomly selected on each trial. The task was to indicate on which side the target was presented. If the target was presented on the left side of the screen, the participant was to respond with a "1." If the target word appeared on the right, they were to respond by typing a "0."

Following each 2AFC test trial, there was an intertrial interval. For the lexical decision condition, a letter string was presented at the center of the screen during the intertrial interval. The task of the participant was to respond "1" if the letter string was an English word and "0" if the letter string was a nonword. In the gender ID condition, a male face or female face was presented at the center of the screen, and the task was to respond "1" if the face was a female face and "0" if it was a male face. For the control condition (also referred to as the "no task" condition), no task was specified during the unfilled 1.0-s ISI.

Results and Discussion

Accuracy and output position. There was no significant main effect of intertrial task on recognition accuracy ($F < 1$; *No Task*: $M = .62$, $SD = .01$; *LD*: $M = .62$, $SD = .01$; *Face*: $M = .62$, $SD = .01$). The left panel of Figure 1 shows a significant main effect of test position on accuracy, $F(9, 1197) = 19.00$, $p < .0005$, where the horizontal axis is binned by the number of 2AFC recognition test trials only. Accuracy decreased with increases in the number of items tested via recognition. There was no significant Task $\times$ Test Position interaction ($F < 1$), so the interpolated task events were not adding to the magnitude of the OI. The right panel of Figure 1 shows recognition accuracy where the binning included all the test events. This plot shows how recognition accuracy is affected by the number of all events, regardless of the interpolated task. For the lexical decision task and face task, one test event bin contained 32 test events (16 2AFC and 16 LD or 16 2AFC and 16 face trials). For the No Task condition, one test event bin was also equal to 32 test events, but all of them were 2AFC trials. Therefore, the number of bins for the No Task condition was half that of the other conditions and consequently the analysis of variance (ANOVA) was conducted using only the first five test event bins of each condition. There was not a significant main effect of intertrial task on accuracy, $F(2, 266) = 2.76$, $p = .065$; if anything recognition accuracy was the worst for the No Task condition. There was a significant main effect of binned test event position, $F(4, 532) = 18.68$, $p < .0005$, but no significant intertrial Task $\times$ Test Position interaction, $F(8, 1064) = 1.35$, $p = .216$. Thus, OI increased with increases in test position; however, forgetting was not affected by including an interpolated task, nor the nature of the intertrial task.

The results of this experiment indicate that the amount of OI observed was restricted to the recognition testing, but not the interpolated task. Therefore, the similarity of the item information used to test recognition memory to the item information used to perform the interpolated tasks had no effect on the accuracy of recognition memory. These results are inconsistent with the simple item-noise models that assume that the storage of episodic information during the course of testing causes OI

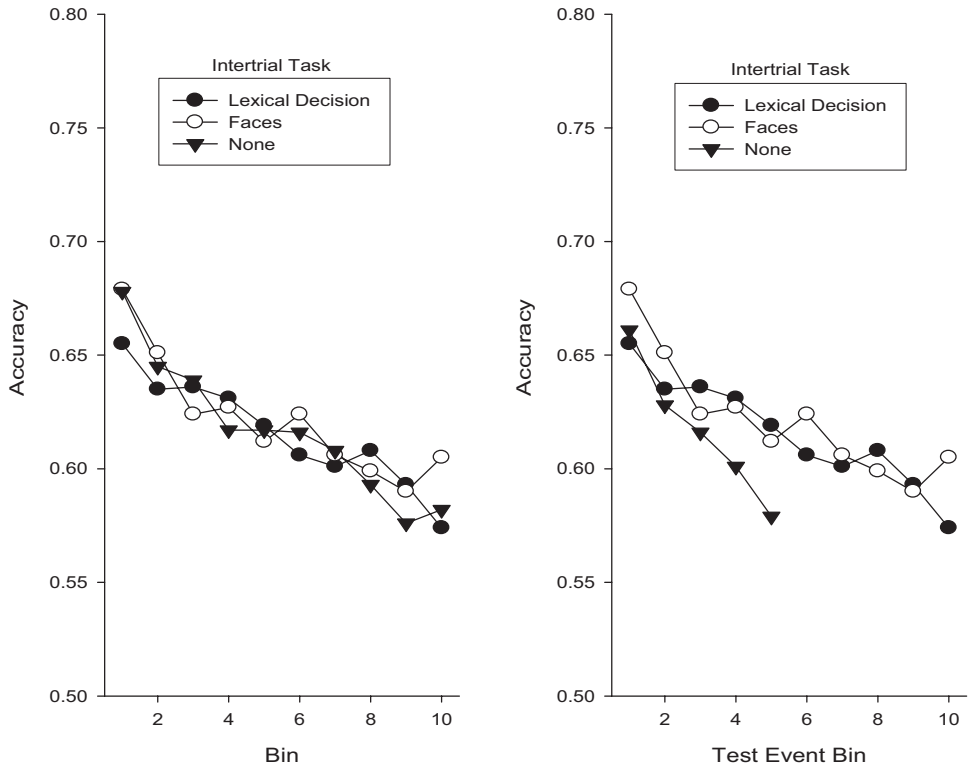

Figure 1. Results of Experiment 1. The left panel plots two-alternative forced-choice (2AFC) word-recognition accuracy as a function of test position and intertrial task. Bin size equals 16 2AFC word-recognition trials. The right panel plots 2AFC accuracy as a function of the number of events (2AFC word-recognition plus the intertrial tasks). Bin size equals 32 events.

under all circumstances. However, this prediction is based on the tacit assumption that the encoding of the new episodic information during the performance of the interpolated task resulted in new traces sufficiently strong to produce interference. If the performance of the intertrial tasks produced only very weak episodic traces, then the similarity of the items used to perform the intertrial tasks would only have a negligible effect. Therefore, the goal of Experiment 2 was to assess the strength of the traces created during the performance of the LD trials.

Experiment 2

This experiment was identical to Experiment 1 with the exception that a third phase was added to the end of the critical LD condition. Participants were given a 2AFC recognition task following a study list. Between each test trial, they received a gender ID task, a LD task, or a blank interval. Following the list containing the intertrial LD task, participants received an additional 2AFC recognition memory task in which the targets were the words used during the LD trials. If the reason for the null effect of LD trials on OI in Experiment 1 was because participants were not storing traces representing the LD trials or storing poor quality information about those items, then we should observe poor memory for those words on the subsequent recognition test in Phase III.

Method

Participants. Fifty students at the University of South Florida participated in exchange for course credit.

Design, materials, and procedure. This experiment was exactly the same as Experiment 1 with one exception. Following the test list in which the LD task was the intertrial task, participants performed a recognition test in which their memory for the English words in the LD task was tested. Participants were informed both at the beginning of the experiment and directly before this test list that their memory for the words on the LD test trials would be tested. In the Phase III test, participants were presented with a word from the LD task and a foil on opposite sides of the screen (order randomly selected). The task of the participant was to respond with a "1" if the word that was presented during the LD task was on the left side of the screen. They were to respond "0" if the target appeared on the right. The 80 test trials in Phase III were separated by a blank screen for 1.0 s with no task required during the ISI.

Results and Discussion

There was no significant main effect of intertrial task ($F < 1$; No Task: $M = .62$, $SD = .01$; LD: $M = .61$, $SD = .01$; Face: $M = .62$, $SD = .02$). This replicates the finding from Experiment 1 indicating that the performance of interpolated tasks has no sig-

nificant effect on overall recognition accuracy. The left panel of Figure 2 plots recognition accuracy as a function of test position. There was a significant main effect of test position on accuracy, $F(9, 441) = 10.03$, $p < .0005$. Recognition accuracy decreased as the number of recognition tests increased. There was not a significant main effect of intertrial task on accuracy ($F < 1$), and no intertrial Task × Test Position interaction, $F(18, 882) = 1.16$, $p = .288$. The middle panel of Figure 2 shows accuracy as a function of binned test event and intertrial task. An ANOVA was conducted using only the first five test event bins of each condition. There was a main effect of test event position on accuracy, $F(4, 196) = 6.74$, $p < .0005$. There was no main effect on intertrial task ($F < 1$), and no interaction, $F(8, 392) = 1.76$, $p = .084$. Thus, the results of Experiment 1 were replicated in their entirety.

The startling result of both Experiments 1 and 2 is that the LD trials did not add to the OI observed during recognition testing. To test the hypothesis that the LD items were being insufficiently encoded, participants were given a recognition task for the English words used in the intertrial LD task. The mean accuracy for this task was .61 ($SD = .08$), which is above chance ($p < .0005$) and almost identical to the level of accuracy achieved for the recognition task during the condition with interpolated LD trials, $t(49) = 0.321$, $p = .749$. The right panel of Figure 2 shows that accuracy during Phase III decreased with increases in output position, $F(9, 441) = 3.19$, $p < .05$. This indicates that the recognition tests of the LD stimuli became less accurate as the number of items tested increased, and the magnitude of OI appears similar to that of the other conditions.

From an item noise perspective the finding that OI does not increase when the items used to perform an interpolated task are similar to those used in recognition testing is surprising. Malmberg et al. (2012) recently reported a release from OI when a switch in the semantic categories was made half way through recognition testing. Within the framework of item-noise models, the switch to a new category allowed for an abrupt increase in the mismatch between the item cue used to probe memory and the item information stored in the episodic traces during the course of testing. The present findings are consistent with the explanation put forth by Malmberg et al. (2012), if one assumes the recognition memory probes are sufficiently dissimilar from the LD test traces. Such dissimilarity could be explained if the recognition task context is very important in the memory probe and is quite different from the LD task context that is presumably stored in the LD test traces. In other words, if participants probe memory with task context and item information, then traces stored during recognition will cause interference, but items stored during LD (or gender ID) will not. In contrast, temporal context would not discriminate between recognition versus LD (or gender ID) trials.

Experiment 3

Experiments 1 and 2 interspersed LD or gender ID tests among the recognition tests. Such task switching could have led to a variety of additional strategies. Thus, Experiment 3 blocked the testing sequence. The study phase is followed by Phase II in which participants received a long block of LD trials, a long block of

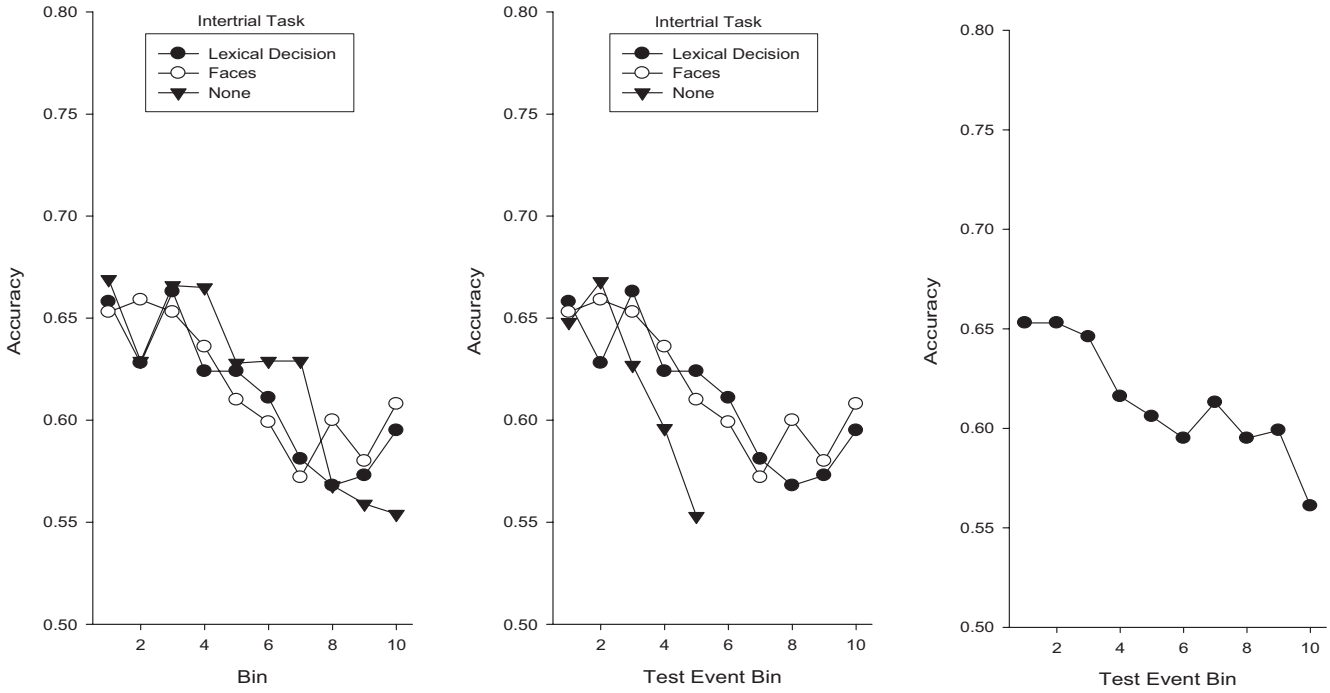

Figure 2. Results of Experiment 2. The left panel plots two-alternative forced-choice (2AFC) word-recognition accuracy as a function of test position and intertrial task. Bin size equals 16 2AFC word-recognition trials. The middle panel plots 2AFC accuracy as a function of the number of events (2AFC word-recognition plus the intertrial tasks). Bin size equals 32 events. The right panel plots 2AFC word-recognition accuracy as a function of test position for the final recognition tests of the word stimuli used in the interpolated lexical decision trials.

gender ID trials, or a long block of 2AFC recognition. We also manipulated the type of stimuli such that participants received either faces or words for study. The task presented in Phase II matched the stimuli studied in Phase I such that those participants receiving LD trials in Phase II were only presented with word stimuli in Phase I while participants in the gender ID condition in Phase II were only presented with face stimuli in Phase I. During Phase III of the experiment, memory for the original study list was tested. The key question is whether the type of task performed during Phase II would affect performance on the recognition task in Phase III. A secondary goal was to measure the strength with which the items in the Phase II test trials, LD and gender ID, were encoded (e.g., insufficiently encoded items may not produce interference). Therefore in Phase IV, recognition memory for Phase II items was assessed.

Method

Participants. Seventy-eight students at the University of South Florida participated in exchange for course credit. Thirty-eight participants were presented with word stimuli, while 40 participants were presented with face stimuli.

Design and materials. The design is illustrated in Figure 3, note that the sets of items were not blocked in the actual experiment and are labeled as such in the figure only to enhance clarity. Each participant was randomly assigned to the word or the face condition. All participants received two rounds of four phases. In one round the participant received 2AFC during Phase II and in the other they received either LD (word condition) or gender ID (face condition) during Phase II, randomly chosen. No stimuli from round one were repeated in round 2. A description of each phase follows.

Word Condition		Face Condition		Both Conditions	
Phase I: Study		Phase I: Study		Phase I: Study	
80 Set A		80 Set A		80 Set A	
80 Set B		80 Set B		80 Set B	
Phase II: LD		Phase II: Gender ID		Phase II: 2AFC	
Word	Nonword			Target	Foil
40 Set A 40 Set C	80 Set D	40 Set A 40 Set C	80 Set D	80 Set A	80 Set C
Phase III: 2AFC		Phase III: 2AFC		Phase III: 2AFC	
Target	Foil	Target	Foil	Target	Foil
80 Set B	80 Set E	80 Set B	80 Set E	80 Set B	80 Set D
Phase IV: 2AFC		Phase IV: 2AFC		Phase IV: 2AFC	
Target	Foil	Target	Foil	Target	Foil
40 Set C	40 Set F	40 Set C	40 Set F	40 Set C	40 Set E

Figure 3. The design of each condition of Experiment 3. During Phase I, subjects were randomly assigned to the word or face condition (between-subjects manipulation) and studied a list of words or a list of faces, respectively. Participants in the word condition completed both the left and the right column, ordered randomly. Participants in the face condition completed both the middle and the right column, ordered randomly. During Phase II, subjects in the face condition performed a two-alternative forced-choice (2AFC) face recognition task following half the Phase I study lists and performed a gender identification (ID) task following the other half of the study lists. During Phase II, subjects in the word condition performed a 2AFC word recognition task following half the Phase I study lists and performed a lexical decision (LD) task following the other half of the study lists. During Phase III, the remaining Phase I stimuli were tested via 2AFC recognition. During Phase IV, items that were used in the gender ID task or the LD task during Phase II were tested via 2AFC recognition. All stimuli from each set were randomly intermixed.

During Phase I, participants studied a list consisting of 160 words or 160 faces (stimulus type manipulated between subjects). Words were randomly drawn from the pool described above. The Phase II task was manipulated within subject, between blocks (LD/gender ID or 2AFC). Participants in the word condition completed 80 2AFC lexical decision trials or 80 2AFC recognition trials of Phase I targets. The LD list in Phase II was composed of 80 nonwords, and 80 English words, where 40 of the English words were drawn from Phase I (since target items were drawn from the study list in the recognition condition) and the other 40 English words were new words drawn from the pool described above. The 2AFC word recognition list in Phase II was composed of 80 words drawn randomly from Phase I and 80 new words. Participants in the face condition completed either 80 gender ID trials or 80 2AFC recognition trials. The gender ID trials consisted of 40 faces drawn from Phase I and 40 new faces drawn from the FERET database. The 2AFC face recognition list in Phase II was composed of 80 faces drawn randomly from Phase I and 80 new faces drawn from the FERET database.

In Phase III, participants completed 80 2AFC recognition trials. The test list in the word condition was composed of 80 words from Phase I, not presented during Phase II, and 80 new words. The test list in the face condition was composed of 80 faces from Phase I, not presented during Phase II, and 80 new faces.

For Phase IV, participants completed 40 2AFC recognition trials for Phase II items. If participants completed a 2AFC task during Phase II, the test list of Phase IV was composed of 40 of the foil items (words in the word condition and faces in the face condition) presented during Phase II and 40 new items. If participants completed a LD or face task during Phase II, the test list of Phase IV was composed of 40 items from Phase II that were not studied during Phase I, and 40 new items. Participants were to indicate the item that previously appeared during the experiment but were not given specific instructions to limit their endorsements to Phase II items.

Procedure. The timing of study, test, and task trials and the response to keyboard mapping were identical to Experiments 1 and 2.

Results and Discussion

Figure 4 plots recognition accuracy as a function of binned test position, stimulus condition (words vs. faces), test phase (II vs. III), and the Phase II task (LD, gender ID, or recognition testing).

Phase II testing. The round in which the subjects were tested interacted with test position, but only in the faces condition, $F(3, 114) = 4.59$, $p < .01$, such that more output interference was observed in the first round of 4 phases than in the second round. There were no other main effects or interactions and subsequent analyses collapse over round. There was no main effect of stimulus on Phase II accuracy ($F < 1$). There was a main effect of test position, $F(3, 228) = 20.25$, $p < .0005$, and the Stimulus $\times$ Test Position interaction was significant, $F(3, 228) = 3.85$, $p < .05$. OI was observed for both word and face recognition, but more OI was observed in the face condition than in the word condition.

Phase III testing. During Phase III, recognition memory for a subset of items appearing during Phase I was tested. These items

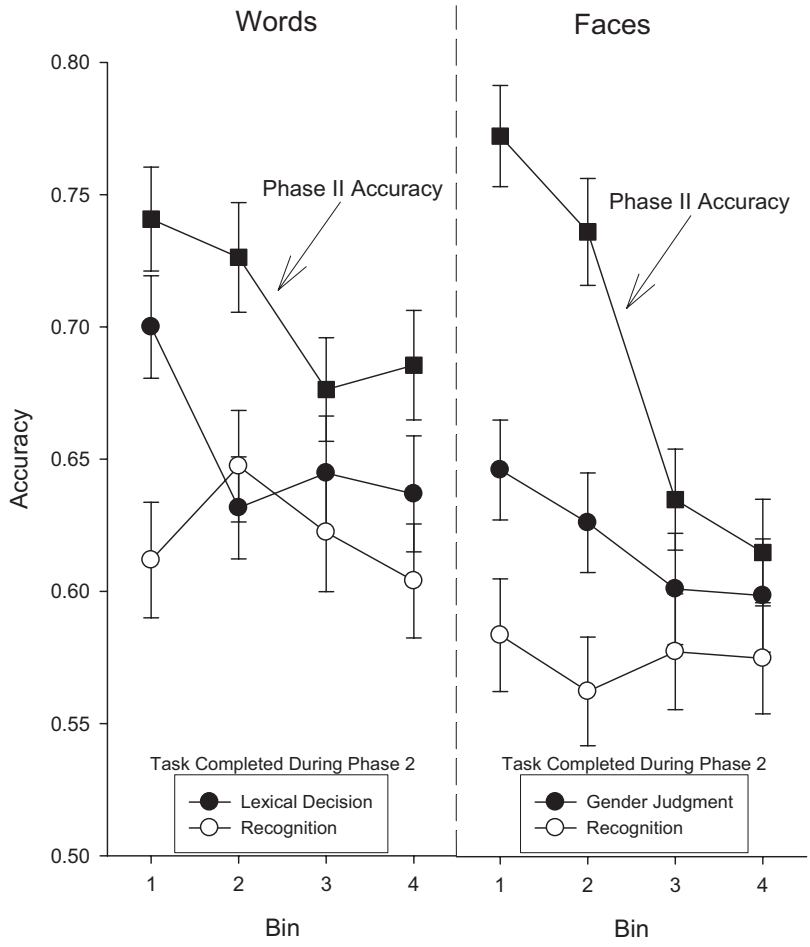

Figure 4. The left panel shows the results of Experiment 3 for the word condition and the right panel shows the results for the faces condition. Both panels plot Phase II accuracy, for those completing recognition in Phase II, as a function test position bin and Phase III accuracy (the lower two curves) as a function of test position bin and the type of task completed during Phase II. The error bars represent the standard error.

were not tested during Phase II. In the faces condition, round interacted with the task completed in Phase II, $F(1, 38) = 5.64$, $p < .05$, such that recognition accuracy decreased more so from Round 1 to Round 2 when the Phase II task was recognition than when it was gender identification. There were no other main effects or interactions with round and further analyses collapse over round. Recognition accuracy was worse for Phase III ($M = .62$, $SD = .08$) than for Phase II ($M = .76$, $SD = .11$) testing, $t(77) = 12.28$, $p < .0005$. This decrease could result from a change in temporal context between Phases II and III, from interference from the stimuli presented in Phase II, or some combination. We note that there is no baseline condition in this experiment (unlike Experiments 1 and 2), and thus, strictly speaking, we cannot evaluate if the cause of lower performance was due to engaging in these tasks or simply the retention interval. However, we note that in prior experiments by Criss et al. (2011) we manipulated the retention interval from 30 s to 20 min, and even after 20 min there was only about a 2–3% decrease in forced-choice accuracy, suggesting that retention intervals of small magnitudes have little effect on 2AFC. Further, the nature of the Phase II task affected performance in Phase III, as described below, and therefore it seems likely that Phase II was the source of forgetting, rather than retention interval.

The relative decrement in accuracy during Phase III for the different tasks in Phase II (LD, Gender ID, or 2AFC) is a critical piece of data because it informs us about the role of task context in interference. The decrease in recognition accuracy from Phase II to Phase III was affected by the nature of the task the subject performed during Phase II. Recognition testing during Phase II produced lower recognition accuracy during Phase II than LD, $F(1, 37) = 4.91$, $p < .05$, or gender ID, $F(1, 39) = 10.27$, $p < .05$. The fact that more interference was caused by recognition testing than by the performance of the LD task or the gender ID task suggests that a task context cue representing the recognition task was used as part of the retrieval cue used to probe memory when recognition testing occurred during Phase III, consistent with the explanation of Experiments 1 and 2.

The main effect of test position during Phase III testing was not reliable, $F(3, 228) = 2.50$, $p = .061$, at conventional levels of significance. The suppressed level of output interference in Phase

III is likely due to a massive amount of interference that seems to be evident even at the start of Phase III testing; performance dropped to near chance, especially for the recognition condition, and had little room to drop further. The Test Position × Stimulus interaction was not significant ($F < 1$). Phase II task did not interact with Phase III test position, $F(3, 228) = 1.83$, $p = .143$, and there was no three-way interaction of Phase II Task × Phase III Test Position × Stimulus, $F(3, 228) = 1.57$, $p = .199$.

Phase IV testing. Phase IV tested memory for items that appeared during Phase II. The key questions for Phase IV is whether the items that were presented during the LD and the gender ID tasks were sufficiently well encoded that they could have caused interference during the test in Phase III. The results of Phase IV are presented in Figure 5. Round did not interact with any other factor; therefore, the following analyses are collapsed across rounds. Accuracy was above chance when the LD task was performed during Phase II ($M = .67$, $SD = .10$), $t(37) = 10.10$, $p < .0005$, and when the task was gender ID ($M = .57$, $SD = .08$), $t(39) = 5.41$, $p < .0005$. There was a main effect of stimulus, $F(1, 76) = 11.90$, $p < .01$. Phase IV word recognition was more accurate than Phase IV face recognition. There was also a main effect of the Phase IV task, $F(1, 76) = 6.91$, $p < .05$, indicating that Phase IV accuracy was worse when recognition was performed during Phase II compared to when either the LD or gender ID was performed, consistent with the pattern of data for Phase III 2AFC. Last, OI was observed, $F(1, 76) = 9.93$, $p < .05$; recognition memory was less accurate during the second half of Phase IV testing than the first half. The Test Position × Stimulus Type interaction was not significant ($F < 1$). The Task × Test Position interaction was not significant, $F(1, 76) = 1.05$, $p = .308$. The Task × Test Position × Stimulus Type interaction was also not significant, $F(1, 76) = 1.41$, $p = .238$.

For further validation of the findings just reported, we conducted separate analysis for each group of participants. For those receiving word stimuli, there was a main effect of task, $F(1, 37) = 18.97$, $p < .0005$. Accuracy was significantly greater for the items used to perform the LD trials during Phase II than the items used to perform the recognition task, suggesting that the lower levels of interference in Phase III arising from LD trials was not due to the LD trials being stored more weakly than the recognition trials. There was no main effect of test position, $F(1, 37) = 1.92$, $p = .174$, and no Task × Test Position interaction ($F < 1$). For participants in the face condition there was no main effect of task ($F < 1$), a main effect of test position, $F(1, 39) = 10.90$, $p < .01$, and no significant interaction, $F(1, 39) = 3.31$, $p = .077$. Again the lower levels of interference during Phase III recognition testing when the gender ID task was performed (compared to 2AFC) during Phase II does not seem to be due to the encoding of weaker face traces.

In Experiments 1 and 2, LD and recognition trials were intermixed but the LD trials did not add to size of the OI effect compared to a control condition with no LD trials. In this exper-

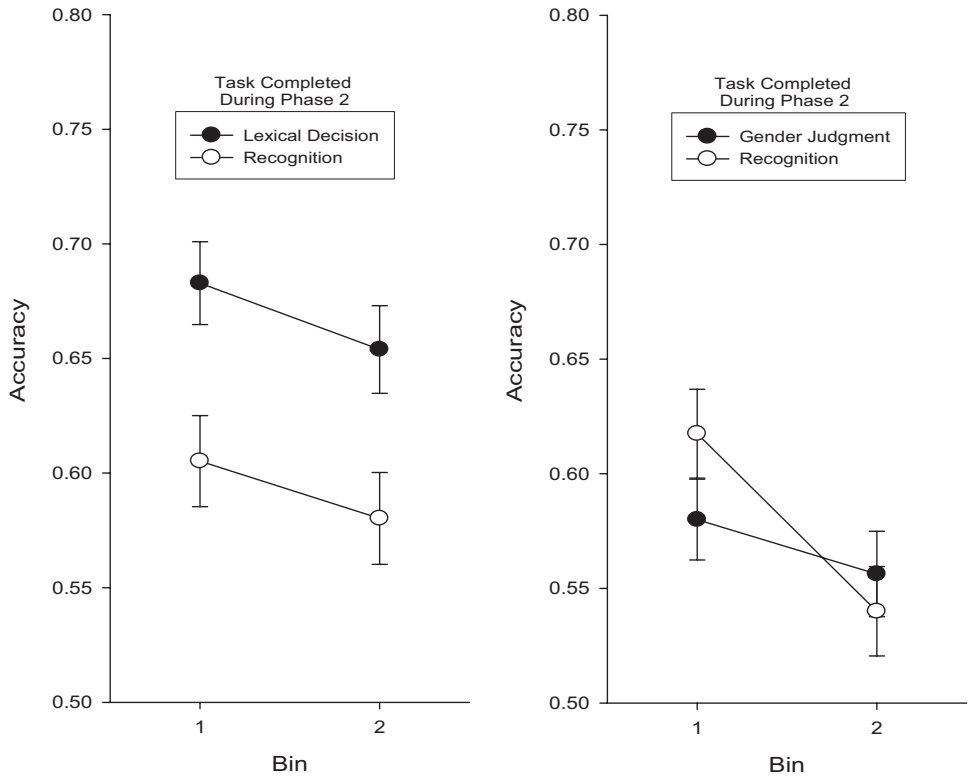

Figure 5. Both figures plot accuracy as a function of task completed during Phase 2 and test position bin. The left panel shows the recognition results of the condition in which Phase II words were tested. The right panel shows the recognition results of the condition in which Phase II faces were tested. The error bars represent the standard error.

iment, we manipulated the orienting tasks in a blocked fashion (e.g., study followed by LD or 2AFC) and did observe a decrease in accuracy for both tasks in a subsequent test, and the decrease in accuracy was larger for 2AFC than LD. That accuracy decreased in Phase III compared to Phase II is not surprising and is likely due to a change in temporal context. However, no simple, generic change of context explanation can explain that Phase II produced different levels of interference depending on the task performed. As a package, these results are consistent with the hypothesis that interference is modulated by the contextual cues used to probe memory and that context contains multiple attributes that can be cued separately (e.g., Underwood, 1969). If only temporal context were used as a probe, then the decrease in accuracy should be similar for the Phase II tasks (LD/gender ID and 2AFC) in Experiment 3 and there should be additional interference when other tasks were interpolated in Experiments 1 and 2. If only task context were used as a probe, then there should be no decrease in accuracy for the LD or gender ID conditions in all experiments. In contrast, the data show both an overall decrease in accuracy and a greater decrease for 2AFC than the other conditions. We suggest that memory was probed with both temporal and task context features, both of which contribute to interference. Specifically, the decrease in Phase III accuracy is due to a change in temporal context and the differential decrease based on the Phase II task as well as the lack of interference caused by interpolated tasks in Experiments 1 and 2 are due to task context.

General Discussion

These experiments place important constraints on the role of output interference in recognition memory. OI during recognition memory testing is unaffected when nonrecognition tasks are inserted between recognition test trials, even when the stimuli used to perform the task are very similar to those used to perform the recognition task. However, when these same tasks are blocked and completed prior to recognition testing rather than intermixed with recognition test trials, interference from these tasks is observed in the form of decreases in overall recognition accuracy. The level of interference is task-specific: Prior recognition testing harms memory more than LD or gender ID. Briefly, these results point to the importance of item information, task context, and temporal context.

We previously specified a simple model that accounts for OI in recognition testing (Criss et al., 2011). According to that model, information about the test items is stored in memory, and these new traces interfere with subsequent testing. The present findings showed that this model is too simple: Interference in recognition testing is produced by the performance of a different task preceding a block of recognition trials but not when the different task was interpolated between recognition trials. The challenges are to explain why recognition testing is only sometimes subject to interference from the performance of other tasks, and why recognition testing causes more interference than other nonrecognition tasks. We believe the findings suggest a more structured and nuanced view of context.

We hypothesize that several forms of contextual information are encoded during the experiment and used to probe memory during recognition testing (see Dennis & Humphreys, 2001; Howard & Kahana, 1999, 2002; Polyn, Norman, & Kahana, 2009, for related discussions); we further hypothesize that the types of context encoded at study and used in later memory probes vary in interesting and understandable ways (e.g., Underwood, 1969). *Temporal context* is associated with the individual and the environment, but not the items presented in the present task; it is correlated over recognition testing and interpolated tasks, albeit the degree of correlation will drop as the passage of time and the introduction of intervening tasks and events cause context change (Estes, 1955). Research has shown that such context changes little during the course of a given list but changes between study and test and when tasks change significantly from one setting to another (Klein, Shiffrin, & Criss, 2007; Jang & Huber, 2008). Generally temporal context is incidental to a given task and is not the focus during storage. Malmberg and Shiffrin (2005) suggested that such context is stored only during the first second or two of storage. Temporal context is however often critical at the time of test, helping to define the memory traces being sought. *Task context* is information about the task at hand and is generally a focus of attention at both study and test. By definition, it is similar for similar tasks. Thus, information concerning one's goal or the procedure used when performing LD, gender ID, or recognition is similar during any given task but different for different tasks. We assume that task context changes quickly when the tasks switch, as is the case for the interpolated paradigm used in Experiments 1 and 2, whereas temporal context does not.

We assume that a probe of memory activates traces to the extent that the contents of the probe are similar to the contents of the traces stored in memory and conversely does not activate traces to the degree that its contents are dissimilar to the probe. If a trace is activated by a probe, the amount of interference is a complex function of the strength of activation. When the probe and trace encode different items, interference will grow with stronger activation initially but then drop as differentiation occurs and the differences between probe and trace grow evident; when probe and trace are very similar then the amount of interference might continue to rise monotonically with activation strength (Criss, 2006, 2009, 2010; Ratcliff, Clark, & Shiffrin, 1990; Shiffrin, Ratcliff, & Clark, 1990; Shiffrin & Steyvers, 1997). Strength of activation is determined by probe-trace similarity, and similarity is determined by the matching and mismatching information in the probe compared with the trace, including the context in both. Thus, interference will often represent a balance of or competition between different types of information in the probe and trace. For instance, Lehman and Malmberg (2011) observed that temporal context cues were less effective than semantic category cues in reducing interference from task irrelevant traces during the performance of free recall when the traces represent items from a common category. In that case, semantic category cues served to better isolate the target traces during retrieval, than the temporal context cues that matched both target and nontarget memory traces.

Temporal context will be an effective and critical retrieval cue especially in cases in which there are no other cues available to perform the task. For instance, when items must be discriminated based solely on their appearance on one of two otherwise similar lists, a temporal context cue is most useful. However, when lists can be distinguished by some other factor, say the orienting task performed during the course of study, or a semantic difference between the items on the lists, then the context cue will be

correspondingly less important (cf. Gruppuso, Lindsey, & Kelley, 1997; Mulligan & Hirshman, 1997).

Thus, to predict the presence or absence of interference, one must assess with great care the various types of context information stored during list study and used as a probe at the time of test. In Experiments 1 and 2, we assume task context is an important component of the information stored during LD, gender ID, or recognition tests, and an important component of the recognition probe. Thus, task context will switch rapidly back and forth as tests alternate between recognition and LD or gender ID. However, temporal context likely will not change much as the mixed testing proceeds. We assume the recognition task context is a strong component of the probe in recognition testing and that this task context differs substantially from the LD and gender ID task contexts stored in the test traces of LD and gender ID tests, respectively. Thus, the traces of these other tests will not be activated and will not produce interference. The similar recognition task context in the recognition test traces will cause those traces to be activated during a subsequent recognition test; such traces will therefore cause interference and produce the OI that is observed.

In Experiment 3 the blocking of LD and gender ID testing in Phase II allows (or causes) temporal context to shift between Phases II and III. Note that the importance of task context, notwithstanding temporal context, is still a component of the memory probe during recognition testing in all three experiments. Thus, in Experiment 3 a shift of context between Phases II and III will cause a general decrement in recognition performance due to the increasing dissimilarity of the probe to initially stored traces in Phase I. Thus, we expect better recognition performance in Phase II recognition than Phase III recognition. Finally, note that task context is still part of the memory probe in Experiment 3. Thus, in Phase III testing, test traces in Phase II will be activated more when Phase II traces are of recognition tests, causing more interference, as observed.

This is a rather simple account for our results. It assumes that LD traces are not activated by recognition tests, due to the mismatch on task context in both study and test. It is possible, however, that some or all of the decrease in performance in Experiment 3 from Phases II to III could be due instead to activation of LD traces. This would be possible if the weight and importance given to task context in Phase III probes was lessened, and the weight and importance of temporal context was correspondingly increased. With less weight given to task context, Phase II LD test traces could then be activated and produce interference. Distinguishing this explanation from the shift-of-temporal-context explanation will have to be reserved for future research.

Consistent with our account for the findings is the observation that recognition testing for LD and gender ID traces shows output interference. The idea is, of course, simple: Each such recognition test lays down a test trace that includes similar context, in this case including both some temporal context and LD task (or gender ID as the case may be) context. Because each subsequent test uses similar context, there is a tendency to activate the previously stored Phase IV test traces, thereby producing OI.

There is a large literature showing that temporal context plays an important role in many episodic memory tasks (Farrell, 2012; Howard & Kahana, 2002; Lehman & Malmberg, 2009, 2011; Malmberg & Shiffrin, 2005; Mensink & Raaijmakers, 1988). This could hardly be otherwise, especially when temporal context is the only basis for performance. The storage of temporal context, and the use of temporal context during retrieval, to varying degrees in different tasks, does not of course preclude the joint storage of other information, and the joint use of other information in retrieval. Such other information in recognition, of course, includes the content of the study or test item, but we argue also includes the task context. We find the task context hypothesis extremely plausible, not only based on our findings, but conceptually: The participant can hardly do otherwise than focus on the task that is required. While the various features making up temporal context (such as the room of testing, the color of the monitor) are rather unimportant and incidental, the task context is critical.

The hypothesis that task context is reinstated as a cue during testing is somewhat related to the source-constrained retrieval hypothesis (Jacoby, Shimizu, Daniels, & Rhodes, 2005). Jacoby et al. (2005) had participants study a list followed by two tests (similar to Phase II and III here). The first test probed memory for the study list, while the second test probed memory for the foils from the first test. When depth of processing of the study items was manipulated between blocks, performance on the third test was better for those foils tested along with targets from the deeply encoded blocks than those tested following a block of shallow encoding. Their explanation was that the encoding task was reinstated during the first test and therefore new memory traces stored during the test for the deeply encoded list were themselves stored in a deep fashion. Because the foils are encoded during test, they are better recognized during the final phase of testing. Our account of the present results shares concepts with this account, in basing the predictions on the mix of different sorts of information, including different sorts of context information, at study and during testing.

What are implications for pure context-noise models? The issue is a bit complex; so initially it is useful to focus only on the present data (although any viable model must account for prior findings as well), which are consistent with the assertion that context information can at times influence the amount of interference observed during recognition testing. That, we believe, is the key implication of the current experiments, since prior findings have already established the contribution of item-noise to OI (Criss et al., 2011; Malmberg et al., 2012; Murdock & Anderson, 1975). A caveat: The existing context-noise model is only a model of word recognition, and therefore it has nothing to say about the output interference we observed in face recognition (Dennis & Humphreys, 2001). Therefore, when we discuss pure context-noise models and their ability to account for our findings, we are only working with hypothetical models. That having been said, a simple context-noise model that assumes that OI is due to changes in temporal context between study and test and that context changes as the result of episodic retrieval but not episodic encoding is insufficient to account for our findings (e.g., Criss et al., 2011). Specifically, LD, gender ID, and recognition caused interference, not just recognition testing. Also note that the impact of Phase II testing is greater when the stimuli consist of faces rather than verbal material. A context-noise model can possibly account for these findings by assuming that context-changes more when making face judgments than making lexical decisions. However, at some point this becomes a circular description of output interference since the model

of context that is being assumed makes no a priori predictions about under what conditions context should change and by how much, hence the importance of developing a model of temporal context like the one we outline. Further, there is evidence that suggests that item noise contributes to OI in recognition testing (e.g., Criss et al., 2011; Criss & Shiffrin, 2004; Malmberg et al., 2012; Murdock & Anderson, 1975; Norman & Waugh, 1968), which is a main point of Experiment 3 where Phase III testing was negatively affected by Phase II testing regardless of what task was performed during Phase II. Thus, while these data are consistent with a role of context in memory and specifically with a differential effect of temporal and task contexts depending on the testing demands, the only explanation of OI in a context-noise model (i.e., in Criss et al., 2011) is insufficient, both in details and conceptually, with the data presented here.

Summary

Items presented during interpolated trials (of LD or gender ID) were recognized as well as items tested during the recognition trials, but the interpolated trials did not affect recognition accuracy (Experiments 1 and 2) or interfered with memory less than additional test trials (Experiment 3). We speculated that these other task trials provided a highly effective task context that distinguished the item information stored during the performance of the interpolated task trials from the item information stored during the performance of the recognition trials. Temporal context is nonetheless part of what is stored and is a component of the retrieval probe. If temporal context shifts between a block of LD or gender ID task trials and a subsequent block of recognition tests, this would produce the decrease in recognition performance that is observed in Experiment 3. We suggest that the present results highlight the importance of distinguishing the various types of context information that are stored in traces at study and that are used in subsequent retrieval probes.

References

Anderson, J. R., & Bower, G. H. (1973). *Human associative memory*. Washington, DC: Winston and Sons.

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

Criss, A. H. (2004). The representation of single items and associations in episodic memory. *Dissertation Abstracts International: Section B. Sciences and Engineering, 65,* 6882.

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

Criss, A. H. (2009). The distribution of subjective memory strength: List strength and response bias. *Cognitive Psychology, 59,* 297–319.

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

Criss, A. H., & Malmberg, K. J. (2008). Evidence in support of the elevated attention hypothesis of recognition memory. *Journal of Memory and Language, 59,* 331–345. doi:10.1016/j.jml.2008.05.002

Criss, A. H., Malmberg, K. J., & Shiffrin, R. M. (2011). Output interference in recognition memory. *Journal of Memory and Language, 64,* 316–326. doi:10.1016/j.jml.2011.02.003

Criss, A. H., & McClelland, J. L. (2006). Differentiating the differentiation models: A comparison of the retrieving effectively from memory model (REM) and the subjective likelihood model (SLiM). *Journal of Memory and Language, 55,* 447–460. doi:10.1016/j.jml.2006.06.003

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,* 800–807. doi:10.1037/0033-295X.111.3.800

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,* 416–424. doi:10.1016/j.jml.2010.06.001

Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review, 108,* 452–478. doi:10.1037/0033-295X.108.2.452

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

Farrell, S. (2012). Temporal clustering and sequencing in short-term memory and episodic memory. *Psychological Review, 119,* 223–271.

Gruppuso, V., Lindsay, D. S., & Kelley, C. M. (1997). The process dissociation procedure and similarity: Defining and estimating recollection and familiarity in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 259–278. doi:10.1037/0278-7393.23.2.259

Howard, M. W., & Kahana, M. J. (1999). Contextual variability and serial position effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 923–941. doi:10.1037/0278-7393.25.4.923

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

Jacoby, L. L., Shimizu, Y., Daniels, K. A., & Rhodes, M. G. (2005). Modes of cognitive control in recognition and source memory: Depth of retrieval. *Psychonomic Bulletin & Review, 12,* 852–857. doi:10.3758/BF03196776

Jang, Y., & Huber, D. E. (2008). Context retrieval and context change in free recall: Recalling from long-term memory drives list isolation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 112–127. doi:10.1037/0278-7393.34.1.112

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). New York, NY: Psychology Press.

Kucera, H., & Francis, W. (1983). *Computational analysis of present day American English*. Providence, RI: Brown University Press.

Lehman, M., & Malmberg, K. J. (2009). A global theory of remembering and forgetting from multiple lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 970–988. doi:10.1037/a0015728

Lehman, M., & Malmberg, K. J. (2011). Overcoming the effects of intentional forgetting. *Memory & Cognition, 39,* 335–347. doi:10.3758/s13421-010-0025-4

Lehman, M., & Malmberg, K. J. (in press). A buffer model of encoding and temporal correlations in retrieval. *Psychological Review*.

Malmberg, K. J., Criss, A. H., Gangwani, T. H., & Shiffrin, R. M. (2012). Overcoming the negative consequences of output interference from recognition memory testing. *Psychological Science, 23,* 115–119. doi:10.1177/0956797611430692

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,* 616–630. doi:10.1037/0278-7393.28.4.616

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

Malmberg, K. J., Steyvers, M., Stephens, J., & Shiffrin, R. M. (2002). Feature frequency effects in recognition memory. *Memory & Cognition, 30,* 607–613. doi:10.3758/BF03194962

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective- likelihood approach to the experience in recognition memory. *Psychological Review, 105,* 724–760. doi:10.1037/0033-295X.105.4.734-760

McGeoch, J. A. (1942). *The psychology of human learning*. New York, NY: Longmans, Green.

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

Mulligan, N. W., & Hirshman, E. (1997). Measuring the bases of recognition memory: An investigation of the process dissociation framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 280–304. doi:10.1037/0278-7393.23.2.280

Murdock, B. B., & Anderson, R. E. (1975). Encoding, storage, and retrieval of item information. In R. L. Solso (Ed.), *Theories in cognitive psychology: The Loyola symposium* (pp. 145–194). Hillsdale, NJ: Erlbaum.

Murnane, K., Phelps, M. P., & Malmberg, K. J. (1999). Context-dependent recognition memory: The ICE theory. *Journal of Experimental Psychology: General, 128,* 403–415. doi:10.1037/0096-3445.128.4.403

Norman, D. A., & Waugh, N. C. (1968). Stimulus and response interference in recognition memory experiments. *Journal of Experimental Psychology, 78,* 551–559. doi:10.1037/h0026637

Phillips, P. J., Wechsler, H., Huang, J., & Rauss, P. J. (1998). The FERET database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing, 16,* 295–306. doi:10.1016/S0262-8856(97)00070-X

Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009). Task context and organization in free recall. *Neuropsychologia, 47,* 2158–2163. doi:10.1016/j.neuropsychologia.2009.02.013

Raaijmakers, J. G., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review, 88,* 93–134. doi:10.1037/0033-295X.88.2.93

Ratcliff, R., Clark, S., & Shiffrin, R. M. (1990). The list-strength effect: I. Data and discussion. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 163–178.

Roediger, H. L., & Schmidt, S. R. (1980). Output interference in the recall of categorized and paired associate lists. *Journal of Experimental Psychology: Human Learning and Memory, 6,* 91–105. doi:10.1037/0278-7393.6.1.91

Schulman, A. I. (1974). The declining course of recognition memory. *Memory & Cognition, 2,* 14–18. doi:10.3758/BF03197485

Shiffrin, R. M., Ratcliff, R., & Clark, S. E. (1990). The list-strength effect: II. Theoretical mechanisms. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 16,* 179–195.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review, 4,* 145–166. doi:10.3758/BF03209391

Shiffrin, R. M., & Steyvers, M. (1998). The effectiveness of retrieval from memory. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 73–95). Oxford, England: Oxford University Press.

Smith, A. D. (1971). Output interference and organized recall from long-term memory. *Journal of Verbal Learning & Verbal Behavior, 10,* 400–408. doi:10.1016/S0022-5371(71)80039-7

Smith, A. D., D'Agostino, P. R., & Reid, L. S. (1970). Output interference in long-term memory. *Canadian Journal of Psychology, 24,* 85–89. doi:10.1037/h0082845

Tulving, E., & Arbuckle, T. Y. (1966). Input and output interference in short-term associative memory. *Journal of Experimental Psychology, 72,* 145–150. doi:10.1037/h0023344

Underwood, B. J. (1969). Attributes of memory. *Psychological Review, 76,* 559–573. doi:10.1037/h0028143

Wickens, D. D. (1970). Encoding categories of words: An empirical approach to meaning. *Psychological Review, 77,* 1–15. doi:10.1037/h0028569

Wickens, D. D., Born, D. G., & Allen, C. K. (1963). Proactive inhibition and item similarity in short term memory. *Journal of Verbal Learning & Verbal Behavior, 2,* 440–445. doi:10.1016/S0022-5371(63)80045-6

Received June 30, 2012
Revision received December 29, 2012
Accepted January 3, 2013 ■