Remembering words not presented in sentences: How study context changes patterns of false memories

LAURA E. MATZEN AND AARON S. BENJAMIN
University of Illinois, Urbana-Champaign, Illinois

People falsely endorse semantic associates and morpheme rearrangements of studied words at high rates in recognition testing. The coexistence of these results is paradoxical: Models of reading that presume automatic extraction of meaning cannot account for elevated false memory for foils that are related to studied stimuli only by their visual form; models without such a process cannot account for false memory for semantic foils. Here we show how sentence and list study contexts encourage different encoding modes and consequently lead to different patterns of memory errors. Participants studied compound words, such as *tailspin* and *floodgate*, as single words or embedded in sentences. We show that sentence contexts led subjects to be better able to discriminate conjunction lures (e.g., *tailgate*) from old words than did list contexts. Conversely, list contexts led to superior discrimination of semantic lures (e.g., *nosedive*) from old words than did sentence contexts.

Memory errors provide a rich source of data for investigating the structure and organization of human memory. Examination of the factors that lead to the creation of false memories tells us about the underlying composition and arrangement of memory representations. A large body of research has investigated memory errors related to aspects of verbal stimuli, and these studies have found a surprising variety of errors for different kinds of materials. Some studies have found that participants make memory errors by endorsing words or sentences that are similar to studied items in terms of meaning (Johnson, Bransford, & Solomon, 1973; Roediger & McDermott, 1995); others have found that participants endorse words that are similar to the studied items only in visual appearance (Jones & Jacoby, 2001; Underwood & Zimmerman, 1973). These two patterns of results have very different implications in terms of what information about a studied word is encoded and retained in memory.

The strategy of using memory errors to infer something about the nature of memory representations is one that has been widely employed in social psychology (e.g., Hastie & Kumar, 1979; B. K. Payne, Jacoby, & Lambert, 2004), as well as in some subdomains of cognitive psychology, such as source memory (e.g., Bayen, Nakamura, Dupuis, & Yang, 2000; Hicks & Cockman, 2003). However, the recent rush in interest in false memory motivated by an article by Roediger and McDermott (1995; see also Deese, 1959) has spurred more theorizing about the retrieval, matching, or decision processes that yield such errors (e.g., Benjamin, 2001; Brainerd & Reyna, 1998, 2001; Gallo & Roediger, 2002; Israel & Schacter, 1997;

Miller & Wolford, 1999; D. G. Payne, Elie, Blackwell, & Neuschatz, 1996) than about how memory errors can be exploited to better understand the types of representations that promote them. Here, we examine the two most commonly employed experimental contexts for the study of verbal stimuli—lists and sentences—and show that those contexts modulate the types of memory errors that arise. By doing so, we show that study context influences the form and not just the strength of memory representations, and we provide a preliminary explanation of how this takes place.

Our primary goal in this article is to demonstrate that different study contexts have predictable effects on patterns of false memory—effects that reveal something about the nature of encoding strategies in those contexts. The results also have several interrelated implications for understanding the nature of memory errors and how to study them. Principally, we show that overemphatic theorizing about the processes at test that promote false memory miss part of the picture: Processes at encoding set the stage for false memory by promoting representations that are biased toward the goals of the learner (see, e.g., Benjamin, 2008). Such bias determines which representations are confusable with each other and, thus, what types of false memory are observed. Thus, a second implication is that it can be misleading to compare or collapse across measures of false memory that appear similar but follow different encoding regimens. We show that very different types of information are extracted from individual words embedded in meaningful sentence contexts and from the same words when they are studied in lists.

L. E. Matzen, lematze@sandia.gov

Types of Memory Errors

First, we review two of the most prominently studied types of memory errors and argue that (1) the presumed bases for these errors are somewhat contradictory, and (2) their coexistence is thus somewhat problematic for current theory. *Conjunction errors* occur when participants mistakenly endorse test words that are perceptually or phonetically recombined versions of actually studied words (Jones & Jacoby, 2001; Reinitz, Lammers, & Cochran, 1992; Underwood, Kapelak, & Malmi, 1976; Underwood & Zimmerman, 1973). The study stimuli in such experiments are usually individual compound words, such as *blackmail* and *jailbird*, and the critical lures at test are rearrangements of those words, such as *blackbird*. An analogous pattern of errors is evident for lures that consist of recombined syllables of shorter study words, such as *instruct* and *consult*, where *insult* is the lure (Underwood & Zimmerman, 1973). In these experiments, semantic relationships between the studied words and the conjunction lures are typically minimized in order to rule out the possibility that semantics underlie the effect. Thus, the high rate of such conjunction errors suggests that the surface forms of words and syllables are maintained in memory for some time, resulting in a misleading sense of familiarity when word components are recombined.

These results are surprising in light of other studies that show that participants remember the semantics of studied items but retain little information about their surface forms. Studies using sentences or stories as stimuli (Bransford & Franks, 1971; Brewer, 1977; Johnson et al., 1973) have found that participants have little or no memory for the surface forms of the words or sentences they have studied. This leads to *semantic errors*. Participants seem to distill these longer stimuli down to their basic meaning, losing any information about the exact structure of the sentences (see Bock & Brewer, 1974). Potter and Lombardi (1990) showed that although readers who were engaged in sentence processing tasks were largely accurate in their recall of the sentences, they seemed to reconstruct each sentence on the basis of their memory for its message-level meaning, rather than using stored information about the exact words and their order. This led participants to substitute semantically similar words for the original words in the sentences, even in immediate recall.

Analogous results have been found in numerous experiments using recognition tasks. For example, people in one study (Bransford & Franks, 1971) were likely to endorse test items that contained the same basic ideas as sentences on the study list, even if those ideas were in very different sentence structures and were combined with additional related sentences. Similarly, when participants were presented short stories describing an event that had a probable but unstated consequence, they incorrectly endorsed test sentences about the implied event (Johnson et al., 1973). For example, having heard "The boy hit the baseball and watched as it flew into the picture window in the house," participants were likely to endorse as previously heard a statement about a baseball breaking a window. Although this statement was never actually heard, it did follow logically from the events of the heard story. Brewer (1977) found that

participants were sometimes even more likely to remember the unstated implication of a sentence than the original sentence itself. After hearing the sentence "The hungry python caught the mouse," participants were far more likely to recall "The hungry python ate the mouse" than they were to correctly recall the original sentence. These memory errors indicate that participants remember the gist of the studied sentences, but little about their surface forms. When asked to recognize or recall the surface forms of the studied sentences, the participants reconstructed the sentences using a combination of the gist information they had stored and their knowledge of likely events in the world.

Similar effects have been found in experiments using word lists containing semantically or associatively related words. Using categorized word lists of common semantic associates, Roediger and McDermott (1995) found that participants often falsely recalled and recognized the unstudied associate from which those stimuli were drawn. This effect occurred even when the relationship between words was purely semantic and not associative (e.g., Benjamin & Bawa, 2004; Shiffrin, Huber, & Marinelli, 1995).

In each of these experiments, the critical memory error—the false recognition of or recall for material semantically related to studied material—revealed that information about the form of the individual sentences, stories, or thematic lists was lost in memory, leaving only abstract representations of their basic meanings. At test, the participants relied on the gist of the items they had studied, often in combination with their own inferences and world knowledge. Without access to any information about the surface forms of the original words and sentences, participants were highly susceptible to memory errors based on similarity in meaning between the study and test items.

Influences of Study Context

On the face of it, the existence of these two types of errors seems paradoxical. If participants rapidly lose information about the surface characteristics of linguistic stimuli, why do they false alarm to semantically dissimilar but physically related lures? Likewise, if they retain those surface characteristics, why are they prone to falsely endorsing semantically related lures? In this study, we consider the question of whether the context of study can modulate the type of encoding and, consequently, the form of memory for words. We start from the perspective of Benjamin (2008), who argued that encoding is always strategic, and that any evaluation of memory performance requires an assessment of the learner's goals and the task affordances. Sentences imply a very different goal set for the learner than do lists of unrelated words. In almost every instance of the participants' lives prior to entering this experiment, their memory for sentences was "assessed" by their ability to recall the semantics of the material. Students are not instructed to repeat the text back verbatim on essay tests—in fact, they may encounter charges of plagiarism if they do. Telling stories among friends requires the adequate reconstruction of events in a series, rather than the reproduction of specific words. In the rare instances in which verbatim reproduction is valued, such as reciting the Gettysburg Address or retelling a joke with peculiar

syntax and specific words that are critical to the humor, such reproduction is difficult and prone to error.

On the other hand, encountering lists of words provides a very different context and set of goals. Grocery lists, to-do lists, and vocabulary terms for a foreign-language test are all contexts that emphasize the need for verbatim retention. Remembering that I need to get "food," but not remembering the set of specific items needed when I reach the grocery store, is useless; the burden is on me to remember the exact individual items, not just their gist.

Because of this accumulation of experience in day-to-day life, participants in an experimental setting are likely to take different approaches to words that are presented in different study contexts. Sentence or story contexts encourage participants to discard surface information, likely because the context implies its lack of future usefulness and because of the considerable demand on the systems underlying encoding and comprehension to retain the surface information for a number of sentences. Participants likely focus instead on the meaning of each item, or on associations between words in the sentence or in the thematic list. From the strategic-encoding perspective, then, we assume that sentences should elicit a relatively greater evaluation of the semantic content of individual words, whereas lists should encourage a lower level retention strategy that promotes greater verbatim recall.

These ideas also relate to the *transfer-appropriate processing* account for memory performance. This account holds that memory performance is enhanced to the degree that the same kinds of processing are used during both study and test. For example, Morris, Bransford, and Franks (1977) showed that participants performed better on a semantic recognition task after doing semantic processing during encoding, but performed better on a rhyme recognition task after doing rhyme processing during encoding. A *levels-of-processing* account predicts that participants would do better on both tests after doing semantic processing during encoding, in which case the words should be more deeply encoded. However, the pattern found by Morris and colleagues showed that the match between the types of processing called for at study and at test outweighed the effects of deeper processing at encoding.

With respect to false memory, the degree of match or mismatch between the type of processing that a participant uses during study and the types of lures presented at test should play a role in determining the participant's susceptibility to the lures. A strategy of discarding surface information and encoding information at a deep, semantic level should give rise to the types of semantic memory errors seen in experiments in which participants remembered the general theme of the items, but little about their exact form. This study strategy should also make participants less susceptible to conjunction errors of the type reported by Jones and Jacoby (2001). With less information about the visual forms of the words being studied, participants should not experience such a high degree of match between physically similar lures and memory for the study list. This should lead to fewer false memories in response to the conjunction lures.

The opposite pattern of false memories would then obtain for participants who study lists of individual, decontextualized words. This study context signals to the comprehension system that extracting meaning is difficult and less useful, and perhaps that no clear gist is being formed across the study session. Thus, surface structure is retained to a greater degree, and the information that participants retain makes them less susceptible to falsely endorsing semantic lures. The cost of this process is that lures that are composed of rearranged surface structures become more alluring by virtue of their relatively greater match with the contents of memory for the study episode. However, without the context provided by a sentence or story, people may be more likely to remember specific details about the word, rather than just the gist of its meaning within a larger unit. In this situation, semantically related test items may be less likely to lead to memory errors, simply because people will have more specific memories about the studied words that could help them reject the lures.

Some prior studies have compared memory for words studied out of context with that for those studied in sentences (e.g., Murnane & Shiffrin, 1991), but their focus has been on how the number of memory traces stored is affected by changes in context. Our focus in the present study is on how changes in context influence the nature of the information that is encoded for studied words and how those changes can account for the seemingly discrepant patterns of memory errors seen in the previous literature on false memory. To our knowledge, this is the first study to investigate the processing of compound words within sentence contexts. In addition, although there has been some investigation of false memories for sentences (Reinitz et al., 1992), the sentences used typically have the same basic frame for all items (such as *the* X *saw the* Y, with different nouns substituted for *X* and *Y*) and provide little meaningful context. In the present study, we used much richer and more natural sentence contexts, more like those a reader would encounter in everyday life.

In summary, the nature of the information that people retain in memory when studying a list of words should influence a trade-off between meaning-based and structure-based false memory. When words are placed in a rich semantic context (such as a sentence), the way in which they are processed and the information that is gleaned from them is likely to change. This change should influence the pattern of false memories, making people more susceptible to semantic lures, but less susceptible to conjunction lures.

In the present study, we conducted three experiments to test these predictions. In the first experiment, participants studied either a list of compound words or a list of sentences in which the same compound words were placed into sentence contexts. Both groups of participants were then given identical memory tests that included conjunction lures that were visually similar to the studied words. In the second experiment, participants studied the same lists of compound words or sentences, but were given a memory test that included semantic lures that were similar in meaning to the studied words. In the third experiment, participants studied both single words and sentences and received a memory test that included both semantic and conjunction lures. We hypothesized that conjunction lures and old items would be less discriminable following word-

list study than following sentence-context study, because those lures would place a premium on memory for surface structure. Similarly, semantic lures should be less discriminable from old items following sentence-context study than following word-list study.

Analytic Techniques in the Measurement of False Recognition

Traditional studies of false memory evaluate false remembering in several ways. Most commonly, they examine mean false alarm rates between conditions. This strategy is appropriate when the response policy is equivalent between the relevant conditions. A detection-theoretic interpretation of this analysis is depicted in the top panel of Figure 1. As long as the rememberer employs the same response criterion for endorsing an item across conditions, the false alarm rate reveals something about the relative proportion of items that surpass that criterion, or how compelling those items are to the rememberer. As can be seen, it is not necessary for this strategy to address the location of the criterion (the dotted line) or the location of the distribution for studied items, because they remain constant across the conditions of interest.

Alternatively, if the response policy is thought to differ between conditions, but overall memory for the studied items does not, the appropriate measure of false memory is not simply the false alarm rate, but rather an estimate of the discriminability of old and new items. This can be seen in the middle panel of Figure 1. Because the criteria differ with the conditions of interest, the false alarm rates reflect a confluence of false memory and different response policies. For example, if one were to compare false memory for different types of lures, one could be reasonably certain that memory for the actually studied old items would not vary with the manipulation, but that the response criterion might. In that case, a measure of discriminability or distance between the distributions circumvents the problems posed by different criteria.

The final case represents the present situation, in which one wishes to compare false memory for different types of items across different conditions. In the experiments in the present study, the participants studied words under experimental conditions that were likely to lead to different levels of overall memory, as well as different response criteria across conditions. The items in the different experimental conditions differ in discriminability, as represented by the two distributions for old items in the lower panel of Figure 1. This makes direct comparison of distances, as shown in the middle panel, inappropriate. In addition, since it is likely that the different lure types promote different criteria, and that the differences in discriminability exacerbate these differences (e.g., Hirshman, 1995), the strategy shown in the top panel is inappropriate as well. Our strategy thus involves comparing the relative distances between distributions and between conditions, as we detail below.

Present Analysis

The measure of relative discriminability used in this study is d_a, which is based on basic assumptions of the
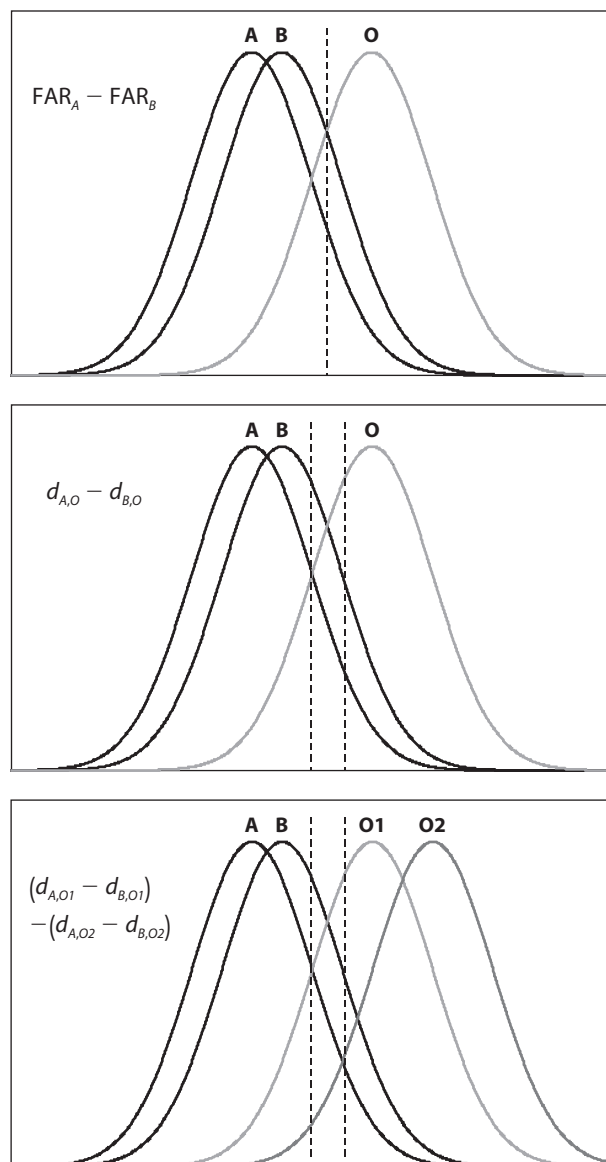

Figure 1. Detection-theoretic representations of the assessment of false memory. Top panel: A direct comparison of false alarm rates is appropriate when a rememberer uses the same response criterion across conditions. Middle panel: Overall memory for the studied items does not differ across conditions, but the rememberer uses different response criteria. In this case, the appropriate measurement of false memory is a measure of discriminability, such as a comparison of the distances between the distributions. Bottom panel: Both the overall memory for the studied items and the placement of response criteria differ across conditions. The appropriate measurement of false memory here is a comparison of the relative distances between distributions across the different conditions. The present experiments make the comparison shown in the lower panel by using Δd_a as a measure of discriminability that can be compared across experimental conditions.

theory of signal detection (Green & Swets, 1966), as applied to recognition memory (Egan, 1958), and effectively handles the evidence that, in recognition, the underlying probability distributions, unlike those shown in Figure 1, differ in variance as well as mean. Participants rated the

test items on the basis of whether they believed the words to be old or new, and the rating data were used to generate isosensitivity functions.[1] d_a represents the shortest distance from the origin of a 2-D space to the isosensitivity function, when plotted in normal–deviate coordinates (scaled by a constant), and is used here (see also Banks, 2000; Benjamin, Diaz, & Wee, in press) as a measure of discrimination between old and unrelated new items, as well as between old items and lures. Its metric properties make it ideally suited for this novel analysis of false memory, in which we compare susceptibility to memory errors across conditions with different overall levels of performance.

A particularly compelling lure leads to responses that are more similar to those seen for old items and less similar to those seen for unrelated new items. By comparing how well participants are able to discriminate lures from old items relative to how well they are able to discriminate new, unrelated items from old items, we are able to determine how compelling semantic and conjunction lures are relative to one another under conditions in which overall levels of performance and response bias are different. A high d_a value indicates that participants were largely successful at discriminating one group of items from another. For example, if the d_a value is higher for the old–new comparison than for the old–lure comparison, this indicates that participants were better at discriminating new items from old items than they were at discriminating lures from old items. In other words, that difference would indicate that participants were more likely to identify a lure as being old than they were to identify a new, unrelated word as being old, indicating that the lures were more compelling and led to more memory errors. Additionally, d_a values have metric qualities that allow us to make direct comparisons across conditions by subtracting the old–lure d_a value from the old–new d_a value for each participant, in order to generate Δd_a values (Green & Swets, 1966; Peterson, Birdsall, & Fox, 1954; Swets, 1986). The resulting Δd_a values indicate how likely the participants are to correctly identify a lure as being a new item. If the Δd_a value is small, it indicates that the participant was largely successful at identifying the lures as new words. A small Δd_a shows that the participant typically responded to lures and to new, unrelated items in the same way. A high Δd_a value indicates that the participant was often unsuccessful at identifying the lures as unstudied items and that he or she was more likely to respond to them as if they were old items. The Δd_a values allow us to determine the relative discriminability of different types of lures from old items following different study conditions.

EXPERIMENT 1

Method

Participants. Sixty-one University of Illinois undergraduates participated in Experiment 1 for credit in an introductory psychology course. Five participants were dropped because they were not native English speakers, leaving 56 participants (23 female), whose data were included in the analysis. The mean age of the participants was 19 years (range = 18–29).

Design. The critical variable was whether items were studied within the context of sentences or not (manipulated between subjects). Item types at test were old, (unrelated) new, and conjunction lures (manipulated within subjects). The remainder of the design variables were for counterbalancing purposes and will be described below. The dependent variable was confidence in the recognition judgment, used to generate individual isosensitivity functions within each condition.

Materials. There were a total of 384 compound words forming 128 triplets, in which 2 parent words (such as *tailspin* and *floodgate*) were recombined to form a conjunction lure (*tailgate*). Eighty of the triplets were from the set used by Jones and Jacoby (2001), although some were slightly modified.

The stimuli were divided into four counterbalanced lists. In each list, there were 64 triplets, for which both parent words were studied and the to-be-rejected conjunction lure was tested. This yielded 128 study items and 64 test items. For the remaining 64 triplets, one parent was studied and served as an old, to-be-endorsed item on the test. The other parent was unstudied and served as a new, to-be-rejected lure on the test. This yielded an additional 64 study items and 128 test items. Thus, both study and test lists were 192 words in length.

Table 1 depicts example items and illustrates the counterbalancing procedure. One counterbalancing variable reversed the sets of old and new items (compare Conditions 1 and 2) and also reversed the study order of the parents for the conjunction lure. The second counterbalancing variable (compare Conditions 1 and 3) swapped the triplets, so that the items that had served as parents for conjunction lures in the other condition now served as the old–new item set, and vice versa. Counterbalancing yielded four unique lists, each of which was assigned a unique study order. The old–new items were placed randomly within a subset of positions reserved for those items. The positions of the parents of the to-be-tested conjunction lure were maintained, but the assignment of Parent 1 (P1) and Parent 2 (P2) to those positions was counterbalanced. For example, *blackmail* appeared before *jailbird* on one list, and vice versa on another. For each pair of parent compounds, P1 and P2 were separated on the study list by 1–5 intervening words, with an average separation of 3 intervening words. The variation in spacing was included so that it would be very difficult for the participants to notice that the parent words could be recombined to form other words.

Four additional experimental lists were created by placing each of the parent compound words in a sentence context, such as "The fighter plane went into a tailspin after it was hit by enemy fire." This produced a total of 256 sentences that were placed into experimental lists using the same pseudorandom order that was created for the original word lists. The test lists that were used in the sentence-study condition were identical to those that were used in the word-list study condition.

Table 1
Examples of Items and Counterbalancing Procedure for Experiment 1

Counterbalancing Condition	Study			Test		
	Parent 1A	Parent 1B	Parent 2A	Conjunction Lure	Old (Parent 2A)	New (Parent 2B)
1	blackmail	jailbird	tailspin	blackbird	tailspin	floodgate
2	jailbird	blackmail	floodgate	blackbird	floodgate	tailspin
3	tailspin	floodgate	blackmail	tailgate	blackmail	jailbird
4	floodgate	tailspin	jailbird	tailgate	jailbird	blackmail

Each of the eight experimental lists was divided into four study blocks containing 48 experimental items and 2 filler items (one at the beginning, and one at the end of the block). In addition to the 2 fillers, each study block contained 32 parents of to-be-tested conjunction lures and 16 parents that were to be tested as old items. Each study block was followed by a test block containing 16 conjunction lures, 16 old items, and 16 unstudied parent items. These were intermixed in a pseudorandom order, so that no more than 4 test items of the same type appeared consecutively. Because some morphemes appeared in more than 1 word—particularly in the sentence-study condition—care was taken to ensure that the 2 morphemes in each of the lure words on the test block appeared the same number of times in the preceding study block. Additionally, the morphemes that formed the new items in the test block did not appear in any of the words in the preceding study block.

Procedure. Participants were seated in front of a computer monitor in a quiet room and were instructed either that they would be studying a list of words or that they would be studying a list of sentences for a subsequent memory test. They were instructed in advance that there would be four study blocks containing 50 items each and four test blocks that would test their memory for the preceding study block. Participants were given a chance to rest between blocks.

All of the words were presented in the center of the computer monitor in black 16-point Times New Roman text on a white background. The compound words in the word lists were presented individually for 2 sec and were followed by a 250-msec interstimulus interval. The sentences in the sentence lists were presented for 8 sec with a 250-msec interstimulus interval. During the test phases, participants saw 1 compound word at a time and were asked to respond by pressing key "1," "2," "3," or "4" on the computer keyboard. A response of "1" indicated that the participants were sure that the word had not appeared on the study list; a response of "2" indicated that they thought the word was new, but were not sure; a response of "3" indicated that they thought they had studied the word, but were not sure; and a response of "4" indicated that they were sure that they had studied the word. Each word stayed on the screen, along with a guide indicating what each response choice meant, until the participant selected his or her response.

The sentence lists had an additional test phase that followed each of the aforementioned test blocks and contained six yes-or-no questions about the content of various sentences in the preceding study block. For example, following a study block containing the sentences "After he discovered evidence of a crime, the butler threatened his employer with blackmail" and "The best player on the little league team was the young boy who played shortstop," the content test posed questions such as "Did the butler find evidence of a crime?" and "Was the pitcher the best player on the little league team?" These comprehension questions varied in difficulty, and some made reference to the compound word in a studied sentence, but others did not, as in the examples above. This test was included because, after seeing the first test phase (which tested only memory for compound words), the participants could have stopped reading the sentences and begun focusing on the compound words embedded within them instead. The comprehension tests after each block were included in an effort to keep the participants reading the sentences as naturally as possible throughout the experiment. On average, the participants responded to these questions correctly 81% of the time, and the percentages of correct responses were similar across all of the blocks (ranging from 76% to 87% correct), indicating that the comprehension test was successful at encouraging the participants to read all of the sentences.

The experiment lasted approximately 20 min for the word lists and 45 min for the sentence lists.

Analysis. The goal of the present experiment was to examine the extent to which sentence contexts modulate the plausibility of conjunction lures. We aimed to do this independently of any effects that context manipulation could have on overall response bias or accuracy. We believed that the raw performance data were likely to reveal an expected, uninteresting advantage for the word-list study condition because of the smaller number of exposed words and the shorter interval between the study items and test, and that this advantage additionally may have encouraged a more liberal response bias (Benjamin & Bawa, 2004). Thus, as discussed above, a direct comparison of false alarm rates across the two experimental conditions would not be meaningful. Instead of analyzing false alarm rates, we calculated d_a values for old–new and old–lure discrimination for each participant. These values were entered into a mixed model ANOVA with discrimination type (old–new vs. old–lure) as a within-subjects variable and study context (sentence vs. word) as a between-subjects variable. The measure of interest is Δd_a, the difference between d_a values for the two different comparisons.

Results

Table 2 provides the mean proportions of each confidence rating for each item type, and Figure 2 shows the Δd_a values for Experiment 1. Individual ratings tables were used to generate isosensitivity functions for the discrimination of old from unrelated test items and conjunction lures for both the sentence-study and word-study conditions. These functions were used to compute d_a values for each participant. All effects described for this experiment, as well as for Experiments 2 and 3, are significant at the $\alpha < .05$ level, unless otherwise noted. Overall, discrimination differed between study context conditions, as evidenced by the differences in d_a in old–new recognition ($d_a = 1.77$ for word-list study; $d_a = 1.33$ for sentence study) [$t(54) = 3.13$]. As noted above, it is not surprising that participants were somewhat less accurate overall in the sentence-study condition, given the much larger amount of information presented to them.

The critical test concerns the discrimination of old items from conjunction lures, which was expected to be relatively superior in the sentence-study condition. Discrimination was only slightly poorer for the old–lure ($d_a = 1.16$; $\Delta d_a = 0.17$) than for the old–new comparison in the sentence condition [$t(54) = 1.33$, n.s.], but was considerably lower ($d_a = 1.34$; $\Delta d_a = 0.43$) in the word-study condition [$t(54) = 3.05$]. This yielded a reliable interaction between lure type and study context [$F(1,54) = 16.25$]. This interaction confirms that old items were more easily discriminated from conjunction lures in the sentence-study than in the word-study condition.

Discussion

Experiment 1 showed that participants were more susceptible to incorrectly endorsing conjunction lures if they studied a list of words, rather than a list of sentences. Although participants in the sentence-study condition

Table 2
Mean Proportions of Each Confidence Rating for Each Item Type in Experiment 1

Confidence Rating	Sentence Condition			Word Condition		
	New	Conjunction Lure	Old	New	Conjunction Lure	Old
1	.27	.23	.07	.47	.35	.07
2	.51	.45	.19	.40	.37	.17
3	.16	.19	.17	.10	.16	.16
4	.07	.12	.57	.03	.12	.60

Note—Confidence rating: 1, *sure new*; 2, *unsure new*; 3, *unsure old*; 4, *sure old*.

Δd_a Values for All Experiments
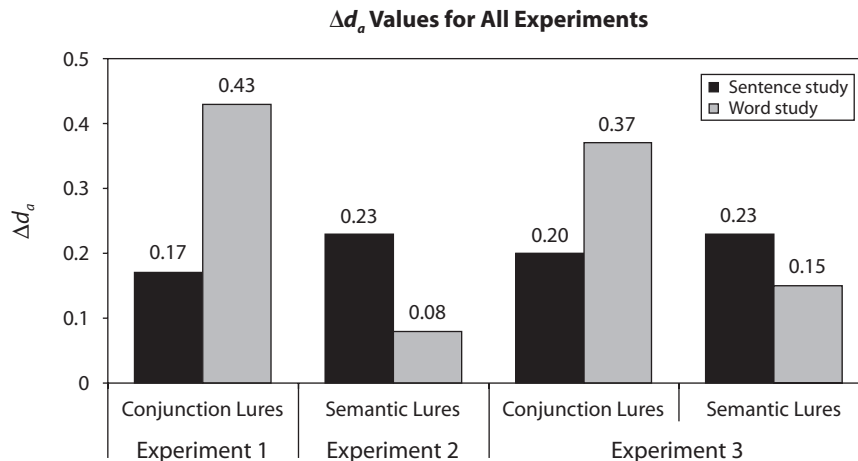

Figure 2. Δd_a values for all three experiments. Δd_a is the difference between the d_a value for old–new discrimination and the d_a value for old–lure discrimination for each condition in each of the experiments.

were presented much more information and had poorer memory for the words overall, participants in the word-study condition experienced relatively more difficulty in discriminating conjunction lures from old items than did participants in the sentence-study condition.

This result supports the hypothesis that the words in the study lists were encoded differently, depending on their context. When the words were studied without sentence contexts, participants retained more information about the surface forms of the words and less information about their meaning. In this case, the conjunction lures provided a better match to the contents of memory for the study episode, and participants were more likely to endorse the lures, even though the meanings of the lures did not match those of any of the original words on the study list. On the other hand, participants who had studied a list of sentences retained little information about the surface form of each word, but they were likely to retain some information about the gist of each sentence. This strategy could help the participants in two ways. First, when there is little surface information encoded in memory, the conjunction lures are poor matches for the contents of memory and are less likely to be endorsed as being old. Second, the information about the gist of each sentence can be used to reject the conjunction lures (Odegard, Lampinen, & Toglia, 2005). A participant could easily reject a lure such as *blackbird*, if he or she remembered that there were no sentences about birds on the study list.

This interpretation of differences in word processing produced by changes in study context has implications for other types of false memories as well. If participants in the sentence-study condition retain information about the gist of the sentences, but not the surface forms of the words, they should be more susceptible to semantic lures that are related to the studied items in meaning, but not in form. Similarly, if participants in the word-study condition retain more information about the surface forms of the words, but relatively less information about their meaning, they should be rela-

tively better at discriminating semantic lures from old items. We tested these predictions in Experiment 2.

EXPERIMENT 2

Method

Participants. Fifty-one University of Illinois undergraduates participated in Experiment 2 for credit in an introductory psychology course. Three participants were dropped because they were not native English speakers, leaving 48 participants (21 female), whose data were included in the analysis. The mean age of the participants was 20 years (range = 18–33).

Design. As in Experiment 1, the critical variable was whether items were studied within the context of sentences or not (manipulated between subjects). Item types at test were old, (unrelated) new, and semantic lures (manipulated within subjects). The dependent variable was confidence in the recognition judgment.

Materials. Experiment 2 used the same compound words and sentences that were used in Experiment 1. In addition to the compound words, there were 128 words that were semantically related to one of the parent words. These words were selected so that they would be interchangeable with the original compound words in both the list and the sentence contexts. For example, the semantic associate for *tailspin* was *nosedive*, and the two words were both appropriate in the sentence context, "The fighter plane went into a tailspin/nosedive after it was hit by enemy fire." Some of the semantic associates were also compound words, but most were not.

As in Experiment 1, the stimuli were divided into four counterbalanced study lists that were 192 items in length. In each list, there were 64 items for which one of the words in the semantically associated pair was studied and the other member of the pair was presented at test as a to-be-rejected semantic lure. An additional 64 items contained compound words (or their semantic associates) that were presented in the same form at test and served as old, to-be-endorsed items. The remaining 64 items were filler items that were taken from Experiment 1. These items were included to make the study phases of Experiments 1 and 2 as similar as possible. The assignment of the pairs of semantic associates to the old or lure conditions was counterbalanced across lists. The assignment of the semantic associates within each pair to study or test was also counterbalanced across lists. The critical items, old items, and fillers were placed in a pseudorandom order, and the experimental items were substituted into the appropriate slots to create four unique study lists.

On each test list, there were 64 to-be-rejected semantic lures, 64 to-be-endorsed old words, and 64 new, unrelated words. Unlike in Experiment 1, the new items for a given list could not be taken from among the old or lure items from other lists, because of the inherent semantic relationships among the critical items. Instead, the new words for each list were drawn from a pool of 107 words that had no semantic association with any of the words on the study lists. As with the lures and old items, slightly more than half of the new words were compound words. These compounds did not share syllables with any of the words on the study lists. The new words used for each test list were matched with the lures and the old words on that list in terms of length and frequency. Across all of the test lists, the average length of the words was 7.82 letters for old and lure items (since these were the same words appearing in different conditions on different lists) and 7.78 letters for the new items. The average frequency of the words was 15.79 for the old and lure items and 16.86 for the new items (based on the Kučera & Francis, 1967, norms included in Balota et al., 2007; a frequency value of 0 was assumed for items not appearing in Balota et al.'s, 2007, database).

The lists for Experiment 2 were divided into study and test blocks in the same way as in Experiment 1. Each was divided into four study blocks containing 48 experimental items and 2 filler items, one at the beginning and one at the end of the block. In addition to the 2 fillers, each study block contained 16 semantic associates of to-be-tested semantic lures, 16 words to be tested as old items, and 16 filler items. Each study block was followed by a test block containing 16 semantic lures, 16 old items, and 16 unrelated new items. These were intermixed in a pseudorandom order, so that no more than 4 test items of the same type appeared consecutively. Care was taken to ensure that the semantic lures on the test list were related to one and only one word in the preceding study list.

Four additional experimental lists were created by placing each of the critical items in the appropriate sentence context. The sentences were identical to those used in Experiment 1, except for a few minor modifications made in order to eliminate words that were semantically related to one of the critical items. The 256 sentences were placed in experimental lists using the same pseudorandom order that was created for the word lists. Each of the four sentence lists was divided into four study blocks containing 50 sentences each. The test blocks were identical to those that were used for the word lists. Again, care was taken to ensure that the semantic lures on the test list were related to one and only one word in the preceding study list.

As in Experiment 1, the sentence lists included sets of comprehension questions at the end of each test block. The questions were identical to those used in Experiment 1, and, on average, participants answered 80% of them correctly. The rate of correct responses was similar across all blocks (ranging from 78% to 83% correct), indicating that the participants had continued to read the sentences throughout the experiment.

The procedure and analysis used in Experiment 2 were identical to those used in Experiment 1.

Results and Discussion

Table 3 provides the mean proportions of each confidence rating for each item type, and Figure 2 shows the Δd_a values for Experiment 2. As in Experiment 1, the overall discrimination differed between study context conditions, as evidenced by differences in d_a in old–new recognition ($d_a = 2.03$ for word study; $d_a = 1.56$ for sentence study) [$t(46) = 2.44$].

The critical test concerns the discrimination of old items from semantic lures, which was expected to be relatively superior in the word-study condition. As expected, discrimination decreased only slightly for the old–lure comparison in the word-study condition ($d_a = 1.95$; $\Delta d_a = 0.08$) [$t(48) = 0.37$, n.s.], and decreased more substantially for

Table 3
Mean Proportions of Each Confidence Rating
for Each Item Type in Experiment 2

| Confidence Rating | Sentence Condition | | | Word Condition | | |
| | | Semantic | | | Semantic | |
	New	Lure	Old	New	Lure	Old
1	.31	.33	.06	.58	.58	.09
2	.52	.40	.20	.30	.29	.11
3	.14	.17	.18	.09	.08	.11
4	.03	.10	.56	.03	.06	.70

Note—Confidence rating: 1, *sure new*; 2, *unsure new*; 3, *unsure old*; 4, *sure old*.

the old–lure comparison in the sentence-study condition ($d_a = 1.32$; $\Delta d_a = 0.23$) [$t(44) = 1.83, p < .05$, one-tailed]. Most important, this yielded a significant interaction between lure type and study context [$F(1,46) = 5.62$].

As predicted, the pattern of results seen in Experiment 2 was the opposite of that seen in Experiment 1. When presented with semantic lures at test, the participants were better able to discriminate the lures from the old items when they had studied a list of words rather than a list of sentences. The kind of word processing involved in studying a list of sentences made participants relatively more susceptible to the semantic lures. In this case, they encoded the gist of each sentence, but less information about the specific words in the sentence. When presented with semantic lures that were consistent with the gist of a studied sentence, participants were likely to endorse them as old items. As in previous studies in which participants endorsed the unstated implications of studied sentences (e.g., Brewer, 1977), the gist information encoded from the sentences in Experiment 2 provided powerful cues that led the participants to endorse the semantic lures. For example, a participant might remember that he or she had read a sentence about a fighter plane crashing. The test item *nosedive* fits very well with this general scenario, and with little information encoded about the surface forms of the words in the original sentence, the participant is unlikely to remember that the word in the original sentence was actually *tailspin*. With gist information supporting the semantic lure and little surface information to contradict it, the participant is likely to endorse the lure.

However, when the participants were presented with a list of individual words rather than a list of sentences, they encoded relatively less semantic information and relatively more of the details of the surface forms of the words. This information benefited the participants both by making the semantic lures less appealing and by providing them details that could help them to reject lures that were consistent with the original items in meaning, but not in form.

It is possible that the between-subjects manipulation of lure type and the use of four separate study and test blocks led the participants to notice the relationship between the studied items and the lures at test. This could have prompted them to develop an unusual study strategy, such as ignoring the sentence contexts and searching for compound words. To eliminate this problem and other confounds that could stem from the between-subjects design used in Experiments 1 and 2, we conducted an ad-

ditional experiment using a entirely within-subjects design. Experiment 3 combined both sentences and isolated words in the study list, as well as both conjunction lures and semantic lures in the test phase. In addition, Experiment 3 used a single study list and a single test list. The single study–test phase design ensured that (1) participants could not tailor their study strategy across items in anticipation of seeing a particular type of lure; (2) there could be no changes in encoding strategy as a function of test experience over multiple blocks; and (3) participants would be forced to read and attend to the full sentences when they were presented.

EXPERIMENT 3

Method

Participants. Twenty-seven University of Illinois undergraduates participated in the experiment for credit in an introductory psychology course. Three participants were dropped because they were not monolingual English speakers, leaving 24 participants (2 female) whose data were included in the analysis. The mean age of the participants was 20 years (range = 18–24).

Design. As in Experiments 1 and 2, one critical variable was whether the items were studied within the context of sentences or not, but in Experiment 3, this variable was manipulated within subjects rather than between subjects. The second critical variable, lure type, was also manipulated within subjects in Experiment 3. The item types at test were old, (unrelated) new, conjunction and semantic lures whose parent words appeared in sentence contexts at study, and conjunction and semantic lures whose parent words appeared as single out-of-context words at study.

Materials. Experiment 3 used a subset of the compound words and sentences that were used in the previous two experiments, plus 5 new items that were created to avoid repetition of morphemes in the study lists. The stimuli were divided into eight counterbalanced study lists containing 160 items each. Ninety-six of these items were rotated through the same experimental conditions that were used in Experiment 1. On each list, 64 of the items from this subset (32 sentences and 32 single words) contained parent words that were recombined at test and presented as to-be-rejected conjunction lures. The other 32 items from this subset (16 sentences and 16 single words) contained compound words that were presented in the same form at test, serving as to-be-endorsed old items. Each study list also contained 64 items that were rotated through the same experimental conditions used in Experiment 2. Thirty-two of the items in this subset (16 sentences and 16 single words) contained one member of a pair of close semantic associates. The other member of this pair was presented at test as a to-be-rejected semantic lure. The order in which the two members of the pair appeared was counterbalanced across lists. The remaining 32 items in this subset (16 sentences and 16 words) contained one member of a pair of semantic associates that was presented in the same form at test, serving as a to-be-endorsed old item.

The 160 study items for each list were placed in a pseudorandom order, with the appropriate versions of each item placed in each slot to create eight unique study lists. Each study list had an associated test list that contained 192 items. Of the test items, 32 were conjunction lures, 32 were semantic lures, 64 were old items, and 64 were new, unrelated items. All of the conjunction lures and approximately half of the semantic lures were compound words, so a similar pattern was created for the old and new items, in which approximately three fourths of the old and new items were compound words, and one fourth were not. The same 64 new items were used for all eight lists. The new items were matched as closely as possible to the old items and lures in terms of length and frequency. The average length of the words on the test list was 8.26 letters for old items, 8.04 letters for lures, and 8.27 letters for new items. The average frequency of

the test items was 12.53 for the old items, 10.55 for the lures, and 6.30 for the new items (based on the Kučera & Francis, 1967, norms included in Balota et al., 2007; a frequency value of 0 was assumed for items not appearing in the Balota et al., 2007, database). The 192 test items for each list were placed in a pseudorandom order, so that no more than 3 items of the same type appeared in a row. The same order was used for all eight test lists, with the appropriate test items substituted into each slot.

Unlike in Experiments 1 and 2, where there were four separate study and test blocks, we used a single study phase in Experiment 3, followed by a single test list. For each test list, care was taken to ensure that the two morphemes in each conjunction lure appeared the same number of times (twice for one item, once for all other items) in the preceding study list. Additionally, none of the morphemes in any of the semantic lures or new items appeared anywhere in the preceding study list. There were no sentence comprehension questions in Experiment 3, because the participants did not know what the test phase would be like until they had completed the entire study block, and it is unlikely that they would have adopted a study strategy in which they ignored the sentence contexts.

Procedure. Participants were instructed that they would be studying a list of intermixed words and sentences for a subsequent memory test. During the study phase, 1 item at a time (a single word or a sentence) was presented on the computer monitor in black 16-point Times New Roman on a white background. Single words were presented for 2 sec, and sentences were presented for 8 sec, with a 250-msec interstimulus interval. The words and sentences were quasirandomly intermixed, with no more than 4 single words or 4 sentences appearing in a row. The test phase was the same as that in Experiments 1 and 2, with the participants rating each test word on a scale of 1 to 4.

Analysis. In the analysis of Experiment 3, d_a values were calculated for old–new, old–conjunction lure, and old–semantic lure discrimination for each type of study context for each participant. Four Δd_a values were calculated for each participant by subtracting the old–lure d_a values for each study context condition from the old–new d_a values for each condition. The Δd_a values for each participant were then entered into a within-subjects ANOVA with lure type (conjunction vs. semantic) and study context (sentence vs. word) as dependent variables.

Unlike in Experiments 1 and 2, the within-subjects design of Experiment 3 allowed for a meaningful comparison of hit rates and false alarm rates for different conditions. High-confidence responses were taken as the best indicator of the participants' performance, so to analyze the false alarm rates, the number of high-confidence "yes" responses for each lure condition for each participant was entered into a within-subjects ANOVA with lure type (conjunction vs. semantic) and study context (sentence vs. word) as dependent variables.

Results and Discussion

Table 4 provides the mean proportions of each confidence rating for each item type, and Figure 2 shows the Δd_a values for Experiment 3. The difference in overall discrimination between study context conditions was marginally significant, as shown by the differences in d_a in old–new recognition ($d_a = 1.25$ for word study; $d_a = 0.96$ for sentence study) [$t(23) = 1.98$, $p = .06$]. This difference in discrimination replicates that in the first two experiments, but is of a smaller magnitude. The fact that this difference is small in the present experiment is to be expected, because of the within-subjects design. In the first two experiments, participants saw only one type of study stimulus and studied more or less information overall, depending on whether they studied sentences or words. Both of these factors make it likely that the participants who studied lists of sentences in the first two experiments would set very different response criteria than those who

Table 4
Mean Proportions of Each Confidence Rating for Each Item Type in Experiment 3

Confidence Rating	Sentence Condition			Word Condition			
	Conjunction Lure	Semantic Lure	Old	Conjunction Lure	Semantic Lure	Old	New
1	.35	.35	.21	.32	.38	.14	.41
2	.41	.40	.24	.39	.40	.24	.42
3	.16	.14	.16	.18	.17	.18	.13
4	.09	.11	.39	.11	.05	.43	.04

Note—Confidence rating: 1, *sure new*; 2, *unsure new*; 3, *unsure old*; 4, *sure old*.

studied lists of words. In Experiment 3, where all of the participants studied both words and sentences and studied the same amount of information overall, it is very likely that their response criteria would be much more similar for the different study context conditions.

Critically, the interaction between study context (word or sentence) and lure type (semantic lure or conjunction lure) was reliable [$F(1,23) = 5.04$], replicating the effects seen in Experiments 1 and 2. There was a bigger decrease in performance for the conjunction lures whose parent items were studied as single words ($d_a = 0.88$; $\Delta d_a = 0.37$) [$t(23) = 4.91$] than there was for the conjunction lures whose parent items had been studied in sentences ($d_a = 0.76$; $\Delta d_a = 0.20$) [$t(23) = 3.20$]. The opposite pattern obtained for the semantic lures, with a bigger decrease in discrimination for lures whose parent items were presented in sentences ($d_a = 0.73$; $\Delta d_a = 0.23$) [$t(23) = 4.09$] than there was for lures whose parent items were presented as single words ($d_a = 1.10$; $\Delta d_a = 0.15$) [$t(23) = 2.34$].

The same pattern holds for the hits and false alarms in Experiment 3. Unlike in Experiments 1 and 2, the within-subjects design in Experiment 3 makes it possible for us to make a direct comparison of hit rates and false alarm rates across study conditions. The number of high-confidence "yes" responses was taken to be the best measure of the participants' susceptibility to the lures, so this number was used to calculate the hit rate and false alarm rates for each condition for each participant. The average hit rate for the old items did not differ significantly across study contexts (39% for old items that were originally studied in sentences and 43% for items that were originally studied as single words) [$t(23) = 0.84$]. As discussed above, the similar hit rates for the two conditions are to be expected in this experiment because of the within-subjects design. For the conjunction lures, the average percentage of high-confidence false alarms was 9% in the sentence condition and 11% in the word-list condition. For the semantic lures, the average false alarm rates were 11% in the sentence condition and 5% in the word-list condition. The interaction between item type (word or sentence) and lure type (semantic lure or conjunction lure) was reliable [$F(1,23) = 6.71$], just as it was for the d_a values.

GENERAL DISCUSSION

The goal of these experiments was to examine the effects of study context on different types of false memories.

The two very different patterns of memory errors seen in these experiments suggest that changes in study context alter the way in which people process words and encode them in memory. When presented a list of sentences, participants were less susceptible to conjunction lures, but more susceptible to semantic lures. The opposite was true following study of decontextualized words.

These findings indicate that participants engage different encoding strategies when studying words in or out of a larger meaningful context. These strategies are not necessarily a conscious decision, but rather are based on the participants' previous experience with verbal materials in everyday life and the information that can be gleaned from the studied items. Before beginning experiments of the type described in this article, participants will have had extensive experience reading and remembering words in numerous contexts, including sentences and lists. Through this experience, they are likely to have developed expectations about what kinds of information about words are most useful in different contexts. When they encounter sentences or out-of-context words in an experimental setting, this previous experience is likely to guide their strategy for encoding the items. In the present experiments, when the participants encountered words in a sentence context, they encoded information about the gist of each sentence, but relatively less information about the exact forms of the words they contained. When they studied isolated words, the participants encoded relatively less semantic information about the words, but more detail about their surface features.

Each type of processing had advantages and disadvantages. Retaining more semantic information through gist processing allowed the participants to reject lures that did not fit with the gist of any of the original sentences. However, semantic lures that were consistent with the meaning of one of the sentences were difficult to reject, especially with little information about the surface features that could help the participant distinguish one semantic associate from another. When the participants retained more information about the surface features of the words and less information about their semantics, they were better able to reject semantic lures that did not match the forms of the original words. Yet they were also more likely to false alarm to conjunction lures that strongly resembled the studied words in form, but not in meaning.

These results can be understood as an example of transfer-appropriate processing. When the participants studied words in sentence contexts, they used encoding processes that were well suited to the kinds of informa-

tion that they would need in order to reject semantic lures presented at test. However, these same encoding processes were poorly matched to the kind of information needed to reject conjunction lures at test. The opposite was true for words that were presented out of context. This kind of stimulus presentation promoted encoding processes that were well matched with the kinds of processing needed for higher performance on a test using semantic lures, but were poorly matched for a test using conjunction lures.

It is important to note that in the experiments described here, the way in which the participants processed the words was more constrained in the sentence-study condition than in the word-study condition. The participants were instructed to read and try to remember the materials. The context provided by the sentences constrained the meaning of the critical words and the way in which they were processed, whereas the out-of-context words remained unconstrained. Our focus in the present study was on naturalistic study strategies that participants might adopt when asked to read and remember different types of verbal materials. We feel that it was important to leave the participants' choices about study strategies as unconstrained as possible in order to gain insight on how strategy choice may have affected the results of previous experiments on conjunction and semantic memory errors. The patterns of memory errors produced in the present experiments helped us infer the kinds of study strategies the participants used during encoding. In future research, it would be beneficial to give participants specific instructions about how to encode the out-of-context words, so that their processing of the words would be similarly constrained across both study context conditions. By manipulating the encoding instructions, it should be possible to alter the resulting patterns of memory errors in very specific ways, which would further strengthen the findings from the present study.

The pattern of errors found in this study can explain the larger pattern of results evident in previous research on false memories for words. The results of studies finding high rates of conjunction errors (Jones & Jacoby, 2001; Reinitz et al., 1992; Underwood et al., 1976; Underwood & Zimmerman, 1973) suggest that the surface forms of words are maintained in memory, and that little semantic information is retained to contradict the sense of familiarity produced by lures that are visually similar to studied words. On the other hand, studies of semantic errors (Bransford & Franks, 1971; Brewer, 1977; Johnson et al., 1973) suggest that only gist information is stored in memory, and that information about the surface forms of words is discarded. Although these results seem to be discrepant, this pattern can be explained by taking into account the types of study materials that have been used in these two different sets of experiments. The studies that found high rates of conjunction errors typically used lists of out-of-context words, whereas those that found semantic errors typically used sentence or story contexts.

As we have shown in the present study, this difference in study context plays a crucial role in how the studied items are encoded. The encoding strategy that participants use when they encounter out-of-context words leads them

to focus relatively more on the exact form of the word and less on its meaning, making the participants less able to reject conjunction errors at test. Conversely, the encoding strategy that participants adopt when studying sentences or stories leads them to encode the gist of the sentences and relatively little information about their forms, making the participants less able to reject semantic lures at test. These different encoding strategies, which arise from the participants' day-to-day experiences with language, lead them to exhibit different patterns of memory errors in experiments with different study contexts.

An exception to this pattern comes from experiments using the Deese/Roediger–McDermott (DRM) paradigm (Roediger & McDermott, 1995), in which high numbers of semantically related false alarms are found in response to lists of out-of-context words. However, the strong semantic relationships between the words in the study-list context are likely to promote semantic processing for the words in a way that studying a list of unrelated words does not. Additionally, the inclusion of phonological associates in DRM lists has been found to greatly increase the number of false memories (Watson, Balota, & Roediger, 2003), leading to errors similar to those seen in studies using conjunction lures. In light of these findings, it seems that the DRM paradigm creates a study context that is like a word list in some ways and like sentence processing in other ways. The close relationships among the words promote semantic processing, making semantically related lures difficult to reject (or easy to generate, as in the case of recall tests). The absence of a larger context for the studied words, such as a sentence or story, also promotes attention to the word forms. Encoding this information may help the participants reject semantic lures at test, but the high number of related words in the DRM lists makes this rejection difficult. Instead, the encoded word form information can make the participants susceptible to form-based memory errors, as well as to semantic errors.

We do not wish to argue that retrieval processes are unimportant with respect to the production of memory errors. However, it is also important to take encoding strategies into account, because they affect what kinds of information are available for retrieval. For example, many previous studies have accounted for conjunction error data using dual-process models (see Jones & Jacoby, 2001; Marsh, Hicks, & Davis, 2002). In this account, conjunction errors occur when familiarity is unopposed by recollection. If the lure at test is a recombined word, it may seem familiar because its syllables had appeared in other words during the study phase. If the participant cannot remember the words that those syllables had actually appeared in, this sense of familiarity could lead to an endorsement of the lure. Manipulations that decrease recollection, such as dividing attention at study (Jones & Jacoby, 2001; Odegard & Lampinen, 2005), imposing a response deadline (Jones & Jacoby, 2001), or placing studied items into very similar contexts (Marsh et al., 2002; Reinitz & Hannigan, 2004; Underwood et al., 1976), consistently increase conjunction error rates. Our present experiments indicate that manipulations of study context also influence error rates by changing what information is encoded and, therefore, what

types of lures seem familiar. When words are presented in sentence contexts rather than in a list, participants engage in gist processing and encode less information about the specific words in the sentences. With this type of information encoded in memory, conjunction lures will not seem familiar at test, and conjunction error rates will go down, as demonstrated in our experiments.

In addition to determining what sorts of information seem familiar, study strategies and the nature of the encoded information also affect recollection attempts. The process of recollection rejection, in which participants are able to reject a lure by recalling its parent item, has been widely studied, with respect to conjunction lures and other types of false memories (Brainerd & Reyna, 2002; Brainerd, Reyna, Wright, & Mojardin, 2003; Hintzman, Curran, & Oppy, 1992; Lampinen, Odegard, & Neuschatz, 2004). When participants are successfully able to recall the originally studied words, they are also much more successful at rejecting conjunction lures (Lampinen et al., 2004; Odegard & Lampinen, 2005; Odegard et al., 2005). Factors such as study strategies that influence encoding can also influence recollection rejection by making the relevant information about the studied items more or less difficult to recall. For example, Lloyd (2007) found that pairing compound words with pictures during study reduced the rate of conjunction errors. Including a picture of the object named by the word provided a richer context, which made participants more likely to remember the word's meaning. This in turn made them less susceptible to conjunction lures. In the present study, placing compound words into sentences may have had much the same effect. If the participants were able to recall the gists of the studied sentences, they could have eliminated lures that did not seem to fit with the gist of any particular sentence. For example, when presented the lure *blackbird*, a participant may be able to determine that there were no sentences on the study list about birds, which would enable him or her to reject the lure.

Building models of retrieval processes is clearly very important for understanding memory errors, but more attention to factors that affect encoding is needed. Our experiments have shown that taking encoding differences into account can resolve some discrepancies in the memory error literature. A closer look at the interplay between encoding and retrieval could also serve to strengthen existing theories of false memory.

In summary, study context plays an important role in determining the motivational factors that influence what information people remember about studied words. Learners adopt different strategies when confronted with meaningful sentences than when presented with a list of unrelated words. When a word is presented in a larger context, such as a sentence or story, people are unlikely to remember the exact features of the word. Instead, they encode the gist of the whole item, a strategy that is less demanding and also likely to be successful in most language processing settings. This strategy makes people more susceptible to errors when they are presented with lures that are semantically similar to words in the original sentences or stories. In a real-world language processing task, such as listening to a conversation, this is unlikely to be problematic, because the basic meaning of the message would be unchanged. Additionally, previous research has found that there are cues in normal language processing situations that can direct the comprehender's attention to the surface features of the words when such attention is necessary. For example, Birch and Garnsey (1995; see also Fraundorf, Watson, & Benjamin, 2008) found that listeners had better memory for the exact forms of words that were prosodically focused than they had for words that were not focused. This indicates that listeners generally process the gist of the sentences they hear, but that they can change their processing strategy and shift their attention to the surface features of the words if the speaker indicates that those words are particularly important. This same sort of change in strategy can account for the high number of conjunction errors found in memory experiments. The present experiments demonstrate that people are highly flexible in their study strategies, and that when presented out-of-context words, they change processing strategies to fit the context. In a situation in which extracting meaning is difficult, people focus more on the surface structures of the words. These structures are subsequently retained in memory and can influence performance at test. The change in study context leads to a change in strategy, and this in turn changes what people are able to encode and remember for later use.

REFERENCES

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., et al. (2007). The English Lexicon Project. *Behavior Research Methods*, **39**, 445-459.

Banks, W. P. (2000). Recognition and source memory as multivariate decision processes. *Psychological Science*, **11**, 267-273. doi:10.1111/1467-9280.00254

Bayen, U. J., Nakamura, G. V., Dupuis, S. E., & Yang, C.-L. (2000). The use of schematic knowledge about sources in source monitoring. *Memory & Cognition*, **28**, 480-500.

Benjamin, A. S. (2001). On the dual effects of repetition on false recognition. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **27**, 941-947. doi:10.1037/0278-7393.27.4.941

Benjamin, A. S. (2008). Memory is more than just remembering: Strategic control of encoding, accessing memory, and making decisions. In A. S. Benjamin & B. H. Ross (Eds.), *Skill and strategy in memory use: The psychology of learning and motivation* (Vol. 48, pp. 175-223). San Diego: Academic Press.

Benjamin, A. S., & Bawa, S. (2004). Distractor plausibility and criterion placement in recognition. *Journal of Memory & Language*, **51**, 159-172. doi:10.1016/j.jml.2004.04.001

Benjamin, A. S., Diaz, M. L., & Wee, S. (in press). Signal detection with criterion noise: Applications to recognition memory. *Psychological Review*.

Birch, S. L., & Garnsey, S. M. (1995). The effect of focus on memory for words in sentences. *Journal of Memory & Language*, **34**, 232-267. doi:10.1006/jmla.1995.1011

Bock, J. K., & Brewer, W. F. (1974). Reconstructive recall in sentences with alternative surface structures. *Journal of Experimental Psychology*, **103**, 837-843. doi:10.1037/h0037391

Brainerd, C. J., & Reyna, V. F. (1998). Fuzzy-trace theory and children's false memories. *Journal of Experimental Child Psychology*, **71**, 81-129. doi:10.1006/jecp.1998.2464

Brainerd, C. J., & Reyna, V. F. (2001). Fuzzy-trace theory: Dual processes in memory, reasoning, and cognitive neuroscience. In H. W. Reese & R. Kail (Eds.), *Advances in child development and behavior* (Vol. 6, pp. 41-100). San Diego: Academic Press.

Brainerd, C. J., & Reyna, V. F. (2002). Recollection rejection: How children edit their false memories. *Developmental Psychology*, **38**, 156-172. doi:10.1037/0012-1649.38.1.156

Brainerd, C. J., Reyna, V. F., Wright, R., & Mojardin, A. H. (2003). Recollection rejection: False-memory editing in children and adults. *Psychological Review*, **110**, 762-784. doi:10.1037/0033-295X.110.4.762

Bransford, J. D., & Franks, J. J. (1971). The abstraction of linguistic ideas. *Cognitive Psychology*, **2**, 331-350. doi:10.1016/0010-0285(71)90019-3

Brewer, W. F. (1977). Memory for the pragmatic implications of sentences. *Memory & Cognition*, **5**, 673-678.

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology*, **58**, 17-22. doi:10.1037/h0046671

Egan, J. P. (1958). *Recognition memory and the operating characteristic* (Tech. Note AFCRC-TN-58-51). Bloomington: Indiana University, Hearing and Communication Laboratory.

Fraundorf, S. H., Watson, D. G., & Benjamin, A. S. (2008, November). Effects of prosodic stress on memory in language comprehension. *Abstracts of the Psychonomic Society*, **13**, 59-60.

Gallo, D. A., & Roediger, H. L., III (2002). Variability among word lists in eliciting memory illusions: Evidence for associative activation and monitoring. *Journal of Memory & Language*, **47**, 469-497. doi:10.1016/S0749-596X(02)00013-X

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Hastie, R., & Kumar, P. A. (1979). Person memory: Personality traits as organizing principles in memory for behaviors. *Journal of Personality & Social Psychology*, **37**, 25-38. doi:10.1037/0022-3514.37.1.25

Hicks, J. L., & Cockman, D. W. (2003). The effect of general knowledge on source memory and decision processes. *Journal of Memory & Language*, **48**, 489-501. doi:10.1016/S0749-596X(02)00537-5

Hintzman, D. L., Curran, T., & Oppy, B. (1992). Effects of similarity and repetition on memory: Registration without learning? *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **18**, 667-680. doi:10.1037/0278-7393.18.4.667

Hirshman, E. (1995). Decision processes in recognition memory: Criterion shifts and the list-strength paradigm. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **21**, 302-313. doi:10.1037/0278-7393.21.2.302

Israel, L., & Schacter, D. L. (1997). Pictorial encoding reduces false recognition of semantic associates. *Psychonomic Bulletin & Review*, **4**, 577-581.

Johnson, M. K., Bransford, J. D., & Solomon, S. K. (1973). Memory for tacit implications of sentences. *Journal of Experimental Psychology*, **98**, 203-205. doi:10.1037/h0034290

Jones, T. C., & Jacoby, L. L. (2001). Feature and conjunction errors in recognition memory: Evidence for dual-process theory. *Journal of Memory & Language*, **45**, 82-102. doi:10.1006/jmla.2000.2761

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

Lampinen, J. M., Odegard, T. N., & Neuschatz, J. S. (2004). Robust recollection rejection in the memory conjunction paradigm. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **30**, 332-342. doi:10.1037/0278-7393.30.2.332

Lloyd, M. E. (2007). Metamemorial influences in recognition memory: Pictorial encoding reduces conjunction errors. *Memory & Cognition*, **35**, 1067-1073.

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. I, pp. 103-189). New York: Wiley.

Marsh, R. L., Hicks, J. L., & Davis, T. T. (2002). Source monitoring does not alleviate (and may exacerbate) the occurrence of memory conjunction errors. *Journal of Memory & Language*, **47**, 315-326. doi:10.1016/S0749-596X(02)00005-0

Miller, M. B., & Wolford, G. L. (1999). Theoretical commentary: The role of criterion shift in false memory. *Psychological Review*, **106**, 398-405. doi:10.1037/0033-295X.106.2.398

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning & Verbal Behavior*, **16**, 519-533. doi:10.1016/S0022-5371(77)80016-9

Murnane, K., & Shiffrin, R. M. (1991). Word repetitions in sentence recognition. *Memory & Cognition*, **19**, 119-130.

Odegard, T. N., & Lampinen, J. M. (2005). Recollection rejection: Gist cuing of verbatim memory. *Memory & Cognition*, **33**, 1422-1430.

Odegard, T. N., Lampinen, J. M., & Toglia, M. P. (2005). Meaning's moderating effect on recollection rejection. *Journal of Memory & Language*, **53**, 416-429. doi:10.1016/j.jml.2005.04.004

Payne, B. K., Jacoby, L. L., & Lambert, A. J. (2004). Memory monitoring and the control of stereotype distortion. *Journal of Experimental Social Psychology*, **40**, 52-64. doi:10.1016/S0022-1031(03)00069-6

Payne, D. G., Elie, C. J., Blackwell, J. M., & Neuschatz, J. S. (1996). Memory illusions: Recalling, recognizing, and recollecting events that never occurred. *Journal of Memory & Language*, **35**, 261-285. doi:10.1006/jmla.1996.0015

Peterson, W. W., Birdsall, T. G., & Fox, W. C. (1954). The theory of signal detectability. *Transactions of the IRE Professional Group on Information Theory*, **4**, 171-212. doi:10.1109/TIT.1954.1057460

Potter, M. C., & Lombardi, L. (1990). Regeneration in the short-term recall of sentences. *Journal of Memory & Language*, **29**, 633-654. doi:10.1016/0749-596X(90)90042-X

Reinitz, M. T., & Hannigan, S. L. (2004). False memories for compound words: Role of working memory. *Memory & Cognition*, **32**, 463-473.

Reinitz, M. T., Lammers, W. J., & Cochran, B. P. (1992). Memory-conjunction errors: Miscombination of stored stimulus features can produce illusions of memory. *Memory & Cognition*, **20**, 1-11.

Roediger, H. L., III, & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **21**, 803-814. doi:10.1037/0278-7393.21.4.803

Shiffrin, R. M., Huber, D. E., & Marinelli, K. (1995). Effects of category length and strength on familiarity in recognition. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **21**, 267-287. doi:10.1037/0278-7393.21.2.267

Swets, J. A. (1986). Form of empirical ROCs in discrimination and diagnostic tasks: Implications for theory and measurement of performance. *Psychological Bulletin*, **99**, 181-198. doi:10.1037/0033-2909.99.2.181

Underwood, B. J., Kapelak, S. M., & Malmi, R. A. (1976). Integration of discrete verbal units in recognition memory. *Journal of Experimental Psychology: Human Learning & Memory*, **2**, 293-300. doi:10.1037/0278-7393.2.3.293

Underwood, B. J., & Zimmerman, J. (1973). The syllable as a source of error in multisyllable word recognition. *Journal of Verbal Learning & Verbal Behavior*, **12**, 701-706. doi:10.1016/S0022-5371(73)80050-7

Watson, J. M., Balota, D. A., & Roediger, H. L., III (2003). Creating false memories with hybrid lists of semantic and phonological associates: Over-additive false memories produced by converging associative networks. *Journal of Memory & Language*, **49**, 95-118. doi:10.1016/S0749-596X(03)00019-6

NOTE

1. These figures are more typically called receiver- (or relative-) operating characteristics. We have chosen to use the transparent nomenclature of Luce (1963; see also Benjamin, Diaz, & Wee, in press).