

Running Head: SDT AND WARNING EFFECTS

Metamnemonic Control Over the Discriminability of Memory Evidence: A Signal-Detection
Analysis of Warning Effects in the Associative List Paradigm

Jeffrey J. Starns

Sean M. Lane

Jill D. Alonzo

Cristine C. Roussel

Louisiana State University

In Press – *Journal of Memory and Language*

Address Correspondence:

Sean M. Lane
Department of Psychology
Louisiana State University
Baton Rouge, LA 70810
(225) 578-4098 (office: voice mail)
(225) 578-4125 (fax)
E-mail: slane@lsu.edu

Abstract

According to signal detection theory (SDT), retrieval warnings may decrease false memory in the associative list paradigm either by inducing a conservative criterion shift or by decreasing the amount of evidence that critical theme words were studied. Fitting a SDT model to 12 existing datasets revealed suggestive evidence that warnings impact critical theme evidence, and two new experiments confirmed this conclusion. We argue that this pattern of results is consistent with warned participants relying less on relational and more on item-specific forms of information at retrieval as compared to unwarned participants. We conclude that warnings enhance metamnemonic awareness, thus allowing participants to select a retrieval strategy that capitalizes on discriminative forms of evidence.

Keywords: signal detection theory; false memory; warning

Metamnemonic Control Over the Discriminability of Memory Evidence: A Signal-Detection
Analysis of Warning Effects in the Associative List Paradigm

Memories sometimes do not correspond to any objective past experience, and a wealth of memory research has explored the factors contributing to false memories (Roediger, 1996). The associative-list paradigm (also known as the “DRM” paradigm; Deese, 1959; Roediger & McDermott, 1995) has been adopted by many researchers as an effective way to induce false memories in a laboratory setting. In this paradigm, participants study lists of words that are organized in terms of their shared association to a non-presented word called the critical theme of the list. On a recognition test, participants are very likely to claim to have studied critical themes, and sometimes show little or no ability to discriminate non-presented critical themes from studied words (e.g., Roediger & McDermott, 1995).

One significant goal of false memory research is to identify strategies that people can use to avoid or edit false memories at retrieval. Some studies in the associative list paradigm have pursued this goal by warning participants about the nature of the paradigm just before a memory test (Anastasi, Rhodes, & Burns, 2000; Gallo, Roberts, & Seamon, 1997; Gallo, Roediger, & McDermott, 2001; McCabe & Smith, 2002; Neuschatz, Payne, Lampinen, & Tolia, 2001). Warnings typically inform participants that the lists that they studied were highly associated to non-presented words, and that they should be careful not to falsely remember these words. It is also typical to provide participants with an example of an associative list and the non-presented critical theme. Warnings enhance participants’ metamnemonic knowledge regarding both the task presented to them (i.e., that the recognition test will contain non-presented words that are highly associated to many studied items) and a property of their memory systems (i.e., that they are likely to experience illusory memories for associated words).

Whereas unwarned participants are likely to greatly underestimate the difficulty of a test requiring the discrimination of studied items from critical theme lures, warned participants have the opportunity to form more realistic expectations about the test. Warned participants may be able to translate their increased metamnemonic awareness into a more effective retrieval strategy, resulting in a heightened ability to distinguish true from false memories.

The results of prior research employing a retrieval warning have been somewhat mixed. Some studies suggest that warnings are only effective when they are provided before encoding, but not after encoding and before retrieval (e.g., Gallo et al., 1997). Other studies show a robust effect of retrieval warnings on false recognition (e.g., McCabe & Smith, 2002). Identifying the factors that contribute to warning effectiveness, as well as the mechanisms of false memory reduction following an effective warning, will provide useful information regarding how retrieval strategies promote or discourage false memories.

The purpose of this paper is to clarify the effects of warnings on false memory by applying an appropriate signal detection model. According to Signal Detection Theory (SDT; Wickens, 2002), recognition decisions are determined by both the evidence stored in memory and the decision processes that are used to translate this evidence into specific responses. The memory evidence used in recognition decisions is represented as a single continuous variable that is usually called familiarity.¹ Both targets and lures vary in the amount of familiarity they inspire, but targets are more familiar on average based on the memory evidence encoded for these items in the study phase. It is typically assumed that the familiarity values of both targets and lures are normally distributed with the mean of the target distribution above the mean of the lure distribution on the familiarity continuum. Also, the variance of the target distribution is regularly greater than the variance of the lure distribution (Glanzer, Adams, Iverson, & Kim

1993). μ is the distance between the means of the target and lure distributions measured in terms of the lure distribution's standard deviation, and this parameter provides a measure of the gain in memory evidence resulting from the encoding episode. σ is the ratio of the target and lure distributions' standard deviations. When the target and lure distributions have equal standard deviations (i.e., $\sigma = 1$), μ is equal to the commonly used memory measure d' . To make recognition decisions, participants set a criterion for the amount of memory evidence they require to claim that an item was studied, and any test item that exceeds this criterion value receives a positive recognition response. The parameter λ expresses the distance of the response criterion from the mean of the lure distribution in terms of the lure distribution's standard deviation.

Applying SDT to the associative list paradigm necessitates a decision space containing three distributions for the three item types on the test (targets, lures, and critical themes; Wixted & Stretch, 2000). Thus, the model requires five parameters to describe recognition performance: μ_T , the distance between the unrelated lure and target distributions; σ_T , the standard deviation of the target distribution relative to the unrelated lure distribution; μ_{CT} , the distance between the unrelated lure and critical theme distributions; σ_{CT} , the standard deviation of the critical theme distribution relative to the unrelated lure distribution; and λ , the position of the response criterion. The μ_{CT} parameter reflects the gain in memory evidence for critical theme words as a result of the presentation of associates in the study phase. This model is graphically displayed in the top panel of Figure 1. For simplicity, this figure displays a situation in which the standard deviations of all distributions are equal.

A signal detection analysis reveals that there are two ways that retrieval warnings can reduce false memory for critical theme words: warnings may lead to a higher response criterion

or warnings may decrease the distance between the unrelated lure and critical theme distributions. A change in response criterion would indicate that warnings prompt participants to use more conservative standards for the evidence required to claim that a retrieval candidate was studied. A change in the position of the critical theme distribution would indicate that warned participants actually retrieve less evidence that critical themes were studied than do unwarned participants. Of course, warnings can simultaneously impact both memory evidence and response criteria, but we will separately consider the implications of a distribution shift and a criterion shift for clarity of exposition.

A criterion shift explanation is displayed in Figure 1. The top panel in this figure shows recognition performance without a warning, and the bottom panel shows recognition performance following a warning. In the figure, a warning induces a shift in response criterion to a higher, more conservative value. This would result if, following a warning, participants decided to avoid errors by requiring a great deal of evidence to claim that an item was studied. In Figure 1, a smaller proportion of the critical theme distribution exceeds the response criterion for the warned condition, resulting in a lower false alarm rate for critical theme words. However, the criterion shift also affects responding for unrelated lures and targets; that is, participants make fewer “yes” responses for all item types.

An explanation of warning effects based on a shift in the position of the critical theme distribution is displayed in Figure 2. According to this explanation, when a warning is provided, the critical theme distribution moves closer to the unrelated lure distribution. As a result, the proportion of the critical theme distribution that exceeds the response criterion decreases. In contrast to the criterion shift explanation, the distribution shift explanation predicts no change in

performance for unrelated lures from the unwarned to warned conditions as well as no change in target performance (assuming that the warning does not affect target discriminability).

The position of the distributions depends on the amount of evidence in memory, so the prediction that the critical theme distribution shifts in response to a variable that is introduced after all encoding procedures have been completed seems unintuitive. However, a distribution shift is possible under the assumption that the amount of evidence retrieved from memory will differ when different types of evidence are considered (Johnson, Hashtroudi, & Lindsay, 1993). Memory traces are blends of various types of information, including conceptual information, perceptual information, and information regarding the cognitive operations or affective states that accompanied encoding procedures (Johnson et al., 1993; Johnson, Foley, Suengas & Raye, 1988). For our analysis, we follow Hunt and Einstein (1981; see also Hunt & McDaniel, 1993) by classifying the various types of information available from memory as either relational or item-specific information. Relational information describes information that is common among a set of items, such as the shared conceptual relationship of each word in an associative list to the critical theme. Item-specific information describes information that is unique to individual items within a set. Perceptual information, such as the orthographic or phonological features of list words, would be considered item-specific information in the associative list paradigm. Other examples of item-specific information include information regarding item-specific elaboration or imagery generated during encoding.

Unwarned participants may heavily rely on relational information in making their decisions; that is, when a test word is consistent with the meaning of the studied words, they regard this as evidence that the test word was studied. Thus, critical theme words, which are highly related to many studied items, should inspire much higher evidence values than unrelated

lures. When participants are warned about the presence of lures that are highly related to the studied words, they may rely less on relational information and more on item-specific information (cf., Neuschatz et al., 2001; Reyna & Kiernan, 1994). In this situation, the evidence retrieved for critical themes should be more similar to that of unrelated lures, because little item-specific information should be available for either item type (Mather, Henkel, & Johnson, 1997; Norman & Schacter, 1997). We will refer to the idea that warned participants reduce critical theme familiarity by altering the types of evidence that they try to retrieve as the *evidence-change hypothesis*.

The criterion shift and distribution shift explanations appear easy to discriminate: the criterion shift explanation predicts that a warning should simultaneously affect critical themes, unrelated lures, and targets, whereas the distribution shift explanation predicts that warnings should uniquely affect critical themes. Some previous experiments employing retrieval warnings demonstrated reductions in the false alarm rate for critical theme words with little or no change in hit rate or unrelated lure false alarm rate (e.g., McCabe & Smith, 2002). Such results seem to be wholly inconsistent with the criterion shift explanation; however, closer inspection reveals that there are several situations in which a criterion shift could influence critical themes to a much greater extent than the other item types. For example, when the critical theme distribution lies between the unrelated lure and target distributions, a criterion shift may “cut off” a much larger proportion of the critical theme distribution than the other distributions. Such a situation is displayed in the top panel of Figure 3. In this figure, the line to the left represents the response criterion without a warning, and the line to the right represents the response criterion with a warning. The proportion of each distribution that lies between the two lines gives the magnitude of the decrease in responding created by the conservative criterion shift. The top panel displays

that a much larger proportion of the critical theme distribution falls between the two criteria compared to either the target or unrelated lure distributions. In fact, the criterion shift displayed in this panel creates a .16 decrease in the false alarm rate for critical theme words (.58 to .42) compared to a .02 decrease in both the hit rate and unrelated lure false-alarm rate. With experimental error added in, the differences in responding for targets and unrelated lures would be very unlikely to be detected statistically; thus, simple analyses on hit and false alarm rates would lead to the conclusion that the warning affected critical themes and no other item type. Thus, in some situations, a pattern of data created by a criterion shift is likely to be interpreted in favor of a distribution shift if hit and false alarm rates are analyzed without an appropriate performance model.

For comparison, the bottom panel of Figure 3 displays a situation in which the positions of the critical theme and target distributions are similar. When this is the case, a similar proportion of the critical theme and list item distributions lie between the criteria. Indeed, the criterion shift displayed results in a .07 drop in responding to critical themes and a .06 drop for list items. Thus, comparing the panels of Figure 3 reveals how the relative position of the distributions influences the extent to which a criterion shift will differentially affect different item types. A model that keeps track of the position of the distributions is needed to determine if a given pattern of data can be interpreted as a criterion shift.

The relative effect of a criterion shift on different item types also depends on the variability in their distributions. Figure 4 displays a situation in which the critical theme distribution is more variable than the unrelated lure distribution, and the target distribution is more variable than the critical theme distribution. The two criteria in Figure 4 display a conservative shift in response to a warning, and, again, the proportion of each distribution

between the two criteria shows the drop in responding induced by a warning. The displayed criterion shift reduces responding for critical themes more than twice as much as it reduces responding for targets or unrelated lures. Notably, the criterion shift has a much larger effect on critical themes than on list items even though responding to these two item types is similar in the unwarned condition: .79 of the target distribution passes the unwarned criteria compared to .73 of the critical theme distribution. A model that assumes that all distributions have the same variability would have to place the critical theme distribution near the target distribution to explain the similar level of responding to these item types. Thus, an equal variance model would predict that a criterion shift should similarly affect the hit rate and critical theme false alarm rate. In contrast, the criterion shift displayed in Figure 4 decreases the critical theme false alarm rate to .57 (from .73) while the hit rate only decreases to .71 (from .79). Thus, Figure 4 demonstrates that the relative variability of the distributions must be considered in addition to their relative positions to determine whether or not a given pattern of hit and false alarm rates can be explained as a criterion shift.

We sought to determine if reductions in false memory following a warning reflect a conservative shift in response criterion, a change in the position of the critical theme distribution, or both. We pursued this goal in two ways. First, we fit a signal detection model to a number of existing datasets in the retrieval warning literature. Second, we collected new data that allowed a more comprehensive application of the model than any of the existing datasets. Evaluating warning effects allowed us to explore which aspects of recognition performance fall under metacognitive control. A wealth of evidence indicates that participants can strategically control the amount of evidence they require to decide that an event was experienced; for example, participants adjust their response criterion based on the relative payoffs and penalties associated

with “yes” and “no” responses (e.g., Koriat & Goldsmith, 1994,1996). Similarly, participants may strategically increase their criterion in response to a warning that highlights the fallibility of their memory. In contrast, finding a critical-theme distribution shift would indicate that the amount of evidence retrieved from memory is also subject to metamnemonic control. The evidence-change hypothesis details how participants can alter the amount of evidence retrieved from memory by selecting the types of information that are sought from memory. Such a finding would be consistent with research that suggests that participants cue their memories differently depending on test instructions (e.g., Herron & Rugg, 2003; Jacoby, Shimizu, Daniels, & Rhodes, 2005).

Fits to Existing Data

Datasets and Fitting Procedure. We fit the signal detection model for associative lists to 12 comparisons of warned versus unwarned recognition performance.² Datasets 1-4 were taken from Neuschatz et al.’s (2001) Experiment 3, and the proportions in these datasets were taken from the immediate test/moderate warning, immediate test/strong warning, delayed test/moderate warning, and delayed test/strong warning conditions, respectively. Datasets 5 and 6 were taken from Gallo et al.’s (2001) conditions in which some critical themes were studied with their lists (dataset 5) and in which no critical themes were studied (dataset 6). The critical theme proportions for dataset 5 were taken only from critical themes not presented in the study phase. Dataset 7 is Anastasi et al.’s (2000) Experiment 3. Dataset 8 consists of the control and cautious conditions from Gallo et al. (1997). Datasets 9-12 were taken from McCabe and Smith (2002). Datasets 9 and 10 are the 4-second and 2-second encoding conditions from the young adult participants in Experiment 1, and datasets 11 and 12 are the corresponding conditions in Experiment 2.

We translated the results of each study into response frequencies by multiplying the proportion of “yes” responses to each type of item by the total frequency of that item type (i.e., the number of items of that type on the test multiplied by the number of participants in each group).³ Although many of the studies included a variety of item types on the test, we chose to model the results only for unrelated lures, critical themes, and targets. We focus on these item types because they are consistently used in all of the existing warning studies and in the present investigation.

As noted, the SDT model uses 5 parameters to fit performance from a given condition. Each data set has three independent response frequencies per condition, i.e., the number of targets, unrelated lures, and critical themes called “studied.” Therefore, the number of parameters exceeds the number of independent response frequencies, and the full 5 parameter model is not identifiable. We chose to reduce the parameter space within each condition to 3 parameters by assuming that the standard deviations of the critical theme and target distributions were equal to the standard deviation of the unrelated lure distribution, eliminating σ_T and σ_{CT} as free parameters. This equal-variance assumption is common in applications of SDT to recognition memory (e.g., Dunn, 2004); however, the Introduction details how differences in variability among the distributions moderate the relative influence of a criterion shift for different item types. The model fits to existing data cannot accommodate this influence. Because we are using an equal variance model, we will notate the distance between evidence distributions as d' as opposed to the more general μ .

Maximum-likelihood estimation was used to find the parameter values that produced a set of predicted response frequencies that most closely matched the response frequencies in each dataset. The criterion-shift and distribution-shift explanations of warning effects were evaluated

based on the model parameters. As evident in Figure 1, the criterion-shift explanation predicts that criterion values (λ) should be higher in warned than in unwarned conditions. Figure 2 shows that the distribution-shift explanation predicts that d'_{CT} values should be lower in warned than in unwarned conditions. We used the G^2 statistic to test the significance of potential criterion and d'_{CT} differences in each dataset. G^2 indexes the overall match of the response frequencies produced by the model to the empirical response frequencies, with smaller values indicating a better match. We began with a full model in which all parameters (d'_T , d'_{CT} , and λ) varied freely for the warned and unwarned conditions. To test the significance of differences in criterion between the warning conditions, we evaluated the change in G^2 from the full model to a model in which the response criterion was constrained to be equal for the warned and unwarned conditions. If the warned and unwarned groups adopt different response criteria, then a model that assumes equal criteria should yield a poor fit to the data in comparison to a model that can accommodate criterion differences. We similarly tested the significance of differences in d'_{CT} between the warning conditions by comparing the fit of the full model to a model in which d'_{CT} was constrained to be equal for the warned and unwarned conditions. If warnings influence the amount of memory evidence retrieved for critical theme words, then a model that assumes the same critical theme discriminability in the warned and unwarned conditions should provide a poor fit to the data in comparison to a model that can accommodate memory differences for critical themes. The G^2 statistics for criterion and d'_{CT} effects were computed by subtracting the G^2 of the full model from the G^2 of the appropriate reduced model, and the significance of the difference in fit was evaluated on a χ^2 distribution with one degree of freedom.

Fitting Results. Table 1 displays the model parameters and G^2 tests for each of the twelve datasets. The d'_{CT} parameters show a consistent effect of warnings on critical theme

memory: critical theme discriminability is lower in the warned than the unwarned condition in ten of the twelve datasets. The G^2 statistics show that equating the d'_{CT} parameters across warning conditions led to a significant reduction in fit for five datasets. In every dataset showing a significant effect of warnings on d'_{CT} , retrieval warnings decreased critical theme discriminability. The λ parameters reveal a somewhat inconsistent pattern: providing a warning had no effect on the criterion in one dataset, led to a conservative shift in seven datasets, and led to a liberal shift in four datasets. The criterion differences in five datasets reached significance, three of which showed a significantly higher criterion for warned participants and two of which showed a significantly lower criterion. Thus, the SDT results from existing studies appear inconsistent with the criterion-shift explanation. We compared the procedural details of the studies showing conservative versus liberal criterion shifts following a warning to determine if the difference in results was systematically related to any methodological differences. The procedural details we considered included the amount of time that each item was studied, the number of studied associates for each critical theme word, the number of associative lists presented in the study phase, and the number of related lures appearing on the recognition test. We found no consistent differences in methodology to discriminate studies showing conservative versus liberal shifts, fortifying our conclusion that the criterion results are inconsistent and do not support the criterion-shift explanation.

The model fits to existing studies suggest that participants can exert metacognitive control over the amount of evidence experienced for false memories, although the decrease in critical theme familiarity induced by retrieval warnings reached significance for only five datasets. There could be several explanations for these equivocal results. One possibility is that warned participants really do experience less evidence in support of their false memories, but the

studies contributing datasets for the model fits did not have adequate power to consistently detect this effect. Another possibility is that, in the datasets apparently showing a critical theme distribution shift, the selective reduction in responding to critical themes following a warning could be explained as a criterion shift with a more comprehensive model. As described in the introduction, the relative effects of a criterion shift on different item types depend on both the position and variability of the underlying distributions. The model that was fit to the existing data could account for the effects of distribution position but not the effects of distribution variability. Figure 4 displays that, with certain variability relationships among the distributions, warned participants can disproportionately decrease critical theme responding merely by becoming more conservative. A model that assumes equal standard deviations among the distributions may be forced to shift distributions to accommodate selective reductions in critical theme responding. In contrast, a model that tracks variability differences among the distributions may be able to more flexibly model these selective reductions as a criterion shift, consequently eliminating the need to change critical theme memory parameters in response to the warning variable. This possibility will be investigated in the current experiments, which will use confidence ratings to permit the application of a complete model that can reveal the potential effects of unequal standard deviations.

Another possibility is that participants disproportionately decrease false memories following a warning by identifying test candidates as critical themes and rejecting them on this basis (see Neuschatz, Benoit, & Payne, 2003 for evidence that such a strategy may operate with pre-encoding warnings). Warned participants are instructed to carefully monitor their responses for a specific type of item on the test, that is, words that are highly related to all of the words on a given list. Thus, participants can reduce false memory if they can identify the test items that are

the special type of item mentioned in the warning. If participants could reject critical themes based on identifiability, it would seriously distort the results of signal detection analyses because the SDT model assumes that the only evidence determining recognition decisions is familiarity, but participants would also be using another type of evidence (i.e., degree of relatedness to multiple studied items) to assess whether or not each item is a critical theme. In other words, the datasets that appear to show an effect of warnings on memory processes may actually show that warnings induce an identification strategy without affecting the familiarity of theme words.

We conducted the following experiments to more definitively test the distribution-shift explanation of warning effects. The model fits to existing data suggest that less evidence supports false memories in the warned versus the unwarned conditions, but alternative hypotheses are able to explain the selective reduction in critical theme responding. We sought to discriminate the distribution-shift explanation from these alternatives. Experiments 1 and 2 used confidence ratings at test to increase the number of independent response frequencies and permit the application of a complete signal detection model including parameters for the standard deviation of the critical theme and target distributions. Thus, if differences in distribution variability underlie the selective reduction in critical theme responding for warned participants, the models applied to our experiments will be able to explain the reduction as a criterion shift without changing critical theme memory parameters. Experiment 2 addressed the identifiability hypothesis by creating a test in which every test candidate was a critical theme word: targets were critical themes studied with their lists, critical theme lures were critical themes not studied with their lists, and unrelated lures were critical themes from entire lists that did not appear in the study phase. Because every test item was a critical theme, participants were not able to reject test candidates by identifying them as the “organizing” word of one of the lists. If selective

reduction in critical theme responding relies on identification, then the selective reduction should not be observed in Experiment 2.

Experiment 1

In this experiment, participants studied associative lists and then completed a recognition test given either standard instructions or instructions that warned them about the false memory illusion. Participants in the warned condition were told that the lists they studied likely made them think about non-presented but highly associated words, and that they should be careful not to falsely recognize these words. Participants made recognition responses on a 4-point confidence scale ranging from “sure new” to “sure old.” The use of confidence ratings increased the number of independent response frequencies in each dataset, thus allowing us to fit a fully specified SDT model with free parameters for the standard deviation of the critical theme and target evidence distributions.

Method

Participants. Ninety-six Louisiana State University undergraduates participated to earn extra credit in their psychology courses. Participants were randomly assigned to the warning or no-warning conditions in equal numbers. One participant in the warning condition was excluded from the dataset for failing to follow instructions.

Design and Materials. This experiment conformed to a 3×2 factorial design, with item type (target, critical theme lure, unrelated lure) and instruction condition (warning, no warning) as factors. Item type was manipulated within-subjects, and instruction condition was manipulated between subjects.

We constructed 36 associative lists by selecting 6 high associates for each of 36 target words in the University of South Florida word-association norms (Nelson, McEvoy & Schreiber,

1998). The lists were divided into 3 sets of 12, and the lists within each set contributed either targets, critical theme lures, or unrelated lures to the recognition test. The set used for each item type was counterbalanced across participants. The lists assigned to contribute targets and critical theme lures to the recognition test were presented in the study phase, and the six items in each list were presented together in blocks. The order of the lists and the order of the items within each list were independently randomized for each participant. The recognition test consisted of one word from each of the 36 lists. The sixth word from 12 of the studied lists served as targets, the critical themes from the other 12 studied lists served as critical-theme lures, and the sixth word from each of the 12 unstudied lists served as unrelated lures.⁴

Procedure. Study instructions informed participants that they were going to see lists of words, and that they would be asked to recall the words following each list. Participants were also told that the single participant who achieved the best performance in the experiment would be awarded \$25 at the end of the semester. During the study phase, the six words from each list were presented for 2200 ms each. Participants heard each word in a female voice through headphones, and they simultaneously saw each word printed on the computer screen. Immediately following each list, participants were prompted to recall as many words as possible from the just-presented list. Participants recorded their responses in a recall booklet that had a new sheet for each list in the study phase. Participants recalled for 25 s following each list. At the end of the recall period, a tone cue was played through the headphones and a message on the screen asked participants to prepare for the next list. Although we were not interested in recall performance, we included recall tests for each list to ensure that participants were effectively encoding the words and to emphasize the thematic organization of the lists.

After the final list was recalled, participants were asked to work on a word-search puzzle that was printed on the final page of their recall booklets. The puzzle was an array of random letters with embedded names of US cities. Participants were instructed to find and circle as many city names as possible, and participants worked on the puzzle for one minute before hearing instructions for the memory test.

Test instructions informed participants in both conditions that they would see words that either were or were not presented to them in the study phase. For each word, they were told to respond to the question “Did you hear this word in the study phase?” by pressing an “SN” sticker (placed on the “S” key) to respond “Sure No,” a “GN” sticker (placed on the “F” key) to respond “Guess No,” a “GY” sticker (placed on the “H” key) to respond “Guess Yes,” and an “SY” sticker (placed on the “K” key) to respond “Sure Yes.” Both groups were also reminded that the participant achieving the highest level of performance would receive \$25 at the end of the semester. Participants in the warning condition were additionally informed about the nature of the false memory paradigm. They were told that lists like the ones that they studied often make people think of words that were not on the list but are highly related to the list words, and they were given the example of the words “mad,” “fear,” “hate,” and “rage” making a participant think of the non-presented word “anger” (the anger list was not in the stimulus set). They were also told that people are very likely to falsely recognize highly-related but non-presented words, and that they should avoid making this error on the upcoming test.

Participants in both groups proceeded through the recognition test at their own pace. When the test was finished, they read a debriefing statement and were allowed to ask questions about the experiment.

Results and Discussion

All statistical tests were conducted with a .05 probability of Type 1 error. For each experiment, we first report analyses on hit and false-alarm rates, and then evaluate the results of the signal detection model.

Table 2 displays the proportion of each item type called studied (i.e., receiving either a “Guess Yes” or “Sure Yes” response) in the warned and unwarned conditions for Experiments 1 and 2. The Experiment 1 data show that providing a warning before the recognition test reduced the critical theme false-alarm rate, but had little effect on the hit rate or the unrelated lure false-alarm rate. Independent-sample t-tests confirmed that warnings had a significant effect on critical theme false-alarm rate [$t(93) = 4.00, p < .05$], but warnings did not significantly affect hit rate [$t(93) = 1.21, ns$] or unrelated lure false-alarm rate [$t(93) = .77, ns$]. Floor effects may have prevented us from observing a decrease in the unrelated lure false alarm rate for warned participants, thus masking potential evidence of a criterion shift. Moreover, the Introduction describes several situations in which a criterion shift can dramatically affect the critical theme false-alarm rate with little change in hit rates or unrelated lure false-alarm rates. Therefore, we used the signal detection model to directly evaluate the distribution-shift and criterion-shift hypotheses.

We fit the SDT model using the total frequency of each rating scale response to each item type in the warned and unwarned conditions. The frequency table for this experiment contains 9 independent response frequencies in each condition; that is, the frequencies of “guess new,” “guess old,” and “sure old” responses to targets, critical themes, and unrelated lures (the “sure new” frequency is not independent because it is fixed once one knows the total frequency and the frequency of the other three responses). To fit the rating scale data, we used a model with three

response criteria: λ_1 was the criterion separating “sure new” from “guess new” responses, λ_2 was the criterion separating “guess new” from “guess old” responses, and λ_3 was the criterion separating “guess old” and “sure old” responses. The full model including standard deviation parameters for the critical theme and target distributions had 7 parameters within each condition (μ_T , σ_T , μ_{CT} , σ_{CT} , λ_1 , λ_2 , and λ_3). Thus, the model used fewer parameters than the available independent response frequencies, allowing us to evaluate the fit of the model to the empirical response frequencies. The μ parameters represent the mean of the target (μ_T) or critical theme (μ_{CT}) distributions, and they replace the d' parameters to reflect the fact that the distributions are no longer assumed to have the same variability. Just like d' , the μ parameters measure the distance of a distribution from the unrelated lure distribution in units of the unrelated lure distribution’s standard deviation. Unlike d' , the μ parameters do not stand alone as memory performance measures. Memory performance is jointly determined by the position and the variability of the underlying distributions, so we combined the μ and σ parameters into the measure A_z to achieve a single memory-performance measure (Wickens, 2002). In studies plotting receiver operating characteristics (ROCs), A_z corresponds to the area under the ROC curve, and A_z also equals the proportion of correct responses that would be achieved if a 2-alternative-forced-choice test were administered (Wickens, 2002). A_z varies from .5 when there is no discriminability between two item types to 1 when the item types can be perfectly discriminated.

We statistically tested whether warnings affected memory evidence for critical themes by comparing the fit of the full model to the fit of a model in which the discriminability of critical themes from unrelated lures was constrained to be equal in the warning and no-warning conditions. Equating critical theme memory across conditions required equating both the

position and standard deviation of the critical theme distributions; therefore, the constrained model had 2 fewer parameters than the full model and the test for differences in model fit had two degrees of freedom. We tested for warning effects on response criteria by comparing the fit of the full model to the fit of a model in which the three λ parameters were constrained to be equal across groups. Equating response criteria eliminated three free parameters; thus, the test for a difference in model fit had three degrees of freedom. The criterion shift explanation clearly predicts that the old/new criterion (λ_2) should be more conservative for warned than for unwarned participants. The other confidence criteria (λ_1 and λ_3) may be expected to follow the movement of the old/new criterion (see Stretch & Wixted, 1998, for a more detailed discussion of shifts in confidence criteria).

Table 3 displays the observed hit and false alarm rates at each response criterion as well as those predicted by the SDT model. Comparing the predicted and observed frequencies reveals that the model produced an impressive fit to the data: the largest deviation of a model prediction from an observed proportion is .01. Table 4 shows the best-fitting model parameters, as well as the A_z measures for true and false memory. The A_z measures reveal that warned participants discriminated targets from unrelated lures as effectively as unwarned participants. The discriminability of critical themes from unrelated lures was lower in the warned than the unwarned condition, indicating that warned participants retrieved less evidence in support of their false memories. Constraining the critical theme memory parameters to be equal between the instruction conditions led to a significant reduction in model fit, $G^2(2) = 17.70, p < .05$.

Equating response criteria for the warned and unwarned groups also significantly reduced model fit, $G^2(3) = 15.38, p < .05$. Table 4 reveals that warned participants employed a slightly more conservative old/new criterion (λ_2); therefore, a change in response criteria somewhat

contributed to the reduction in false memory for warned participants. However, this slight criterion shift alone could not have created the magnitude of false memory reduction displayed in Table 2, and the model results clearly indicate that the warning also decreased the memory evidence experienced for critical theme words. The criterion distinguishing high and low confidence “new” responses (λ_1) followed the old/new criterion in that it had a more conservative value for warned participants. However, the criterion separating high and low confidence “old” responses (λ_3) was actually slightly more liberal for warned than for unwarned participants.

Experiment 1 offers strong evidence in favor of the distribution-shift explanation. Warned participants were able to selectively reduce false memory for critical theme lures without significantly decreasing their responding to either targets or unrelated lures. Furthermore, the results of the signal detection model demonstrate that this selective reduction cannot be explained as a criterion shift. The old/new criterion was slightly more conservative for warned participants, but the reduction in false memory was too large to be explained as a mere shift in criterion. The Introduction shows that the selective impact of a criterion shift is based on the relative position and the relative standard deviation of the critical theme distribution; therefore, a model that takes these factors into consideration is needed to determine if a given pattern of hit and false alarm rates reflects a shift in distributions. Although the signal detection model for this experiment was free to adjust both the position and variability of the critical theme distribution, it could not mimic the observed pattern of data unless different false memory parameters were permitted in the warning and no-warning conditions. Thus, Experiment 1 rules out the possibility that the selective reduction in critical theme false-alarm rate results from a global criterion shift alone.

However, the results of Experiment 1 do not rule out the possibility that participants reject critical themes by identifying them as the special type of item mentioned in the warning. Subjects are able to discover the critical theme for a high proportion of associative lists when they are informed of the nature of the paradigm before encoding (McDermott & Roediger, 1998; Neuschatz et al., 2003), and it is possible that participants can also do this retrospectively when they are warned after encoding. We explored the possibility of an identification strategy in Experiment 2.

Experiment 2

The current experiment eliminated the possibility of an identification strategy by using a more extreme version of a technique developed by McDermott & Roediger (1998; see also Gallo et al., 2001). These researchers presented some critical themes in the study list, thus creating a situation in which the “special” items mentioned in the warning were sometimes targets. If participants reject items that are highly related to a number of studied words, this will affect not only the false-alarm rate to critical theme lures, but also the hit rate to critical theme targets. Thus, there would be no possibility of selectively reducing false memory without also impairing veridical memory. In Experiment 2, every item on the test was a critical theme word: targets were critical themes studied along with their lists, critical theme lures were critical themes not studied with their lists, and unrelated lures were critical themes from lists that were not studied.

Method

Participants. This experiment included 119 Louisiana State University undergraduates who did not participate in Experiment 1. Participants were randomly assigned to conditions, with 59 participants in the warning condition and 60 participants in the no-warning condition.

Design and Materials. The design of this experiment was identical to Experiment 1, and the same materials were used with the exception that only critical themes appeared on the recognition test. For the lists assigned to contribute targets, the critical theme of each list was included along with the other list words in the study phase, creating lists with seven items. Thus, critical themes from these lists were actually studied items. For lists assigned to contribute critical theme lures, the list words, but not the critical theme itself, appeared in the study phase. Thus, critical themes from these lists were not studied but were highly related to studied words. For lists assigned to contribute unrelated lures, neither the critical theme itself nor its list words appeared in the study phase. Thus, critical themes from these lists were not studied and were not highly related to any of the studied words. The recognition test consisted of the critical theme from each of the experimental lists.

Procedure. All procedural details matched those of the first experiment.

Results and Discussion

Inspection of Table 2 reveals that, as in the first experiment, providing a warning noticeably influenced responding for critical themes but had little effect on responding to targets or unrelated lures. The decrease in critical theme false-alarm rate for the warned participants was significant, $t(117) = 2.37, p < .05$. Warnings did not have a significant influence on either the hit rate [$t(117) = 1.38, ns$] or the unrelated lure false-alarm rate [$t(117) = .95, ns$]. Again, floor effects may have obscured the t-test results for the unrelated lure false alarm rate; however, warned participants were actually slightly more likely to false alarm to unrelated lures, providing little evidence that the warning decreased the unrelated lure false alarm rate as predicted by the criterion-shift explanation.

Table 3 shows that the signal detection model closely reproduced the data from Experiment 2: a single predicted proportion deviates from the observed proportion by .02, and all other deviations are .01 or less. Table 4 displays the parameter values of the best-fitting model. The A_z values show that warnings had almost no effect on the discriminability of targets from unrelated lures, but reduced the “false” discriminability of critical themes from unrelated lures. Model fit worsened significantly when critical theme memory was constrained to be equal across instruction conditions, $G^2(2) = 12.91, p < .05$. Assuming equal response criteria for the warned and unwarned groups did not significantly reduce model fit, $G^2(3) = 2.77, ns$. Thus, the conservative shift in the old/new criterion (λ_2) observed for warned participants in Experiment 1 was not replicated in Experiment 2; in fact, the old/new criterion was actually nominally more liberal for warned participants.

The results of Experiment 2 replicate the critical theme distribution shift found in the first experiment and eliminate the possibility that the selective reduction in the critical theme false-alarm rate is actually due to an identification strategy. If participants rejected words that seemed to capture the core concept of one of the studied lists, then warnings should have decreased positive recognition decisions for both critical theme targets and critical theme lures, as these stimulus classes were identical in every sense except their presentation status in the study phase. We observed decreased responding only for critical theme lures, which is inconsistent with the identification hypothesis.

General Discussion

Our results support the claim that retrieval warnings can change the amount of illusory memory evidence experienced for critical theme words, which is modeled as a shift in the critical theme distribution in signal detection theory. Fitting an equal variance model to 12 datasets

taken from existing literature on retrieval warnings revealed that the critical theme distribution was consistently closer to the unrelated lure distribution for warned versus unwarned participants, and this difference reached significance for five of the datasets. Experiment 1 showed that even a complete model that tracks the relative standard deviation of each distribution must move the critical theme distribution to accommodate the effects of a retrieval warning. Experiment 2 showed that the selective reduction in critical theme false alarm rate is not created by an identification strategy. Thus, our results support the conclusion that a critical theme distribution shift underlies participants' ability to reduce false memory following a retrieval warning. Interestingly, while preparing our manuscript we learned of a study employing SDT analyses in a paradigm in which warnings came before encoding (Westerberg & Marsolek, 2006), and results showed that pre-encoding warnings can also influence critical theme evidence distributions.

In contrast to the distribution-shift explanation, the hypothesis that warnings reduce false memory by promoting a more conservative response criterion was not consistently supported by either the fits to the existing studies or by our own experiments. Out of the 12 existing datasets, 3 showed a significantly more conservative response criterion for warned participants and 2 showed a significantly more *liberal* response criterion for warned participants. Our first experiment did show a significant effect of warning instructions on response criteria, and the old/new criterion was slightly more conservative for warned participants. However, the warning did not significantly affect response criteria in Experiment 2, and the old/new criterion was actually numerically more liberal for warned participants. Although warnings may sometimes induce more conservative responding, criterion shifts alone do not appear to fully explain the reductions in false memory that have been observed in the warning literature.

The models that we used assumed stable criteria across the recognition test. Neely and Tse (in press) discuss how violations of this assumption can cloud the results of an SDT analysis. For example, parameter values will be distorted if participants systematically shift their criterion based on the type of item tested (e.g., list word versus critical theme), as may result if different items types vary in certain characteristics (e.g., frequency of occurrence). Notably, our second experiment eliminated the possibility of systematic within-test criterion shifts: The words used as targets, critical theme lures, and unrelated lures were fully counterbalanced. Thus, the item types did not differ in terms of any characteristic other than their presentation history. The results of the second experiment were fully consistent with the conclusions that we drew from the existing studies and our first experiment, so systematic within-test criterion shifts do not appear to play a role in the results of warning studies.

Neely and Tse (in press) also discuss how unsystematic variability in criteria can impact the results of an SDT analysis. These researchers note that random fluctuations in criterion placement can decrease estimates of memory sensitivity (e.g., d'), such that a manipulation that impairs participants' ability to maintain stable criteria throughout a test may appear to negatively impact memory sensitivity. How might criterion fluctuation influence our results? Most significantly, our finding that the warning reduced memory evidence for critical themes may in fact only signify that the warning engendered variability in response criteria. Although our model results cannot eliminate this possibility, we consider it unlikely. We can think of no compelling reason why a warning would impair participants' ability to maintain stable criteria. Moreover, if inflated criterion variability biased the critical theme memory parameters for warned participants, then one would expect a similar bias to appear in the target memory parameters. That is, target memory should appear to be worse in the warned condition. Table 4

reveals that this pattern did not emerge in either of our experiments. For these reasons, we are confident that our critical theme results reflect true differences in memory evidence.

Nevertheless, we acknowledge that this claim could be more definitively established by a model that accommodates the influence of criterion variability.

The evidence-change hypothesis holds that critical themes are less familiar for warned participants because these participants alter their retrieval searches to capitalize more on item-specific and less on relational types of information in memory. This process would be considered an “early selection” form of cognitive control [Jacoby, Kelley, & McElree (1999); cf. retrieval orientation (Rugg & Wilding, 2000; Herron & Rugg, 2003) or retrieval focus (Schacter, Norman, & Koustaal, 1998)]. Our results suggest not only that memory performance is a function of the types of evidence sought from memory (Johnson et al., 1993), but also that participants can strategically control the content of retrieval cues to optimize discrimination among items with different origins. To select appropriate retrieval cues, participants must have knowledge regarding both the types of items that need to be distinguished and the forms of evidence that best discriminate these items. Thus, warned participants may be able to select a more effective retrieval strategy because the warning alerts them to the presence of highly related lures and to the fact that relational evidence could lead to memory errors. By this account, the warned participants’ advantage is their enhanced metamnemonic knowledge of the types of information that are most useful in discriminating studied from unstudied items. Future research can explore other variables and individual difference factors that moderate participants’ ability to discover the forms of evidence that most effectively meet the goals of a particular memory task, and how this ability impacts memory performance. One interesting possibility is that

participants can make on-line adjustments in the types of evidence that they use on a memory test based on performance feedback (Lane, Roussel, Villa & Morita, 2006).

Besides retrieval warnings, several other manipulations investigated in the false memory literature suggest that changing the types of evidence considered at retrieval can impact false memory (Hicks & Marsh, 1999; Mather et al, 1997; Multhaup & Conner, 2002). For example, Mather et al. (1997) showed that requiring participants to rate phenomenological details associated with recognized items reduced false recognition when associative lists were mixed at encoding. Similarly, some studies suggest that adding source judgments to a memory test can reduce false memory (e.g., Hicks & Marsh, 1999; Multhaup & Conner, 2002), although the opposite finding has also been reported (Hicks & Marsh, 2001). It is likely that both of these manipulations highlight item-specific information at retrieval, which may explain the false-memory reductions that they can induce. Moreover, studies in which participants rated the phenomenological characteristics of their memories have consistently shown that critical themes are rated much higher on certain characteristics than others; for example, critical theme ratings for associative detail are much higher than ratings for auditory detail (e.g., Mather, et al., 1997; Norman & Schacter, 1997). Such findings directly support the contention that the amount of evidence retrieved for critical themes depends on the types of information used to cue memory.

Our encoding procedures differed in some respects from the “standard” associative list paradigm. We used 6-item lists compared to the typical 15-item lists. We also had a recall period following each list, which is not unusual in the associative list literature in general (e.g., Roediger & McDermott, 1995) but is atypical of studies investigating retrieval warnings. Probably reflecting both of these methodological details, we observed false memory rates that were much lower than is found in the typical false memory study. In light of these deviations

from the “standard” false memory paradigm, one may wonder whether our findings are peculiar to our methodology. A set of closely related experiments that we conducted as part of another project (Lane, Roussel, Villa, Starns & Alonzo, 2006) suggests that our results can be obtained with a more typical false memory methodology. For example, Experiment 3 in Lane et al. involved 15-item lists with recall following half of the lists. Consistent with the current study, results from college-aged participants showed that instructions that alerted participants to the presence of highly related lures led to a selective decrease in critical theme false alarm rate. This pattern emerged even though the false memory rate following standard instructions was above .80, indicating a strong false memory effect. Furthermore, this pattern was consistent for recalled and non-recalled lists. As in the current study, the results of Lane et al. (2006) were inconsistent with the claim that warned participants decrease false memory simply by becoming more conservative. Specifically, results showed a crossover pattern in which the critical theme false alarm rate was higher than the hit rate in the unwarned condition, but lower than the hit rate in the warned condition. For example, for unrecalled lists in Experiment 3, young adult participants accepted .86 of the critical themes and .77 of the list targets in the unwarned condition, compared to a critical theme false alarm rate of .66 and a hit rate of .72 in the warned condition. This crossover pattern is difficult to reconcile with a mere shift in criterion, thus suggesting that the warning impacted the amount of evidence retrieved for critical theme lures. In light of the consistency in findings across studies using different encoding procedures, we are confident that our conclusions are not limited to a particular methodology.

The incidence of false memory depends on both the amount of illusory evidence retrieved from memory and the standards applied to evaluate that evidence (e.g., Johnson et al., 1993; Schacter et al., 1998). To develop successful theories of false memory, researchers must be able

to independently assess how various manipulations impact illusory evidence versus retrieval standards. Many studies in the false memory literature demonstrate the utility of signal detection theory in meeting this goal (Miller & Wolford, 1999; Westerberg & Marsolek, 2003, 2006; Wickens & Hirshman, 2000; Wixted & Stretch, 2000). Continuing this theme, our study shows how SDT models can be used to more definitively establish the effects of retrieval warnings, and the model results indicated that warnings influence the amount of illusory evidence retrieved for critical themes.

References

- Anastasi, J. S., Rhodes, M. G., & Burns, M. C. (2000). Distinguishing between memory illusions and actual memories using phenomenological measurements and explicit warnings. *American Journal of Psychology, 113*, 1-26.
- Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology, 58*, 17-22.
- Dunn, J. C. (2004). Remember-know: A matter of confidence. *Psychological Review, 111*, 524-542.
- Gallo, D. A., Roberts, M. J., & Seamon, J. G. (1997). Remembering words not presented in lists: Can we avoid creating false memories? *Psychonomic Bulletin & Review, 4*, 271-276.
- Gallo, D. A., Roediger, H. L., III, & McDermott, K. B. (2001). Associative false recognition occurs without strategic criterion shifts. *Psychonomic Bulletin and Review, 8*, 579-586.
- Glanzer, M., Adams, J. K., Iverson, G. J., & Kim, K. (1993). The regularities of recognition memory. *Psychological Review, 100*, 546-567.
- Herron, J. E., & Rugg, M. D. (2003). Strategic influences on recollection in the exclusion task: Electrophysiological evidence. *Psychonomic Bulletin & Review, 10*, 703-710.
- Hicks, J. L., & Marsh, R. L. (1999). Attempts to reduce the incidence of false recall with source monitoring. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 25*, 1195-1209.
- Hicks, J. L., & Marsh, R. L. (2001). False recognition occurs more frequently during source identification than during old-new recognition. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 27*, 375-383.

- Hunt, R. R., & Einstein, G. O. (1981). Relational and item-specific information in memory. *Journal of Verbal Learning & Verbal Behavior*, 20, 497-514.
- Hunt, R., & McDaniel, M. (1993). The enigma of organization and distinctiveness. *Journal of Memory & Language*, 32, 421-445.
- Jacoby, L. L., Kelley, C. M., & McElree, B. D. (1999). The role of cognitive control: Early selection vs. late correction. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp.383-400). NY: Guilford.
- 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.
- Johnson, M. K., Foley, M. A., Suengas, A. G., & Raye, C. L. (1988). Phenomenal characteristics of memories for perceived and imagined autobiographical events. *Journal of Experimental Psychology: General*, 117, 371-376.
- Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin*, 114, 3-28.
- Koriat, A., & Goldsmith, M. (1994). Memory in naturalistic and laboratory contexts: Distinguishing the accuracy-oriented and quantity-oriented approaches to memory assessment. *Journal of Experimental Psychology: General*, 123, 297-315.
- Koriat, A. & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review*, 103, 490-517.
- Lane, S. M., Roussel, C. C., Villa, D., & Morita, S. (2006). *Features and feedback: Reducing source monitoring errors in an eyewitness paradigm*. Manuscript in preparation.

- Lane, S. M., Roussel, C. C., Villa, D., Starns, J. J. & Alonzo, J. D. (2006). *Changing expectations at retrieval reduces false recognition*. Manuscript under review.
- Mather, M., Henkel, L. A., & Johnson, M. K. (1997). Evaluating characteristics of false memories: Remember/know judgments and memory characteristics questionnaire compared. *Memory & Cognition*, *25*, 826-837.
- McCabe, D. P., & Smith, A. D. (2002). The effect of warnings on false memories in young and older adults. *Memory & Cognition*, *30*, 1065-1077.
- McDermott, K. B., & Roediger, H. L., III (1998). Attempting to avoid illusory memories: Robust false recognition of associates persists under conditions of explicit warnings and immediate testing. *Journal of Memory and Language*, *39*, 508-520.
- Miller, M. B., & Wolford, G. L. (1999). Theoretical commentary: The role of criterion shift in false memory. *Psychological Review*, *106*, 398-405.
- Multhaup K. S., & Conner C. A. (2002). The effects of considering nonlist sources on the Deese–Roediger–McDermott memory illusion. *Journal of Memory and Language*, *47*, 214-228.
- Neely, J. H., & Tse, C-S. (in press). Semantic relatedness effects on true and false memories in episodic recognition: A methodological and empirical review. To appear in J. S. Nairne (Ed.), *The foundations of remembering: Essays in honor of Henry L. Roediger III*. New York: Psychology Press.
- Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms. <http://www.usf.edu/FreeAssociation/>.

- Neuschatz, J. S., Benoit, G. E., & Payne, D. G. (2003). Effective warnings in the Deese-Roediger-McDermott false-memory paradigm: The role of identifiability. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *29*, 35-41.
- Neuschatz, J. S., Payne, D. G., Lampinen, J. M., & Toggia, M. P. (2001). Assessing the effectiveness of warnings and the phenomenological characteristics of false memories. *Memory*, *9*, 53-71.
- Norman, K. A., & Schacter, D. L. (1997). False recognition in younger and older adults: Exploring the characteristics of illusory memories. *Memory & Cognition*, *25*, 838-848.
- Reyna, V. F., & Kiernan, B. (1994). Development of gist versus verbatim memory in sentence recognition: Effects of lexical familiarity, semantic content, encoding instructions, and retention interval. *Developmental Psychology*, *30*, 178-191.
- Roediger, H. L., III (1996). Memory illusions. *Journal of Memory and Language*, *35*, 76-100.
- Roediger, H. L., III, & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 803-814.
- Rugg, M. D., & Wilding, E. L. (2000). Retrieval processing and episodic memory. *Trends in Cognitive Neurosciences*, *4*, 108 –115.
- Schacter, D. L., Norman, K., & Koutstaal, W. (1998). The cognitive neuroscience of constructive memory. *Annual Review of Psychology*, *49*, 298-318.
- Stretch, V., & Wixted, J. T. (1998). Decision rules for recognition memory confidence judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 1397-1410.

- Westerberg, C. E., & Marsolek, C. J. (2003). Sensitivity reductions in false recognition: A measure of false memories with stronger theoretical implications. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *29*, 747-759.
- Westerberg, C. E., & Marsolek, C. J. (2006). Do instructional warnings reduce false recognition? *Applied Cognitive Psychology*, *20*, 97-114.
- Wickens, T. D. (2002). Elementary signal detection theory. New York, NY: Oxford.
- Wickens, T. D., & Hirshman, E. (2000). False memories and statistical design theory: Comment on Miller and Wolford (1999) and Roediger and McDermott (1999). *Psychological Review*, *107*, 377-383.
- Wixted, J. T., & Stretch, V. (2000). The case against a criterion-shift account of false memory. *Psychological Review*, *107*, 369-376.
- Wixted, J. T., & Stretch, V. (2004). In defense of the signal detection interpretation of remember/know judgments. *Psychonomic Bulletin & Review*, *11*, 616-641.
- Yonelinas, A. P. (1997). Recognition memory ROCs for item and associative information: The contribution of recollection and familiarity. *Memory & Cognition*, *25*, 747-763.

Author Note

Jeffrey J. Starns, Sean M. Lane, Jill D. Alonzo, & Cristine C. Roussel, Department of Psychology, Louisiana State University.

The authors thank Courtney Bourgeois, Dana Ellis, David Marshall, Stephanie Martin, Valerie MacNeill, Brent Nobles, Christina Peairs, and Paige Raschke for their invaluable help collecting the data. This research was supported by State of Louisiana Board of Regents grant LEQSF(2004-07)-RD-A-12 awarded to S.M. L.

Footnotes

¹We use the term familiarity only to refer to evidence on a single continuum (Wickens, 2002), not as an alternative to recollection as in dual-process theories (e.g., Yonelinas, 1997). Further, we assume that evidence of various types, some of which would be considered “recollective” (e.g., contextual detail), can be integrated into a single continuous variable for use in decision-making (see Wixted & Stretch, 2004).

²We made no attempt to include every existing study using retrieval warnings in the fitting procedure; our goal was simply to assemble a representative sample that could be used to explore the veracity of the criterion-shift and distribution-shift explanations.

³McCabe and Smith (2002) and Neuschatz et al. (2001) reported only the total number of participants in each of their experiments, so we assumed that an equal number of participants were assigned to each of the between-subject conditions.

⁴Recognition studies in the associative list paradigm do not typically use only one item from each list on the test, but we did this to make Experiment 1 more comparable to the second experiment in which only critical themes were tested (making it impossible to have more than one item from each list).

Table 1

Parameter values from the SDT model fit to existing studies on retrieval warnings.

Dataset	Parameter and Warning Condition							
	d'_T		d'_{CT}			λ		
	NW	W	NW	W	ΔG^2	NW	W	ΔG^2
1 - Neuschatz (2001)	1.26	1.26	1.15	.94	2.050	.74	.74	.000
2 - Neuschatz (2001)	1.26	1.51	1.15	1.16	.001	.74	.95	7.109*
3 - Neuschatz (2001)	.85	.75	1.05	.72	5.109*	.47	.44	.144
4 - Neuschatz (2001)	.85	.72	1.05	.80	2.939	.47	.39	1.273
5 - Gallo (2001)	1.33	1.47	1.98	1.72	1.623	.81	.92	1.239
6 - Gallo (2001)	1.33	1.02	1.98	.91	37.412*	.81	.50	11.281*
7 - Anastasi (2000)	.98	1.25	.98	1.10	1.261	.95	1.23	14.449*
8 - Gallo (1997)	1.74	1.56	1.91	1.82	.212	1.04	1.18	1.482
9 - McCabe (2002)	2.30	2.08	1.76	1.15	7.230*	.95	1.00	.105
10 - McCabe (2002)	1.83	1.58	1.69	1.33	2.568	.95	1.00	.105
11 - McCabe (2002)	1.96	1.97	1.96	1.41	7.990*	.88	1.13	5.815*
12 - McCabe (2002)	2.08	1.58	2.28	1.30	23.709*	1.00	.77	5.070*

Note: NW = no warning; W = warning; ΔG^2 statistics report the change in fit from the unconstrained model to a model in which either the d'_{CT} or λ parameters were constrained to be equal in the warning and no-warning conditions. Values significant at the .05 level are marked with an asterisk. Studies are listed by first author and year (see the text for a detailed description of the datasets).

Table 2

Recognition Performance for Each Instruction Condition in Experiments 1 and 2.

Recognition Measure	Experiment and Instruction Condition			
	Experiment 1		Experiment 2	
	No Warning	Warning	No Warning	Warning
Hit Rate	.81 (.02)	.78 (.02)	.77 (.02)	.81 (.02)
Critical Theme FAR	.31 (.03)	.15 (.03)	.28 (.02)	.20 (.02)
Unrelated FAR	.04 (.01)	.03 (.01)	.02 (.01)	.03 (.01)

Note: FAR denotes false-alarm rate; values in parentheses are standard errors.

Table 3

Predicted and Observed Recognition Performance Measures at Each Criterion Level Across the Instruction Conditions of Experiments 1 and 2.

Criterion Level and Recognition Measure	Experiment and Instruction Condition							
	Experiment 1				Experiment 2			
	No Warning		Warning		No Warning		Warning	
	P	O	P	O	P	O	P	O
Criterion 1								
HR	.882	.880	.849	.848	.875	.872	.884	.878
CT FAR	.497	.495	.279	.278	.474	.474	.376	.378
UL FAR	.201	.203	.131	.131	.180	.181	.183	.184
Criterion 2								
HR	.805	.811	.773	.778	.762	.771	.795	.815
CT FAR	.297	.306	.145	.147	.278	.278	.209	.199
UL FAR	.048	.038	.035	.030	.028	.024	.040	.034
Criterion 3								
HR	.730	.730	.724	.723	.683	.683	.719	.717
CT FAR	.175	.175	.095	.094	.192	.192	.128	.132
UL FAR	.011	.017	.014	.018	.007	.011	.010	.014

Note: P = predicted value; O = observed value; HR denotes hit rate; CT FAR denotes critical theme false-alarm rate; UL FAR denotes unrelated lure false-alarm rate. Criterion 1 values are the proportion of items receiving “Guess New,” “Guess Old,” and “Sure Old” responses. Criterion 2 values are the proportion of items receiving “Guess Old” and “Sure Old” responses. Criterion 3 values are the proportion of items receiving “Sure Old” responses.

Table 4

Parameters of the SDT Model across the Instruction Conditions in Experiments 1 and 2.

Model Parameter	Experiment and Instruction Condition			
	Experiment 1		Experiment 2	
	No Warning	Warning	No Warning	Warning
Targets				
μ_T	3.85	3.63	3.52	3.65
σ_T	2.54	2.43	2.27	2.30
A_z	.92	.92	.92	.93
Critical Themes				
μ_{CT}	.82	.27	.79	.36
σ_{CT}	1.57	1.46	1.89	1.72
A_z	.67	.56	.64	.57
λ_1	.84	1.12	.92	.91
λ_2	1.66	1.81	1.90	1.76
λ_3	2.29	2.19	2.44	2.32

Figure Captions

Figure 1. Demonstration of a reduction in false memory for warned participants due to a criterion shift. In each panel, the leftmost distribution is the unrelated lure distribution, the middle distribution is the critical theme distribution, and the rightmost distribution is the target distribution. The vertical line represents the response criterion.

Figure 2. Demonstration of a reduction in false memory for warned participants due to a shift in the critical theme distribution. In each panel, the leftmost distribution is the unrelated lure distribution, the middle distribution is the critical theme distribution, and the rightmost distribution is the target distribution. The vertical line represents the response criterion.

Figure 3. Displays the effects of a conservative criterion shift on different item types in a situation in which the critical theme distribution lies halfway between the target and unrelated lure distributions (top panel) and a situation in which the critical theme distribution is close to the target distribution (bottom panel). In each panel, the leftmost distribution is the unrelated lure distribution, the middle distribution is the critical theme distribution, and the rightmost distribution is the target distribution. The vertical lines represent response criteria, and the arrows show the direction of a conservative criterion shift.

Figure 4. Displays the effects of a conservative criterion shift on distributions with different standard deviations. The leftmost distribution is the unrelated lure distribution, the middle distribution is the critical theme distribution, and the rightmost distribution is the target distribution. The vertical lines represent response criteria, and the arrow shows the direction of a conservative criterion shift.







