Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Memory and Language 51 (2004) 473–493
Journal of Memory and Language
www.elsevier.com/locate/jml

Reducing false recognition with criterial recollection tests: Distinctiveness heuristic versus criterion shifts☆

David A. Gallo,* Jonathan A. Weiss, and Daniel L. Schacter

Psychology Department, Harvard University, 33 Kirkland St., Cambridge, MA 02138, USA

Received 15 April 2004; revision received 1 June 2004
Available online 8 July 2004

Abstract

We devised criterial recollection tests to investigate why testing memory for pictures elicits lower false recognition than testing memory for words. Subjects studied unrelated black words paired either with the same word in red font, a corresponding picture, or both. They then took three memory tests, always using black words: a recognition test (say "yes" to all studied items), a red word-test, and a picture-test (say "yes" only if you recollect a red word or a picture, respectively). Regardless of whether pictures were more or less familiar than red words, false recognition was lowest on the picture test. These results cannot be explained easily by familiarity or strength-based criterion shifts. Instead, they suggest that subjects expected more detailed recollections for pictures, thereby facilitating a diagnostic monitoring process (the "distinctiveness heuristic"). This recollective difference also influenced source monitoring errors (an "it-had-to-be-a-word" effect), again suggesting that detailed recollective expectations influence monitoring processes.
© 2004 Elsevier Inc. All rights reserved.

Introduction

Much recent theorizing in memory research has focused on how episodic memory accuracy can be improved through metacognitive monitoring processes (e.g., Brainerd, Reyna, Wright, & Mojardin, 2003; Johnson, Hashtroudi, & Lindsay, 1993; Koriat, Goldsmith, & Pansky, 2000; Roediger, Watson, McDermott, & Gallo, 2001; Schacter, Norman, & Koutstaal, 1998). Schacter and colleagues have argued that such processes can help to understand why more distinctive events lead to lower false recognition than less distinctive events in false memory tasks (for a review see Schacter & Wiseman, in press). Consider a popular example from the Deese (1959)/Roediger and McDermott (1995) task (DRM). In this task subjects falsely remember a nonstudied word (e.g., *chair*) because it is associated with a list of studied words (e.g., *table*, *desk*, *couch*, ...). Israel and Schacter (1997) found that pairing each studied word with a pictorial representation significantly reduced this illusion. They argued that studying pictures led to more distinctive recollections than did studying words, and as a result, subjects in the picture condition expected to retrieve more distinctive recollections about studied events. Here we use "distinctiveness" to refer to the complexity and uniqueness of the perceptual features of a stimulus (cf. Nelson, 1979). Because nonstudied events (e.g., *chair*) would not be accompanied with distinctive picture recollections, the failure to recollect the expected features would suggest that they were not studied, thereby facilitating rejection.

Schacter, Israel, and Racine (1999) called this monitoring process "the distinctiveness heuristic" to highlight the role of recollective expectations in guiding memory decisions (see also Ghetti, 2003; Strack & Bless, 1994). Exactly how one conceptualizes such a process depends on the underlying theory of false recognition. One interpretation of the distinctiveness heuristic comes from

☆ This research was supported by National Institute on Aging Grants AG021369 and AG08441.
* Corresponding author.
E-mail address: dgallo@wjh.harvard.edu (D.A. Gallo).

0749-596X/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.jml.2004.06.002

dual-process theory, which assumes that recognition memory is influenced by recollection (i.e., the recall of details of the prior occurrence of an event) and familiarity (i.e., the feeling that an event had previously occurred, without the recall of detailed information). According to the distinctiveness heuristic, when subjects expect more distinctive recollections (e.g., pictures) they are less likely to rely on familiarity than when they expect less distinctive recollections (e.g., words). As a result, they are less prone to familiarity-based false recognition. Another interpretation focuses on the idea that recognition can be based on the recollection of qualitatively different types of information (or "multiple dimensions"), each of which can differentially contribute to performance depending on task-relevant decision processes (e.g., each dimension is weighted via the setting of "decision axes," see Banks, 2000; Johnson & Raye, 1981). By these views, false recognition can be caused by the attribution of an event to the wrong source, due to insufficient, degraded, or "illusory" recollection of the features of the event (see Gallo & Roediger, 2003; Schacter et al., 1998, for relevant discussions). According to the distinctiveness heuristic, such false attributions are less likely when illusory recollection fails to correspond to the subject's recollective expectations. When the subject expects more distinctive recollections (e.g., pictures), illusory recollection is less likely to conform to expectations than when they expect less distinctive recollections (e.g., words), and false recognition is reduced. Note that these interpretations are not mutually exclusive—false recognition could be driven by a vague feeling of familiarity or by a strong sense of illusory recollection.

Regardless of the hypothetical cause of false recognition (familiarity and/or illusory recollection), the distinctiveness heuristic explains the picture/word effect on false recognition by appealing to the notion of recollective expectations. With the present investigation we questioned this basic idea. Do we need to appeal to a decision process based on expected recollections, or could a purely familiarity-based or strength-based account provide a simpler explanation of these effects? According to classic unidimensional signal-detection theories, recognition memory is guided by familiarity and a response criterion. By this view, the picture/word effect on false recognition could be caused by a more conservative familiarity-based response criterion after studying pictures, relative to words. Such unidimensional models of picture/word effects in recognition memory are not uncommon (e.g., Hintzman, Curran, & Caulton, 1995; Morrell, Gaitan, & Wixted, 2002; see Glanzer & Adams, 1985, for relevant discussion of the "mirror effect"), and criterion shifts along a single dimension have been used to explain between-subjects or between-list reductions in false recognition that are based on stimulus "strength" (e.g., Hirshman, 1995; Stretch & Wixted, 1998). It therefore seems reasonable that familiarity-based criterion shifts could provide an alternative explanation to the distinctiveness heuristic. Of course, both criterion shifts and the distinctiveness heuristic can be considered monitoring or decision processes, but the critical difference is that criterion shifts are based on expected *levels of familiarity*, whereas the distinctiveness heuristic is based on expected *types of recollections*.[1]

To understand how familiarity-based criterion shifts could explain the picture/word effect on false recognition, consider a unidimensional model of DRM false memories similar to the model proposed by Wixted and Stretch (2000, see also Wickens & Hirshman, 2000). For the sake of argument, assume that recognition performance is driven purely by familiarity and also that related lures are more familiar than unrelated lures but not quite as familiar as list items. The familiarity of related lures could be increased through processes such as associative-activation, semantic feature overlap, or relatedness to the thematic gist of the list (for simplicity, we assume these processes would not be greatly affected by picture or word presentation). Fig. 1 presents such a model. For ease of illustration, items are normally distributed with equal variance, and the response criterion (vertical line) is set near the intersection of the curves (indicating no bias). In order to recognize most of the studied words, this criterion must be set somewhere to the left of the mean of the target distribution. Because related lures are almost as familiar as targets, this criterion will also cause many false alarms to related lures, as is typically observed.[2]

[1] Some unidimensional models use the terms "strength" or "strength of evidence" instead of "familiarity," in order to remain agnostic with regard to the subjective experience of retrieval (i.e., recollection or familiarity). However, by definition, unidimensional models do not allow different types of retrieval to differentially contribute to performance (as in dual process models or multidimensional models). As a result, unidimensional models based on strength or familiarity make identical predictions in the present task, and unless otherwise noted, we treat the two synonymously. We use the term "familiarity" instead of "strength" to leave open the possibility that recollection can contribute to performance, even if familiarity is modeled with classic signal-detection procedures (see Yonelinas, 2002).

[2] This unidimensional model is different than that proposed by Miller and Wolford (1999), who argued that DRM false recognition is due mostly to criterion shifts to related lures, as opposed to the familiarity of these lures. Objections to Miller and Wolford's item-specific criterion shift model have been made on theoretical grounds (Roediger & McDermott, 1999; Wickens & Hirshman, 2000; Wixted & Stretch, 2000) and empirical grounds (e.g., Gallo, Roediger, & McDermott, 2001; Roediger et al., 2001), but in general, familiarity-based criterion shifts remain a critical concern in false memory research.

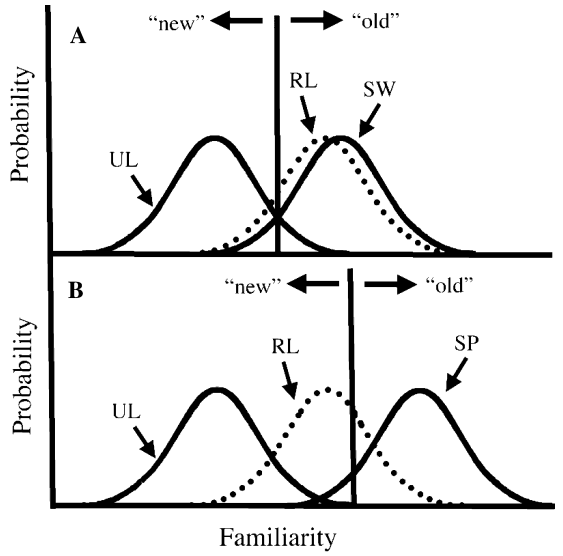

Fig. 1. An idealized familiarity-based model of the DRM task. (A) On average, studied words (SW) and related lures (RL) are more familiar than unrelated lures (UL), yielding greater hits and related FAs than unrelated FAs. (B) When subjects study pictures, the target distribution (SP) is more familiar than when they study words (SW). As a result, subjects can use a more conservative response criterion relative to the word condition, thereby lowering false alarms to related and unrelated lures with little or no cost to hits.

The familiarity-based model in Fig. 1 can readily account for reductions in false recognition after studying pictures, relative to words, via a conservative criterion shift. Consider the younger adult results of Schacter et al. (1999, Experiment 1), in the condition where study format was manipulated between-subjects and items were presented verbally at test. In this study, false alarms to related and unrelated lures were lower in the picture condition (mean = .41 and .08, respectively) than the word condition (.64 and .26), but hits to pictures (.71) did not differ from hits to words (.77). If one assumes that picture targets were more familiar than word targets (an assumption owing to the picture-superiority effect in memory, e.g., Paivio, 1971), then subjects in the picture condition could have set a more conservative response criterion along the familiarity dimension relative to subjects in the word condition (see Fig. 1B). As a result, subjects in the picture condition would have reduced false alarms to both related and unrelated lures relative to the word condition, at little or no cost to their hit rate, thereby explaining the full pattern of results (see Israel & Schacter, 1997; and Schacter, Cendan, Dodson, & Clifford, 2001, for similar results). The only assumption made by this model is that subjects set their familiarity-based response criterion depending on the relative positions (or levels of familiarity) of the memory distributions.

This simple familiarity-based model also can explain the results from Schacter et al. (1999, Experiment 2), when study format of the lists was manipulated within-subjects. Subjects studied several lists, as words or pictures, and took a final recognition test for all lists. Under these conditions subjects would use the same response criterion for all test words, regardless of the study format of the corresponding list, because recollections of study format could not influence a purely familiarity-based account. As a result, false alarms from picture lists would equal those from word lists, assuming the lures were equally familiar. Consistent with this prediction, there was no difference in false alarms to related lures from picture lists (.55) and word lists (.54) when format was manipulated within-subjects. Thus, when the same familiarity-based response criterion was used for related lures from picture or word lists, there was no picture/word effect on false recognition. There was an overall reduction in false alarms in this experiment (collapsing across study formats), relative to the between-subjects experiment, but as discussed next this also is consistent with the familiarity-based criterion shift account.[3]

Three other pieces of evidence that have been interpreted in terms of a recollection-based distinctiveness heuristic also can be explained by a familiarity-based criterion shift. First, according to the distinctiveness heuristic, subjects should not demand distinctive recollections when only some of the studied materials are studied as pictures (i.e., a within-subjects manipulation). Nevertheless, overall levels of false recognition were suppressed even when the two formats were mixed at study, with the exact level of suppression varying across studies and tasks (Dodson & Schacter, 2001, 2002a, 2002b; Schacter et al., 1999, 2001). One explanation is that these subjects inappropriately used the distinctiveness heuristic even when picture recollections were not perfectly diagnostic of study presentation. An equally viable explanation, though, is that the average familiarity of the study stimuli in these mixed picture/word

[3] Interestingly, a picture-superiority effect was not obtained even under these within-subject conditions (picture hits = .77, word hits = .79). This result is inconsistent with both the distinctiveness heuristic (which predicts more distinctive recollections for picture targets than word targets) and the criterion shift account (which predicts more familiarity for picture targets). However, as Dodson and Schacter (2002b) point out, picture-superiority effects in hit rates are not always demonstrated on verbal recognition tests (see Mintzer & Snodgrass, 1999). One potential reason, at least for the conditions discussed here, is that test items were presented only as words. Thus, the study/test match for items studied as words may have counteracted the benefits of studying pictures on true memory. We avoided this complication in the present series, and a picture superiority effect was obtained on a verbal recognition test (see Experiment 1).

conditions was greater than that in the word-only conditions. As a result, subjects in the mixed conditions suppressed false recognition by using a more conservative response criterion, with the exact size of the suppression effect depending on how accurate subjects were in setting their response criterion (which, like the distinctiveness heuristic, would not necessarily be perfect). Second, Dodson and Schacter (2002b) found that instructing subjects that they would be tested only for items studied as words, even though they had studied pictures and words, reduced the false recognition suppression effect. They argued that the instructions caused subjects to avoid a distinctiveness heuristic, but it also could be argued that the instructions caused them to avoid a conservative criterion shift. Third, estimates of bias (B″d or C) have been more conservative in picture study conditions than word conditions in several studies (Dodson & Schacter, 2001, 2002a, 2002b; Schacter et al., 1999). Changes in measures of bias are not sufficient evidence for criterion shifts (see Wixted & Stretch, 2000), but they are consistent with the criterion shift model as depicted in Fig. 1B, where the response criterion is located to the right of the intersection of the two distributions.

In sum, all of the aforementioned evidence that has been provided in support of the distinctiveness heuristic also can be interpreted in terms of a simple familiarity-based criterion shift. By raising this alternative we are not proposing that recall or recollection does not contribute to recognition memory performance—there is ample evidence that it does (e.g., Yonelinas, 2002). Instead, we are arguing that there is no evidence to favor the distinctiveness heuristic account over a familiarity-based criterion shift account of the picture/word effect on false recognition. Even in dual process theories that allow for recollection and familiarity (e.g., Rotello, Macmillan, & Van Tassel, 2000; Yonelinas, 1997), the picture/word effect on false recognition could be caused by a familiarity-based criterion shift alone. The criterion shift and distinctiveness heuristic theories offer equally viable explanations of previous results because, in all prior experiments, studying pictures relative to words could have led to increases in both recollective distinctiveness and familiarity. In the present study we sought to decouple presentation format (picture vs. word) with levels of familiarity, and thereby test the two theories.

For the present study we devised a task that could distinguish between the distinctiveness heuristic and criterion shift accounts of the picture/word effect on false recognition. Rather than inferring that subjects are using different retrieval expectations following the study of pictures and words, as in previous studies of the distinctiveness heuristic, in Experiments 1 and 2 we directly manipulated the recollective demands of the recognition test (which we call a criterial recollection test). In brief, subjects studied pictures and red words. On the picture test, subjects said "yes" to items studied as pictures and "no" to items studied as red words and to nonstudied items, and vice versa on the red word test. Importantly, we presented some items as both pictures and red words during the study phase, so that subjects could not use a recall-based exclusion strategy to reject a test item (i.e., "I recall that this was a red word, so it couldn't have been a picture," see Jacoby, 1999). In lieu of such a recall-to-reject strategy, they had to search memory for evidence that the stimulus was presented in the relevant format (e.g., pictures on the picture test). Experiment 3 provided a manipulation check for Experiment 2. In Experiments 4 and 5 we used more typical source tests to investigate further the differences between pictures and words in memory decisions, and to provide an additional test between the distinctiveness heuristic and criterion shift accounts.

Experiment 1

The goal of Experiment 1 was to demonstrate the false recognition pattern that has been attributed to the distinctiveness heuristic using criterial recollection tests. Subjects studied a list of unrelated black words. Each black word (e.g., *dragon*) was followed either by a picture (a picture of a dragon), or a red word (the word "dragon" printed in a larger red font). Some of the black words were presented once (followed by either a picture or a red word), and some were presented twice (once followed by a picture, and once followed by a red word). Subjects were then given a standard recognition memory test and two criterial recollection tests (the red word test and the picture test), with all test words presented in the same black font as used at study. On the standard test, they were instructed to say "yes" to any item that was studied (regardless of whether it was a red word or a picture). On this test, we expected to find the typical within-subjects picture superiority effect—greater hits to black words studied with pictures (hereafter we simply call these "pictures") than to black words studied with red words (hereafter "red words"). Of greater interest was the pattern of false alarms on the criterial recollection tests. Exclusion errors are defined as erroneously accepting pictures on the red word test and red words on the picture test. These errors would be driven by study-induced familiarity (and/or illusory recollection) and by a failure of monitoring processes that operate on this familiarity (e.g., familiarity-based criterion setting and/or a distinctiveness heuristic). The other type of false alarms are those to nonstudied words, which would be driven by the idiosyncratic familiarity of these words and a failure of monitoring processes operating on this familiarity (again, criterion shifts or the distinctiveness heuristic).

The distinctiveness heuristic predicts that it will be easier to reduce both types of errors on the picture test, because pictures afford more distinctive retrieval expectations than red words. The criterion shift account also predicts fewer errors on the picture test, because pictures should be more familiar than red words, and thus should engender a more conservative response criterion. That is, a similar criterion shift as described in Fig. 1 could occur in this situation, thereby lowering both types of false alarms. The exact implementation of the familiarity-based model in this task might involve multiple criteria, because there are multiple item types to distinguish, but the logic and the predictions are the same (we discuss the multiple-criteria model more thoroughly in the results section). Note that, under any explanation, the familiarity difference of the stimuli alone (pictures > red words) would predict that exclusion errors will be greater on the red word test (false alarms to pictures) than on the picture test (false alarms to red words), in the absence of monitoring processes. More direct evidence for one of the two types of monitoring processes will be in whether false alarms to nonstudied items will be lower on the picture test than the red word test. Both the distinctiveness heuristic and the criterion shift account predict that they should be.

Method

Subjects

Twenty-four Harvard University undergraduates participated for $10. Data from one subject were replaced because they did not finish the experimental session.

Materials and design

Study materials were 288 common words and corresponding colored pictures obtained from the internet (we thank Rachel Garoff for supplying the materials). Average word length was 6.1 letters ($SD = 1.7$), and the average printed word frequency (Kucera & Francis, 1967) was 21.49 per million ($SD = 46.52$). Frequency information was not available for 14% of the words. Each picture represented a single isolated object on a white background.

Stimuli were divided into 12 sets of 24 items. Across 12 counterbalancing conditions, each set occurred once in each of the 12 study/test combinations, which were obtained by crossing the four study conditions (pictures, red words, both, or nonstudied) with the three test conditions (standard test, red word test, or picture test). The standard test always came first, to provide an estimate of recognition memory for the different classes of stimuli. The order of the two criterial recollection tests was counterbalanced across subjects, resulting in a total of 24 counterbalancing conditions.

Study and test materials were presented via computer. Subjects studied 216 unique items, with 1/3 presented as red words, 1/3 presented as pictures, and 1/3 presented as both red words and pictures (for a total of 288 events). Each studied item first was presented in black lowercase letters using Courier font for 700 ms. It was then replaced with either a picture referent or with the same word in the red font, each for 2000 ms. Pictures ranged in size from $1'' \times 1''$ to $3'' \times 3''$. Red words were presented in red-colored Sand font that was visibly larger and distinct from the Courier font. A 700 ms blank screen separated each picture or red word from the next item. Items were randomly presented during study, with the exception that 1/3 of the items from each study/test combination were presented in the beginning, middle, and end of the phase. This was done to ensure an even sampling of the different types of items across the three sections of the study phase, which were separated by two rest prompts. For items that were presented as both a picture and a red word, the two occurrences were randomly spaced in the corresponding third of the study list.

Test items were presented using the same black font that had been used for each item at study, so that the perceptual overlap between study and test could not serve as a cue for whether the item had been studied with a red word or with a picture (or both). Each test contained four item types: items studied with red words, items studied with pictures, items studied with both red words and pictures, and nonstudied items. On the test with standard recognition instructions (the standard test), 3/4 of the items were targets and 1/4 were lures, whereas on the criterial recollection tests, 1/2 the items were targets and 1/2 the items were lures. For each of the three tests, items were freshly randomized for each subject.

Procedure

Subjects were told that they would study a list of items presented on the computer screen. They were told that some items would be presented as red words, others as pictures, and others as both red words and pictures, and that they should pay close attention to both the words and pictures, because their memory would later be tested (the exact nature of the tests was not revealed at this time). The total study phase took approximately 15 min, with two break prompts (''Rest briefly. Press space to resume study phase.'') separating the beginning, middle, and end of the study list.

At the end of the study phase, the experimenter read the instructions for the standard recognition memory test. Subjects were told that they would be presented with test words, one at a time, on the computer screen, and that some of these words were studied (with red words or pictures) and some were not studied (new). If they remembered studying either a red word or a

picture, they pressed the key labeled "yes," and if they thought the word was never studied, as either a picture or a red word, they pressed "no." At the end of the standard test, the experimenter read the instructions for the first criterial recollection test. For the red word test, subjects were told that their memory would be tested for the red words. They were instructed to respond "yes" only if they remembered studying the test word in red letters. They were further reminded that some red words also were studied as pictures, and other red words were never studied as pictures. Thus, whether or not they remembered studying a picture was irrelevant to the red word test. Instructions for the picture test were identical, except subjects were instructed to say "yes" only to words that they had studied as pictures, and that their memory for red words was now irrelevant. All test decisions were self-paced, and the experimenter ensured that the subjects understood each of the sets of instructions. Following the final test phase, subjects were given a brief questionnaire regarding their strategies used on each of the tests.

Results and discussion

Recognition performance is presented in the left column of Table 1. Consider first the results for those stimuli tested on the standard recognition test. The main point to notice is that the typical picture-superiority effect in within-subjects designs was obtained (picture hits = .66, red word hits = .45, $t[23] = 7.62$, $SEM = .028$, $p < .001$), and each of these hit rates was significantly greater than new-FAs, or false alarms to nonstudied lures (.10; all p's $< .001$). As expected, both-hits (the only items that were presented twice) were greater than red word hits (.70 vs. .45, $t[23] = 7.42$, $SEM = .034$, $p < .001$), and were marginally greater than picture hits (.70 vs. .66, $t[23] = 1.95$, $SEM = .022$, $p = .06$).

Table 1
Mean recognition of each item type as a function of test type in Experiments 1 and 2

	Experiment 1 (Red words 1×)	Experiment 2 (Red words 3×)
Standard test		
Both hits	.70 (.04)	.82 (.02)
Red word hits	.45 (.04)	.72 (.03)
Picture hits	.66 (.04)	.61 (.04)
New FAs	.10 (.03)	.10 (.02)
Red word test		
Both hits	.46 (.04)	.70 (.03)
Red word hits	.40 (.04)	.61 (.03)
Picture FAs	.31 (.04)	.35 (.03)
New FAs	.11 (.02)	.11 (.02)
Picture test		
Both hits	.56 (.03)	.54 (.04)
Red word FAs	.14 (.02)	.10 (.02)
Picture hits	.51 (.04)	.46 (.04)
New FAs	.02 (.01)	.01 (.00)

Note. Standard errors of each mean are in parentheses. FAs, false alarms.

Consider next the results from those items that were tested on the criterial recollection tests (the red word and picture tests). In general, hit rates tended to be lower on these tests than on the standard test, which would be expected if subjects were relying less on familiarity and more on criterial recollection. On the red word test, red word hits (.40) were greater than picture FAs (.31), $t(23) = 2.58$, $SEM = .036$, $p < .05$. A single-criterion familiarity-based model cannot explain this result. By that model, responses to pictures always should have been greater than those to red words, because pictures were "stronger" than words on the standard test. To account for this reversal, one would need to assume that subjects were relying on criterial recollection (more than familiarity) on the criterial recollection tests, or that subjects were using multiple familiarity-based response criteria (discussed more below). On the picture test, picture hits (.51) were greater than red word FAs (.14), $t(23) = 9.97$, $SEM = .037$, $p < .001$. This result also is consistent with the idea that subjects were using criterial recollection, although familiarity differences (pictures > red words) could explain this difference.

Evidence that subjects were influenced by familiarity (or some form of illusory recollection) on the criterial recollection tests comes from the pattern of false positive errors. False alarms for to-be-excluded studied items were more likely than false alarms for nonstudied items on each criterial test (.31 vs. .11 on the red word test, $t[23] = 5.04$, $SEM = .039$, $p < .001$, and .14 vs. .02 on the picture test, $t[23] = 6.29$, $SEM = .018$, $p < .001$). These effects suggest that subjects were influenced by the prior presentation of these to-be-excluded items. Finally, as on the standard test, both-hits tended to be greater than those to red words or pictures. This effect was significant on the red word test (.46 vs. .40, $t[23] = 2.10$, $SEM = .03$, $p < .05$), but failed to reach significance on the picture test (.56 vs. .51, $t[23] = 1.75$, $SEM = .031$, $p = .09$).

Directly comparing the criterial tests

The most important aspect of this task is that it allows a direct comparison of performance when decisions were based only on memory for one of the study formats (either red words or pictures). In this regard, it can be seen that picture hits on the picture test (.51) were greater than red word hits on the red word test (.40), $t(23) = 2.89$, $SEM = .037$, $p < .01$, indicating that the picture superiority effect that was observed on the standard test was replicated across the criterial

recollection tests. Consider next the hit rates to "both" items. Because these items had the same presentation history, they should have been equally familiar on the two tests. However, both-hits were greater on the picture test (.56) than on the red word test (.46), $t(23) = 3.28$, $SEM = .03$, $p < .01$. To account for this result, a single-criterion familiarity-based model would need to assume a conservative response criterion on the red word test, but this model predicts a criterion shift in the opposite direction (because pictures were stronger than red words, see below). An alternative explanation is that some combination of recollection and familiarity contributed to hits on these criterial recollection tests (and picture recollection > red word recollection), and/or that subjects had used multiple-criteria along the familiarity dimension (instead of a single-criterion).

Most important, false alarms were lower on the picture test than on the red word test. Red word FAs on the picture test (.14) were lower than picture FAs on the red word test (.31), $t(23) = 4.64$, $SEM = .037$, $p < .001$, and a similar pattern was observed for false alarms to nonstudied lures (.02 vs. .11), $t(23) = 5.07$, $SEM = .016$, $p < .001$. As discussed in the introduction to this experiment, these effects are consistent with a recollection-based distinctiveness heuristic. Under this hypothesis, subjects should expect more distinctive recollections on the picture test than the red word test, thereby lowering false alarms for to-be-excluded studied items and for nonstudied lures. However, a familiarity-based model also could explain these effects. Even if one assumes that recollection contributed to hits on the criterial recollection tests, false alarms could still have been driven by familiarity (i.e., a dual-process account such as those reviewed by Yonelinas, 2002). Because pictures were "stronger" in memory than words, this account could predict a conservative familiarity-based criterion on the picture test, thereby lowering false alarms relative to the red word test. This criterion shift also would lower familiarity-based hits on the picture test, relative to the red word test, but recollection could compensate for this effect, leading to the observed pattern of greater hits on the picture than the red word test.

A multiple-criteria model

As discussed above, a single-criterion familiarity-based model could not account for all of the data from Experiment 1. One solution is to assume that subjects relied on recollection (in addition to familiarity) on the criterial recollection tests. Another solution would be to allow multiple response criteria along the familiarity dimension (e.g., Donaldson, 1996; see Hirshman, Lanning, Master, & Henzler, 2002, for recent discussion). Such a model for the criterial recollection tests of Experiment 1 is presented in Fig. 2 (top). In this figure, each of the four item types is represented by a different distribution, and their relative placement (or strength) is

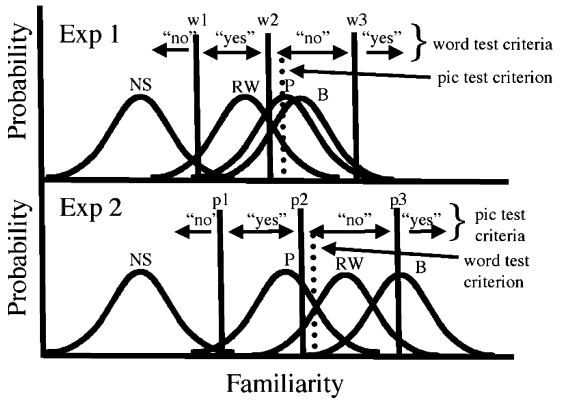

Fig. 2. An idealized familiarity-based model of the criterial recollection tests. In Experiment 1 (A), one criterion (dashed) is used for the picture test, but three criteria (w1, w2, w3) are used for the red word test. In Experiment 2 (B), red words are more familiar than pictures, so one criterion is needed for the red word test (dashed), and three (p1, p2, and p3) are needed for the picture test.

determined by the differences in hit rates on the standard test of Experiment 1 (nonstudied [NS] < red words [RW] < pictures [P] < both [B]). This strength difference also was reflected in the hit rates on the criterial recollection tests. For simplicity, we have drawn normal distributions with equal variance and minimal overlap. These assumptions may be violated in practice, but that would not change the logic or key conclusions of the present analysis.

The important feature of this model is in the setting of the response criteria, which were roughly drawn to illustrate how the obtained pattern of recognition responses could be obtained (again, for ease of illustration, these figures were not drawn perfectly). On the picture test, only one criterion (the dashed line) is needed to discriminate between the two target distributions (P & B) and the two lure distributions (NS & RW). This criterion is drawn to bisect the picture distribution, reflecting the obtained hit rate for pictures (.51), a slightly greater hit rate for both-items (.56), a small FA rate for red words (.14), and a negligible false alarm rate for new items (.02). Multiple criteria (w1, w2, and w3) are needed on the red word test. W1 is used to discriminate between nonstudied lures and red word targets, yielding the low nonstudied FA rate (.11). A second criterion (w2) is also needed, because subjects need to exclude items studied only as pictures. By placing this criterion between RW and P, the subjects could respond positively to many of the red word targets (by responding "yes" to items between w1 and w2), and negatively to many of the picture lures (by responding "no" to items above w2). In the figure, w2 is drawn so that many of the red words fall between w1 and w2 (yielding the obtained hit rate of .40), and fewer pictures fall in this range

(yielding the obtained FA rate of .31). Finally, a third criterion (w3) can be added, to explain why both-hits (.46) were not lower than picture-hits (.40). In short, much like the dual-process model with a single familiarity-based criterion shift, this multiple-criteria model could explain the data obtained in Experiment 1 without postulating the use of a recollection-based distinctiveness heuristic.

Experiment 2

In Experiment 1 we found that testing subjects on pictures led to fewer errors than testing subjects on red words, considering both exclusion errors and false alarms to nonstudied words. This pattern is predicted by the distinctiveness heuristic, but it also could be explained by familiarity-based criterion shifts (under either the dual process model or the multiple-criteria model). Experiment 2 provides a critical test between these competing explanations. This experiment differed from Experiment 1 in only one feature: All red words were presented three times at study, so that red words would be more familiar than pictures. According to the single response-criterion model, this manipulation should reverse the pattern of false alarms across the two criterial recollection tests. When red words are more familiar, subjects should use a more conservative familiarity-based criterion on the red word test than on the picture test, leading to lower false alarms on the red word test. For the same reasons, this model also predicts that false alarms on the red word test of Experiment 2 would be lower than the corresponding false alarms in Experiment 1—if red word familiarity is greater in Experiment 2, then subjects should use a more conservative response criterion. Similar criterion shifts would be involved in a multiple-criteria model, potentially resulting in similar predictions, although we withhold making exact predictions in order to allow some flexibility in the setting of the various criteria of this model (this model is discussed more in the results). In contrast to both of these explanations, the distinctiveness heuristic account makes the opposite prediction. This view states that pictures will engender more distinctive recollections than red words, regardless of the relative levels of familiarity, because picture stimuli will still be more distinctive (i.e., have fewer overlapping features) than red word stimuli. As a result, false recognition of to-be-excluded items and of nonstudied lures should again be lower on the picture test.

Before proceeding, an important distinction to make is between quantitative differences in recollection (recalling more events within a stimulus class) and qualitative differences in recollection (the different types of features that can be recalled from different classes of stimuli or events). If repeating red words makes them easier to recall, then subjects might expect to recall more red words in this experiment than in the previous experiment (i.e., a quantitative difference in the number of recalled events). Based on this quantitative difference, one might predict that the absence of red word recall would be more diagnostic of nonoccurrence in this experiment, leading to fewer false alarms on the red word test in this experiment than in the last. Similarly, if repeating red words made them easier to recall than pictures in this experiment, then one might predict that false alarms on the red word test would be lower than false alarms on the picture test. These predictions are similar to those made by a criterion shift explanation, except the former relies on quantitative differences in recall, whereas the latter focuses on quantitative differences in familiarity. The point to stress here is that all of these assumptions are different from those made by the distinctiveness heuristic. Although quantitative differences in recall might influence monitoring processes, the distinctiveness heuristic focuses only on qualitative differences in recollection across types of events. In this regard, pictures should still elicit more distinctive recollections than red words, because repeating red words will not change the fact that pictures have more complex and unique perceptual features. If the distinctiveness heuristic is the critical mechanism through which subjects are monitoring false recognition in this task, then in contrast to these other explanations, the pattern of false alarms should be similar to that of the previous experiment.

Method

Subjects

Twenty-four Harvard University undergraduates participated for $10. Data from one subject were replaced because they were tested in the wrong experimental session.

Procedure

The procedures were similar to those of Experiment 1, except items studied as red words were repeated three times. Repetitions were randomly spaced throughout the corresponding section of the study phase (beginning, middle, or end). Under these repetition conditions, pilot testing indicated that it was not necessary to present each red word for the entire two seconds in order to achieve a greater hit rate to red words than to pictures. We therefore used a 1500 ms presentation duration for each red word, instead of the 2000 ms used in Experiment 1, which shortened the study phase by a few minutes (to a total of approximately 29 min). In the present experiment, each red word had been presented for a total of 4500 ms (summing the three repetitions) whereas each picture had been presented for 2000 ms. All other procedures were identical to those of Experiment 1.

Results and discussion

Recognition performance is presented in the right column of Table 1. The first point to notice is that red words were recognized more often than pictures, as indexed on the standard test (red word hits = .72, picture hits = .61, $t[23] = 3.97$, $SEM = .029$, $p < .01$). This same difference was also obtained on the criterial recollection tests (red word hits = .61, picture hits = .46, $t[23] = 4.72$, $SEM = .032$, $p < .001$). These findings indicate that repetition was successful at reversing the picture-superiority effect. We assume that repetition influenced both recollection and familiarity (e.g., Jacoby, 1999; see Yonelinas, 2002, for a review), and that red words were at least as familiar as pictures in this experiment (additional evidence for this assumption is provided in Experiment 3). Importantly, only those conditions where subjects should have relied on red word memory actually increased relative to Experiment 1 (i.e., both-hits and red word hits on the standard and red word tests, all p's $< .01$). There were no significant differences across experiments between any of the other conditions (all p's $> .05$). (These analyses are follow-ups to Item Type x Experiment ANOVAs conducted on each of the three tests, which revealed a significant interaction on the standard test and the red word test [both p's $< .001$], but not on the picture test [$F < 1$]). These data strongly suggest that the same processes were operating in the two experiments with the only exception being that memory for red words was greater in Experiment 2.

As in Experiment 1, subjects correctly responded to the target items on the criterial recollection tests. Red word hits (.61) were greater than picture FAs (.35) on the red word test, $t(23) = 5.48$, $SEM = .048$, $p < .001$, whereas picture hits (.46) were greater than red word FAs (.10) on the picture test, $t(23) = 8.81$, $SEM = .041$, $p < .001$. Also as in Experiment 1, subjects made false alarms based on familiarity (or illusory recollection) on the criterial recollection tests. Red word FAs (.10) were greater than new FAs (.01) on the picture test, $t(23) = 5.04$, $SEM = .018$, $p < .001$, and picture FAs (.35) were greater than new FAs (.11) on the red word test, $t(23) = 8.05$, $SEM = .029$, $p < .001$. Familiarity effects also were observed on both-hits, which were greater than picture hits on the picture test (means = .54 and .46, $t[23] = 3.52$, $SEM = .025$, $p < .01$) and red word hits on the red word test (means = .70 and .61, $t[23] = 2.74$, $SEM = .032$, $p < .05$). Finally, "both" hits were now greater on the red word test (.70) than on the picture test (.54), $t(23) = 4.85$, $SEM = .032$, $p < .001$, even though these were the same types of items. As in Experiment 1, this effect could be due to the use of some combination of recollection and familiarity on these tests (e.g., due to the distinctiveness heuristic, subjects would be less likely to rely on familiarity on the picture test relative to the red word test), and/or the use of multiple familiarity-based criteria. Overall, these patterns replicated those patterns observed in Experiment 1, with the only difference being that red words were more likely to be recognized in this experiment.

The most important finding from this experiment is that both types of false alarms were lower on the picture test than on the red word test. Red word FAs on the picture test (.10) were lower than picture FAs on the red word test (.35), $t(23) = 8.80$, $SEM = .028$, $p < .001$, and the same pattern was observed for new FAs (.01 vs. .11), $t(23) = 4.51$, $SEM = .023$, $p < .001$. This pattern runs contrary to the prediction made from a single familiarity-based criterion shift process (either in a unidimensional model, or couched within in a dual process model). According to this hypothesis, subjects in Experiment 1 used a more conservative response criterion on the picture test because pictures were more familiar than red words. Because red words were more familiar than pictures in the present experiment, subjects should have used a more conservative response criterion on the red word test in this experiment, and thus the pattern of false alarms should have reversed. Critically, this reversal did not occur. As can be seen from Fig. 3, the pattern of false alarms on the criterial recollection tests was the same across Experiments 1 and 2.

This pattern of false alarms is consistent with—and was predicted by—the distinctiveness heuristic. According to this view, pictures elicit more distinctive recollections than red words, regardless of the relative level of familiarity of the two types of stimuli. Repeating red words would not change the fact that pictures are more distinctive than red words (i.e., they contain more complex perceptual features that could be retrieved from memory). Even if repetition had made it easier to recall or recollect red words than pictures (a quantitative difference), it is the qualitative difference between the types

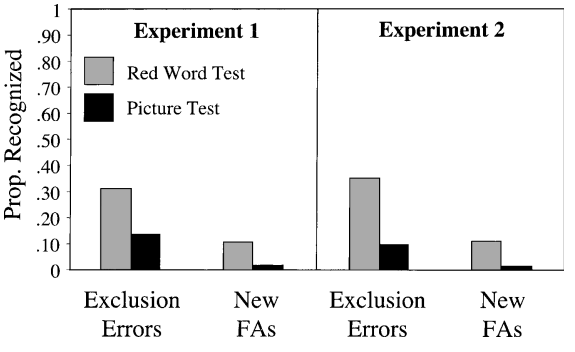

Fig. 3. Errors on the criterial recollection tests in Experiment 1 (red words presented once/pictures presented once) and Experiment 2 (red words presented thrice/pictures presented once). On the picture test, exclusion errors are false alarms to items studied only as red words, on the red word test, exclusion errors are false alarms to items studied only as pictures.

of expected recollections that is critical for the distinctiveness heuristic. As a result, subjects in both experiments would have expected more distinctive recollections on the picture test than on the red word test, leading to fewer errors on the picture test than the red word test in both experiments.

Multiple-criteria model

Can multiple response criteria on a familiarity (or strength) dimension explain the obtained pattern of results? To address this question, a multiple criteria model for the criterial recollection tests of Experiment 2 is presented in the bottom half of Fig. 2, using the same principles that were outlined in the context of Experiment 1. The nonstudied (NS) and picture (P) distributions are identical to those of Experiment 1, because as in that experiment, the former were never studied and the latter were only studied once. However, the red word (RW) and both (B) distributions are shifted to the right relative to their positions in Experiment 1, because red words were repeated in Experiment 2. The RW distribution is now drawn to be more familiar than the P distribution (reflecting the obtained reversal of the picture superiority effect), and the B distribution is strengthened to the same degree (because "both" items also were studied as repeated red words). In this experiment, subjects would only need one familiarity-based criterion on the red word test (the dashed line), in order to discriminate between lures (nonstudied items and to-be-excluded pictures) and targets (red words and "both" items). In contrast, three criteria would now be needed on the picture test, in order to discriminate between nonstudied lures and picture targets (p1), picture targets and to-be-excluded red words (p2), and to-be-excluded red words and to-be-included "both" items (p3).

These response criteria were drawn to roughly correspond to the obtained patterns of false recognition in Experiment 2. We do not consider how this model could explain the hit rates, because a dual process model could always appeal to the target recollection to explain discrepancies between the familiarity-based process and the hit rates. Our interest is in whether familiarity-based processes can explain the observed patterns of false recognition. If one allows some freedom in the setting of these criteria, then some of the false recognition results of the present experiment can be modeled (especially if hit rates do not have to be modeled). For instance, the criteria could be set in such a way that picture-FAs (on the red word test) are greater than red word-FAs (on the picture test). The fact that false alarms to new items were greater on the red word test than on the picture test also can be explained, although it involves different assumptions than were used in Experiment 1. In Experiment 1, the model explained this effect by assuming that the portions of the NS curve that fell in the "yes" regions for the weaker class of items (i.e., red words, which involved three criteria) was *greater* than the portion that fell in the "yes" regions for the stronger items (i.e., pictures, which involved one criterion). To explain this same result in Experiment 2, the model would have to offer the reverse explanation. That is, the portions of the NS curve that fell in the "yes" regions for the weaker class of items (i.e., pictures, which involved three criteria) was *lower* than the portion that fell in the "yes" regions for the stronger items (i.e., red words, which involved one criterion). In general, the fact that the false recognition data were identical across Experiments 1 and 2 is difficult to explain by the multiple-criteria model. As can be seen in Fig. 2, this model uses a different combination of criteria across the tests in the two experiments, so that identical false alarm rates across experiments would not necessarily have been predicted.

This model's clearest prediction concerns the pattern of false alarms to nonstudied items across Experiments 1 and 2. As can be seen in Fig. 1, the response criterion on the picture test of Experiment 1 (the dashed line in the top of the figure) is more liberal (or left) of the response criterion on the red word test of Experiment 2 (the dashed line in the bottom of the figure). This difference arises from the fact that repeated red words (in Experiment 2) were more familiar than pictures (which were equivalent across experiments), and thus afforded a more conservative familiarity-based response criterion than pictures. Said differently, the dashed criterion in either experiment is used to discriminate between the red word (RW) and picture (P) distributions, and so should be placed somewhere in between the means of the two distributions. Because the red word distribution shifts from the left to the right of the picture distribution across experiments, the dashed criterion also should shift from left to right across experiments.

This particular criterion shift predicts that nonstudied-FAs on the picture test of Experiment 1 should be greater than nonstudied-FAs on the red word test of Experiment 2. In fact, the opposite effect was found. False alarms to nonstudied lures were greater on the red word test of Experiment 2 (.11) than on the picture test of Experiment 1 (.02), $t(46) = 3.58$, $SEM = .024$, $p < .01$. This effect cannot be explained through floor effects in nonstudied FAs in Experiment 2, because as was discussed, these FAs were found to be significantly lower in other conditions of that experiment (plus, a replication of this inequality will be reported in Experiments 4 and 5). It also cannot be explained by appealing to differences in the shape or variance of the NS distribution across experiments—because these items were never studied, the distribution should be identical across experiments. Because none of the parameters in the multicriteria model can explain this inequality, additional processes need to be considered. We propose that this effect is most consistent with a distinctiveness heuristic account. Picture recollections were more detailed

than red word recollections, in both experiments, so that false alarms to lures on the picture tests always should have been lower than false alarms to lures on the red word tests.

Noncriterial recall

One final issue concerns the role of noncriterial recall on the criterial recollection tests (i.e., recalling a picture for an item on the red word test). By design, noncriterial recall did not afford a recall-to-reject strategy, because the inclusion of items studied as both red words and pictures made it so that the recall of one format (e.g., a picture) did not disqualify an item as also having been presented in the other format (e.g., a red word). Nevertheless, subjects may have used noncriterial recollection to inform the setting of their familiarity-based response criteria in a dual process account. For instance, if subjects could recall a red word on the picture test, then they could eliminate the "picture-only" distribution as relevant to the decision, so that they would only have to decide if the test item had come from the "red word-only" or "both" distributions (and reset their response criterion accordingly). Although such complicated criterion shifts are difficult to completely rule out, we do not believe that they are a critical concern. First, research by Stretch and Wixted (1998) and Morrell et al. (2002) suggests that subjects are reluctant to shift response criteria on a trial-by-trial basis. Second, even if our subjects did shift criteria from trial to trial, such criterion resettings would have had to occur more often or be more conservative for red words (on the picture test) than for pictures (on the red word test) in order to yield the obtained pattern of false alarms (red words $<$ pictures). We can think of no good reason that this would be the case. Finally, these shifts could not account for the pattern of false alarms to nonstudied items. These items were never presented and thus could not elicit recollection. Thus, the finding that false alarms to nonstudied lures was lower on the picture test than the red word test in both experiments cannot be explained by such criterion shifts.

Experiment 3

The critical finding from Experiments 1 and 2 was that, consistent with a distinctiveness heuristic account, false recognition was lower on the picture test than on the red word test. Familiarity-based criterion shift accounts could explain the results of Experiment 1, but not those of both Experiments 1 and 2. In Experiment 1 we assumed that pictures were more familiar than red words, because picture hits were greater than red word hits on the standard recognition test and on the criterial recollection tests. In Experiment 2 we assumed that repetition of red words had reversed these familiarity differences, because the red word hits were greater than picture hits. As a result, any explanation of false recognition that was based only on relative levels of familiarity would have difficulty explaining how the pattern of false alarms was identical across experiments.

Our interpretation of the results of Experiment 2 hinges on the assumption that repetition of red words made them more familiar than pictures. According to some dual process theories, though, it is possible that repetition may have increased recollection more than familiarity (e.g., Jacoby, 1999; Yonelinas, 2002). Thus, the obtained differences in hit rates (red words $>$ pictures) could have been driven entirely by recollection, while pictures may still have been more familiar than red words. This scenario would be problematic for our interpretation of the results. If pictures were still more familiar than red words, then the same familiarity-based explanation of false recognition could be proposed for each experiment, and similar results would be predicted. Note that this is only a problem if one assumes that recollection and familiarity provide independent bases of responding. If the two are combined into a single "strength" parameter, as in some unidimensional models of recognition, then our reversal of the picture-superiority effect in Experiment 2 unambiguously indicates that red words were stronger than pictures. Thus, a unitary strength-based criterion shift hypothesis can be rejected from these results alone.

The goal of Experiment 3 was to provide two manipulation checks for our claim that repeating red words made them more familiar than pictures. The study phase of this experiment was identical to that of Experiment 2. Subjects then received three tests. The first test was a standard recognition memory test, as in Experiment 2. The other two tests provided different means of measuring familiarity-based responding. The speeded test was similar in all respects to the standard test except subjects were forced to respond within a very brief window of time (700 ms). Prior research suggests that such rapid responses are more likely to be based on familiarity than recollection (e.g., Hintzman & Curran, 1994; McElree, Dolan, & Jacoby, 1999; see Yonelinas, 2002, for discussion). If repeating red words was successful in making them more familiar than pictures, then we expected to replicate our reversal of the picture superiority effect on the speeded test. The subjective test also was similar to the standard test, but instead of speeding responses we asked subjects to take their time and to make subjective judgments for each item that they recognized (e.g., Tulving, 1985). In particular, we asked them to indicate whether they "actually recollected" the presentation of the item in one of the study formats, or whether they "just knew" that the item was studied because it was very familiar. Using the independent-remember-know procedure (IRK, Yonelinas, 2002), which is most relevant to the dual process theory

under consideration, Jacoby, Jones, and Dolan (1998) found that familiarity estimates were sensitive to changes in familiarity due to study repetitions of words. We therefore used this procedure as a second means of estimating the relative familiarity of test items corresponding to red words and pictures in our repetition condition.

Subjects

Twenty-four Harvard undergraduates participated in each testing condition for $10.

Methods

The study phase of this experiment was identical to that of Experiment 2. Stimuli were rotated through the four item types (both, picture, red word, and new) and were counterbalanced across the three test blocks, and the order of the speeded and subjective tests was counterbalanced across subjects. On the standard test, subjects were given the same instructions as in the previous experiments. On the speeded test, subjects were told that we were interested in speeded recognition decisions. With the initiation of each test trial (by pressing the spacebar), a series of simultaneous auditory beats and visual fixation cues was presented to establish a response tempo (700 ms), using recognition-tempo procedure described by Balota, Burgess, Cortese, and Adams (2002). On the third beat, the test item appeared in the center of the computer screen, and subjects were told to respond on the fourth beat. They were told to press the "yes" button if they thought the test word had been presented in the study phase (regardless of whether it was studied as a red word or a picture), and the "no" button if they thought it was nonstudied. If they responded within 600 ms a "TOO FAST" error message appeared, and if they responded after 800 ms a "TOO SLOW" error message appeared. If they responded on time (600–800 ms) then a "GOOD!" message appeared.

On the subjective test, subjects were told that we were interested in how they made their recognition decision. These instructions were modeled after the "remember"/"know" distinction (e.g., Tulving, 1985), but were modified so that the remember judgment would only reflect the recollection of a red word or a picture. For each test item, instead of making a "yes"/"no" judgments, subjects made a "Just Know," "Actually Recollect," and "New" decision. They were instructed to press "JK" if they were sure that the item had been presented because it was very familiar, but they did not have a vivid memory of a red word or a picture, and to press "AR" if they had a vivid recollection of the test word being presented as either a red word or a picture at study (i.e., they could retrieve a clear image of the red word or picture on the computer screen). A "guess" judgment (e.g., Gardiner & Conway, 1999) was not included because we did not want subjects to guess "yes." Note that "yes" judgments based on noncriterial recollection (e.g., remembering a personal association that was made to a studied item) did not easily fit into either of the subjective categories. It was unclear to us whether noncriterial recollection would occur often, and if it did, whether it would have influenced our criterial recollection tests (in Experiment 2) via recollection or familiarity (e.g., Yonelinas & Jacoby, 1996). We therefore left it to subjects to decide which judgment to use in these cases.

Results and discussion

The main results of this experiment are summarized in Fig. 4. Results of the standard test (leftmost bars) replicated those of the standard test in Experiment 2. Both-hits (.89) were greater than red word hits (.80), red word hits were greater than picture hits (.70), and all hit rates were greater than new FAs (.12), all p's $< .01$. Results form the speeded test can be found in the second set of bars. Response latencies indicated that subjects were very good at keeping the 700 ms response tempo (mean latency across all item types and responses = 704 ms, compared to 1079 ms in the standard test and 1604 on the subjective test). To avoid inevitable issues of response selection, all trials were included in the analysis regardless of whether subjects responded on time. As can be seen from the figure, hits to all types of targets were lower on the speeded test (mean = .70) than on the standard test (.80), and new FAs were greater on the speeded test (.24) than on the standard test (.14), all p's $< .01$. This pattern demonstrates a significant speed-accuracy tradeoff, and is consistent with the idea that speeding subjects' responses forced them to rely less on

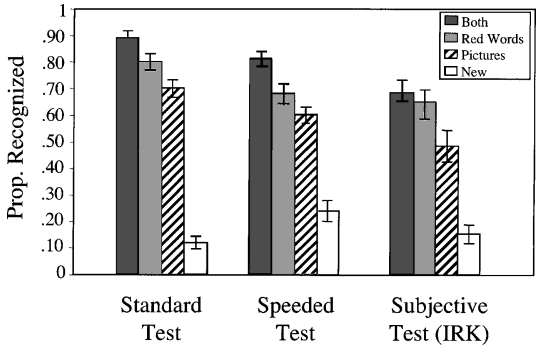

Fig. 4. Results from the three tests of Experiment 3. For the subjective test, the raw proportions of familiar ("JK") and recollect ("AR") judgments, respectively, were .18 and .70 for both hits, .31 and. 51 for red word hits, .20 and .53 for picture hits, and .11 and .04 for nonstudied FAs. Estimates of familiarity on the subjective test were calculated using the IRK procedure (see text). Bars represent standard error of the mean.

recollection (thereby lowering hits) and more on familiarity (thereby increasing false alarms). Most important, hits to red words (.68) were greater than hits to pictures (.60) on the speeded test, $t(23) = 3.58$, $SEM = .024$, $p < .01$. Assuming that subjects relied mostly on familiarity on the speeded test, these results indicate that repetition made red words more familiar than pictures.

Turn lastly to the results from the subjective test. If one infers the recognition rates on this test (p"old" = p"AR" + p"JK", see figure note), then the results replicated those of the standard recognition test (both hits = .88, word hits = .82, picture hits = .73, and new FAs = .15, compared to .89, .80, .70, and .12 on the standard test, respectively). The estimates of familiarity that were based on the subjective judgments can be found in the rightmost bars of the figure. These estimates used the independent-remember-know procedure, which defines familiarity (F) as the probability of making a "just know" judgment in proportion to the number of opportunities to make such a judgment (F = "JK"/[1-"AR"]). If it is assumed that recollection and familiarity are independent, then this calculation adjusts for the fact that familiarity judgments can only be made when recollection (measured as an "AR" judgment) is absent, even though a test item might elicit both recollection and familiarity. A few subjects always gave "AR" judgments for some test items ($n = 4$ for both items, 1 for red words, and 2 for pictures), which paradoxically translates into zero estimates of familiarity with this procedure. These cells were excluded from the familiarity estimates reported in the figure.

As can be seen from the figure, the familiarity estimates further confirmed that red words were more familiar than pictures. Excluding those three subjects who never made "JK" responses to red words or pictures, the red word estimate (.68) was significantly greater than the picture estimate (.50), $t(20) = 3.62$, $SEM = .051$, $p < .01$. Even when these three subjects were included (with "0" estimates of familiarity), the means were different (.63 and .45, respectively, $t[23] = 3.31$, $SEM = .053$, $p < .01$). The difference between red words and pictures also was obtained in the raw (unadjusted) proportion of "just know" judgments (means = .31 vs. .20, $t[23] = 2.94$, $SEM = .037$, $p < .01$), and also when these judgments were expressed as a proportion of overall recognition (i.e., F = "JK"/["JK"+"AR"]; means = .40 and .30, $t[23] = 2.10$, $SEM = .044$, $p < .05$). In short, regardless of how one estimates familiarity from these subjective judgments, estimates for red words were significantly greater than those for pictures.

The results of the speeded test and the subjective test provide converging evidence that our repetition manipulation made red words more familiar than pictures. These findings bolster the conclusion that, if anything, subjects should have used a more conservative familiarity-based criterion on the red word test of Experiment 2, relative to the picture test of Experiments 1 or 2. Criterion shift accounts are therefore unable to explain how false recognition was lower on the picture test than on the red word test in Experiment 2. In contrast, these results were predicted by the distinctiveness heuristic hypothesis, which focuses on qualitative differences in recollective expectations. Finally, notice that "AR" judgments were greater for "both" items (.70) than for pictures (.53) or red words (.51), which in turn were greater than those for nonstudied items (.04). The fact that "AR" judgments were similar for pictures and red words suggests that these two classes of stimuli elicited the same quantity (or amount) of recollection. This result is consistent with the distinctiveness heuristic idea that the picture/word effect on false recognition is due to qualitative differences in the types of recollections that these two types of stimuli afford, as opposed to quantitative differences in recollection.

Experiments 4 and 5

The results of the previous three experiments were more consistent with a recollection-based distinctiveness heuristic than with familiarity-based criterion shifts. Experiments 4 and 5 provided an additional test between these two accounts. Experiments 4 and 5 were similar to Experiments to 1 and 2, respectively, with the only difference being that we tested memory with more typical source tests as opposed to criterial recollection tests. In Experiment 4 red words were studied once, and in Experiment 5 red words were studied three times. In each experiment, only one test was given, in which subjects had to attribute each item to one of the four possible sources (both, picture only, red word only, and nonstudied).

In order to model performance on a four-alternative source test, a multiple-criteria familiarity model would need three response criteria. Fig. 5 provides such a model for the source tests in Experiments 4 (top) and 5 (bottom). For consistency, the distributions are drawn as they were in Fig. 3 (for Experiments 1 and 2). In order to discriminate between the four classes of items, a response criterion is placed somewhere between each adjacent set of distributions. The model in Fig. 5 generates straightforward predictions for familiarity-based source monitoring errors. For example, consider source judgments for nonstudied items (NS). Most of these items will fall to the left of the first criterion, eliciting correct "nonstudied" judgments, but some of these items will fall to the right of this criterion. In Experiment 4, because the red word distribution is the next distribution, more of the items from NS will fall within the "red word" response range than in the "picture" response range (which is farther to the right), eliciting more "red word" errors than "picture" errors. This pattern should

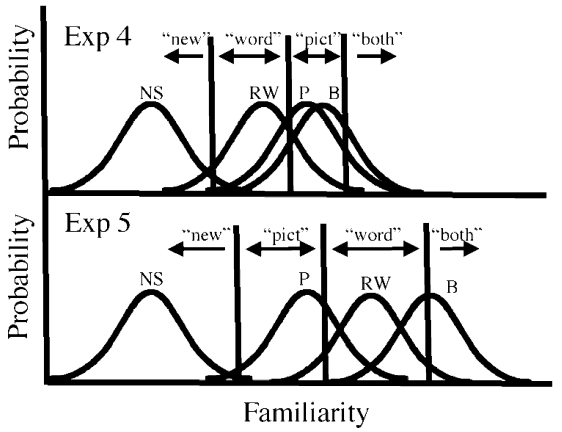

Fig. 5. An idealized familiarity-based model of the source memory tests. Distributions for Experiment 4 (A) and Experiment 5 (B) are identical to those in Experiment 1 and 2, respectively (see Fig. 2). Subjects use three response criteria to distinguish between the four sources.

reverse in Experiment 5. Here the order of the red word (RW) and picture (P) distributions is reversed, so that nonstudied items will be more likely to fall within the "picture" response range than within the "red word" response range. Importantly, this same pattern would be predicted even from a dual process model that allowed recollection to contribute to correct source attributions, as long as source errors are thought to be driven solely by familiarity-based processes. (Again, because some items were studied as both red words and pictures, a recall-to-reject strategy would not contribute to performance, and in any event, such a process could not operate for nonstudied lures).

The distinctiveness heuristic makes a different prediction for source memory errors. According to this theory, items that only elicit familiarity are less likely to be attributed to the format that is expected to elicit more distinctive recollections. As a result, source errors for nonstudied lures should be more likely to take the form of "red word" judgments than of "picture" judgments, in either Experiment 4 or Experiment 5. Along these lines, Dodson and Schacter (2002a) noted that the picture/word effect in false recognition is related to the "it-had-to-be-you" effect in source monitoring research, in which subjects attribute familiar events (whose source cannot be recollected) to the source that is thought to be less memorable (e.g., Johnson, Raye, Foley, & Foley, 1981; Hicks & Marsh, 2001). In those studies, the two potential sources were self-generated words and words presented by the experimenter. Because self-generation was more memorable, subjects decide that uncertain events probably were presented by the external source. Under this framework, memorability is not always conceived as quantitative memory "strength," but also can refer to the qualitatively different features that could be recollected from different types of events (see also Bink, Marsh, & Hicks, 1999; and the multidimensional model of Banks, 2000). If this interpretation is true, and if subjects expect to recollect more distinctive information from pictures than red words (as predicted by the distinctiveness heuristic), then we should find an "it-had-to-be-a-word" effect on source attributions regardless of whether red words were more or less familiar than pictures. For convenience, we report the methods and results of the two experiments together.

Method

Subjects

Twelve Harvard University undergraduates participated for $10 in each experiment. Data from two subjects were replaced in Experiment 5, one because they were tested in the wrong counterbalancing condition, and the other due to computer failure.

Procedure

The study phases of Experiments 4 and 5 were identical to those of Experiments 1 and 2, respectively. In Experiment 4 red words and pictures were presented once, in Experiment 5 red words were presented three times and pictures were presented once. Immediately after the study phase, subjects were given a source memory test. All of the items were presented on a single test, and items were counterbalanced across the four conditions (both, picture, red word, and new). Subjects had to decide whether each test word had been studied as a red word, a picture, both, or was nonstudied, and indicated their response by pressing one of four buttons. The test was self-paced.

Results and discussion

Source attributions in both experiments can be found in Table 2. The main point to notice is that, in both experiments, erroneous attributions were more likely attributed to the "red word" source than to a source associated with a picture (i.e., "picture" or "both" sources). This effect is most easily seen for the nonstudied lures. Although subjects often attributed these items to the correct "nonstudied" source (.83 in Experiment 4 and .82 in Experiment 5), erroneous attributions to the "red word" source in Experiment 4 (.13) were more likely than to the "picture" source (.03; $t[11] = 5.32$, $SEM = .019$, $p < .001$) or to the "both" source (.02; $t[11] = 5.01$, $SEM = .021$, $p < .001$). The same pattern was found in Experiment 5, where nonstudied item attributions to the "red word" source (.16) were greater than those to the "picture" source (.01; $t[11] = 3.04$, $SEM = .049$, $p < .05$) and the "both" source (.01; $t[11] = 3.43$, $SEM = 4.39$, $p < .01$). These

Table 2
Mean proportion of each source attribution for each item type in Experiments 4 and 5

	Experiment 4 (Red words 1×)	Experiment 5 (Red words 3×)
Both items		
"Both"	.34 (.04)	.52 (.05)
"Red word"	.19 (.03)	.28 (.06)
"Picture"	.22 (.04)	.12 (.04)
"New"	.25 (.05)	.09 (.02)
Picture items		
"Both"	.19 (.04)	.16 (.03)
"Red word"	.19 (.03)	.24 (.03)
"Picture"	.33 (.07)	.36 (.07)
"New"	.29 (.04)	.24 (.03)
Red word items		
"Both"	.06 (.02)	.05 (.01)
"Red word"	.46 (.04)	.72 (.03)
"Picture"	.05 (.02)	.02 (.01)
"New"	.44 (.04)	.21 (.03)
New items		
"Both"	.02 (.01)	.01 (.00)
"Red word"	.13 (.02)	.16 (.05)
"Picture"	.03 (.01)	.01 (.00)
"New"	.83 (.04)	.82 (.05)

Note. For each item type, the four attribution proportions sum to 1. Standard errors of each mean are in parentheses.

patterns demonstrate an "it-had-to-be-a-word" effect, where erroneous source attributions are more likely to be made to the red word source than to the picture source. Models that assume that source errors are influenced only by familiarity (as in Fig. 5) cannot explain these results, and instead these results suggest that a recollection-based distinctiveness heuristic had influenced source decisions.

These "it-had-to-be-a-word" effects also were found for the other types of source errors. In Experiment 4, pictures were erroneously attributed to the "red word" source (.19) more often than red words were attributed to the "picture" source (.05; $t[11] = 6.02$, $SEM = .024$, $p < .001$). Pictures also were attributed to the "both" source more often than were red words (.19 vs. .06, $t[11] = 4.25$, $SEM = .031$, $p < .01$). Both of these findings demonstrate that subjects were more likely to claim that a picture had been studied as a red word than vice versa. Similarly, in Experiment 5, pictures attributed to the "red word" source (.24) were greater than red words attributed to the "picture" source (.02; $t[11] = 6.03$, $SEM = .036$, $p < .001$), and pictures were more likely to be attributed to the "both" source (.16) than were red words (.05; $t[11] = 4.24$, $SEM = .027$, $p < .01$). A similar effect probably contributed to source attributions for "both" items, although it is difficult to tell from the data because these items were presented as both red words and pictures, and thus an erroneous source judgment of "picture only" or "red word only" could have been based on correct recall of one of those sources.

Comparing criterial recollection to source tests

It is well established that memory performance can differ depending on how source monitoring questions are asked (e.g., Dodson & Johnson, 1993; Lindsay & Johnson, 1989; Marsh & Hicks, 1998). In the present experiments, one could conceptualize criterial recollection tests as source memory tests in which memory for only one of the formats is queried. It was unclear to us whether (or how) this difference would affect performance. By one view, only having to consider one format (i.e., pictures on the picture test) might be "easier" than having to simultaneously consider both formats. That is, on the criterial recollection test, subjects could narrow their retrieval orientation and selectively "search" memory for the relevant format, leading to greater accuracy on the criterial recollection tests (see Marsh & Hicks, 1998). By another view, source judgments might lead to better performance than binary decisions (e.g., Dodson & Johnson, 1993; Multhaup, 1995), because source judgments force subjects to consider all possible sources of information for every decision.

To directly compare the results from the source tests to the criterial recollection tests (Experiments 1 and 2), we estimated how subjects in Experiments 5 and 6 would have responded had we given them a standard recognition test or the criterial recollection tests, using source judgments as an index of subjective beliefs (e.g., Bink et al., 1999). To infer recognition on the standard test, studied items that were attributed to a studied source (correct or incorrect) were tallied as hits, and nonstudied items that were attributed to a studied source were tallied as false alarms. To infer recognition on the red word test, we assumed that source judgments of "red word" or "both" would have led to positive responses. Thus, these judgments were summed to infer hits rates for targets (i.e., red words or "both" items) and false alarm rates for lures (i.e., pictures and nonstudied words). Inferred recognition rates for the picture test were calculated in this same way, with red word and picture judgments reversed. The resulting estimates can be found in Table 3, with corresponding data from Experiments 1 and 2 in parentheses.

In general, the inferred recognition data followed similar patterns as the actual recognition data, especially with regard to the patterns of false alarms. In experiment 4, inferred false alarms to red words on the picture test (.10) were lower than inferred false alarms to pictures on the red word test (.38), $t[11] = 6.83$, $SEM = .04$, $p < .001$, and inferred nonstudied FAs were lower on the picture test (.04) than on the red word test (.14),

Table 3
Inferred recognition means from the source test in Experiments 4 and 5

	Experiment 4 (Red words 1×)	Experiment 5 (Red words 3×)
Standard test		
Both hits	.75 (.70)	.91 (.82)
Red word hits	.56 (.45)	.79 (.72)
Picture hits	.71 (.66)	.76 (.61)
New FAs	.17 (.10)	.18 (.10)
Red word test		
Both hits	.53 (.46)	.80 (.70)
Red word hits	.52 (.40)	.77 (.61)
Picture FAs	.38 (.31)	.40 (.35)
New FAs	.14 (.11)	.17 (.11)
Picture test		
Both hits	.56 (.56)	.63 (.54)
Red word FAs	.10 (.14)	.07 (.10)
Picture hits	.52 (.51)	.52 (.46)
New FAs	.04 (.02)	.02 (.01)

Note. Actual recognition from Experiments 1 (words 1×) and 2 (words 3×) are in parentheses. FAs, false alarms.

$t(11) = 5.32$, $SEM = .019$, $p < .001$. Similarly, in Experiment 5, inferred red word FAs on the picture test (.07) were lower than inferred picture FAs on the red word test (.40), $t[11] = 6.99$, $SEM = .047$, $p < .001$, and inferred nonstudied FAs were lower on the picture test (.02) than the red word test (.17), $t[11] = 3.04$, $SEM = .048$, $p < .05$. These patterns replicate those observed in Experiments 1 and 2, and highlight how fundamental differences between classes of stimuli can influence false recognition and source misattribution errors in a similar way.

The major difference between actual and inferred recognition was that the latter tended to be greater than the former on some tests. From Table 3 it can be seen that, on the standard and red word tests, inferred recognition (from the source tests) tended to be greater for all item types than was actual recognition, but these differences were minimal on the picture test. To investigate these effects we conducted an Item Type × Repetition × Response Format (inferred recognition vs. actual recognition) ANOVA on each of the three tests (standard, red word, and picture). On the standard test, there was a main effect of item type, a main effect of repetition, and an interaction between the two (all p's $< .01$), which reflects differences that have already been discussed in the context of Experiment 2. More important, there was a main effect of response format, $F(1, 68) = 8.85$, $MSE = .051$, $p < .01$, demonstrating that there were more positive responses to all item types on the source test (Experiments 3 and 4) than on the standard "yes"/"no" test (Experiments 1 and 2). There were no other significant effects or interactions. Analysis of the red word test yielded an identical pattern of results. There was a main effect of item type and repetition, and a significant interaction between the two (all p's $< .001$). There also was a main effect of response format, $F(1, 68) = 10.43$, $MSE = .042$, $p < .01$, and no other significant effects or interactions. The picture test was the only test that did not show effects of response format. On this test, there was only a main effect of item type ($p < .001$), and no other significant effects.

We did not expect to find such an effect of response format, but other findings in the literature are relevant. Hicks and Marsh (2001) reported very similar findings using a verbal variant of the DRM task. In three experiments, they found that inferred recognition judgments (on a source test) led to greater claims that targets and lures were studied compared to a standard recognition test. One explanation that they offered is that giving subjects three "old" response options (i.e., the three sources) encouraged them to search for evidence that an item had occurred in one of the studied sources (rightly or wrongly), compared to simply giving them a binary "yes"/"no" decision. This could explain our results from the standard test (which are analogous to those of Hicks & Marsh, 2001) and the red word test. The failure to find such an effect for the picture test suggests that the more distinctive the to-be-remembered events, the less likely performance will be affected by such response-option effects. We admit that these conclusions are only speculative, and future research will be needed to more thoroughly understand the cause of these response format effects. The main point we wish to make from this comparison is that, even though the source tests do not appear to be identical to the criterial recollection tests in all respects, the critical pattern of false alarms from the criterial recollection tests was replicated on the source tests.

General discussion

In the present experiments we used instructions to manipulate the recollective demands of the recognition test so that we could investigate the effects of recollective expectations on false recognition. The main results were that, regardless of the relative familiarity or "strength" of the stimuli, subjects made fewer memory errors (i.e., false recognition or false source attributions) when claiming something occurred as a picture than as a red word. Decision processes based on the level of familiarity or strength of the stimuli (conceptualized as unidimensional criterion setting) cannot easily explain these results, but these results were consistent with the distinctiveness heuristic hypothesis. This hypothesis proposes that subjects rely on expectations about the quality of information that should be recollected from pictures to help make their recognition decisions (e.g.,

Dodson & Schacter, 2002b; Schacter et al., 1999). Regardless of the familiarity of the questionable event, if it fails to evoke distinctive pictorial recollections then this failure is taken as evidence that it did not occur. Unlike a recall-to-reject process, which is based on the recollection of information that *disqualifies* an event as having occurred (due to task-specific exclusion rules), the distinctiveness heuristic is based on the absence of expected recollections, and thus is only *diagnostic* of nonoccurrence (see Gallo, 2004, for discussion of these two types of recollection-based monitoring).

The notion that subjects relied on recollection-based expectations to inform their memory decisions can be couched within the source monitoring framework (e.g., Johnson & Raye, 1981; Johnson et al., 1993; Mitchell & Johnson, 2000). According to this framework, qualitatively different types of information (or features) can be recalled or recollected for different types of events (e.g., Bower, 1967; Underwood, 1969), and the attribution of a questionable event to a source depends, in part, on the amount of information retrieved for an event and also on the relative weightings of the different types of information in the decision (for a multidimensional SDT model that incorporates similar views, see Banks, 2000). This framework predicts that the qualitative difference between picture and word recollections should influence a variety of memory judgments, and not just "yes"/"no" recognition decisions. Consistent with this view, the source monitoring tests in Experiments 4 and 5 demonstrated an "it-had-to-be-a-word" effect on source misattributions following picture and word study. Systematic differences existed between the quality of information that could be recollected from pictures and red words, and subjects used this information to inform diagnostic monitoring processes (the distinctiveness heuristic). Of course, source memory tests such as these involve the simultaneous consideration of several sources, so that the use of different recollective expectations for different sources can only be inferred from the results. The criterial recollection tests used in the present study were designed to more clearly isolate the distinctiveness heuristic process, by directly querying one source at a time and by precluding a recall-to-reject strategy.

Although the source-monitoring framework is relevant, we are not advocating any particular model of recollection. As discussed in the Introduction, the distinctiveness heuristic can be applied to dual process models or to multidimensional models. Many (but not all) dual-process models assume that recollection is a threshold or all-or-none process, so that one either does or does not have the subjective experience of recollecting an event. Other models, such as multidimensional SDT models, instead propose that recollection can be a continuous process, so that one can have the experience of recollecting an event to varying degrees (e.g., recollecting a varying amount of its features, or recollecting each feature to a varying degree). None of these assumptions about the nature of recollection is incompatible with the logic of a distinctiveness heuristic. The distinctiveness heuristic focuses on the quality of recollected information—the number of unique perceptual features that could be recalled for an event—regardless of whether the recollection of these features occurs in an all-or-none or a continuous fashion, and regardless of whether events from one class are recalled more or less often than those from another. With regard to qualitative differences, it is clear that pictures afford the recollection of more unique and complex perceptual features than words, on average. Our data indicate that these qualitative differences can influence recollective expectations, and that expecting more distinctive recollections can reduce false recognition.

A more important question is how the distinctiveness heuristic reduces false recognition: Does it facilitate post-retrieval process, or does it provide a more optimal pre-retrieval process? This distinction is most easily conceptualized within a dual-process framework, although it might apply to other frameworks, too. According to the post-retrieval view, the questionable event feels equally familiar on the picture test and the red word test, but this familiarity is more likely to be successfully rejected on the picture test due to decision processes that are based on recollective expectations. To use the terminology of Jacoby, Kelley, and McElree (1999), this would be a "late correction" form of memory monitoring, because the questionable event is experienced as familiar and subsequently needs to be monitored via post-retrieval processes. An alternative interpretation of these effects is akin to what Jacoby et al. called an "early-selection" form of monitoring. According to this interpretation, the different criterial recollection tests might afford different retrieval orientations (cf. Herron & Rugg, 2003). If subjects are better at constraining recollection for more distinctive information (e.g., picture recollections on the picture test), then the feeling of familiarity arising from presentation in the noncriterial format (e.g., red words on the picture test) might not be as great. Said differently, when the relevant source is more distinctive, the questionable event might not seem "as familiar" as when the relevant source is less distinctive, and thus false recognition would be avoided by a lack of familiarity. Such a possibility is consistent with attribution-based theories of familiarity, which propose that the subjective experience of familiarity (or illusory recollection) can be influenced by expectations (e.g., Jacoby, Kelley, & Dywan, 1989; see also Gallo & Roediger, 2003; Whittlesea & Williams, 1998).

Our experiments were not designed to test between these two alternatives, but they do provide some insights. If the distinctiveness heuristic operated through

post-retrieval monitoring processes, then one might not expect differences in false recognition suppression across the criterial recollection tests and source tests. Subjects should have been able to recall the same types of information regardless of the type of test, and hence could have recruited the same post-retrieval monitoring processes in either case. In contrast, if retrieval orientation and pre-retrieval processes were the critical factor, then one would expect that the criterial recollection tests would elicit more false recognition suppression than the source test. This prediction is based on the idea that criterial recollection tests could potentially constrain retrieval to a single source, whereas source tests require the simultaneous consideration of all sources. Our finding of similar false recognition patterns between the criterial recollection tests and the inferred recognition on the source tests is consistent with the post-retrieval monitoring view. On the other hand, there were some effects of test format on performance (i.e., inferred recognition led to more "old" responses than did actual recognition on the standard and red word tests), and these might reflect differences in retrieval orientation. Further work aimed at directly testing these ideas is needed.

We conclude by commenting on the generality of the present results. First, although we were able to rule out familiarity-based criterion shift accounts of the present results, these results do not speak directly to the mechanism of the picture/word effect in false recognition that has been observed in the DRM task (e.g., Israel & Schacter, 1997) and others (e.g., Dodson & Schacter, 2002a, 2002b). Familiarity-based criterion shifts could still have played some role in those other tasks. That being said, given that the subjects in the present experiment used the distinctiveness heuristic in two different situations (criterial recollection and source tests), we see no reason to believe that subjects could not take advantage of this process in other tasks, too. To argue otherwise, one would need to assume that subjects do not use this sort of recollection-based rejection process when taking a standard "old"/"new" recognition test for pictures, even though they could, and instead rely only on strength or familiarity. This view assumes that, even when stimuli afford more distinctive recollections, subjects use familiarity-based responding because it is quicker or easier than recollection-based rejection. Although we are sympathetic to this view, the data from our criterial recollection tests indicate otherwise. Even when red words were more familiar than pictures (Experiment 2), so that familiarity-based discriminations should have been easier on the red word test (see Fig. 2, bottom panel), subjects were quicker to respond to all items on the picture test (mean = 1218 ms) than on the red word test (1407 ms), $p < .001$. These latency differences (picture test faster than red word test) were observed for every item type, including correct exclusions (1251 and 1491 ms, $p = .02$) and correct rejections of nonstudied words (1153 vs. 1253, $p = .09$). Further, on an open-ended questionnaire given at the end of the experiment, most of our subjects in Experiment 2 ($n = 22$) indicated that the picture test was "easier" than the red word test (the other two subjects did not complete the questionnaire).

If subjects prefer to rely on familiarity, and if familiarity discrimination should have been easier on the red word test, then why was the picture test quicker and easier than the red word test? Although these data are only suggestive, we believe that they indicate that subjects do not always prefer to rely on familiarity or strength. Instead, when stimuli afford richly detailed recollections that can inform memory decisions, people naturally take advantage of these differences to facilitate the rejection of lures. Of course, when the recognition situation involves only the discrimination of stimuli that afford relatively impoverished, nondistinctive, or homogenous recollections (e.g., nonsense syllables or superficially processed words), familiarity-based responding might be the major determinant of performance (see Donaldson, 1996, for discussion). We suspect, though, that the types of non-laboratory events that we remember from our daily lives afford richer recollections, on average, than do these other types of events. As such, recollection-based monitoring processes such as the distinctiveness heuristic should play an important role in many memory decisions.

A second issue of generality is that the diagnostic monitoring processes that underlie the picture/word distinctiveness heuristic may be involved across a variety of stimulus dimensions that could potentially afford such monitoring. For instance, Dobbins, Kroll, Yonelinas, and Liu (1998) have shown that performing different types of cognitive operations on words can later help to reject words from a to-be-excluded source. The idea was that the absence of memory for those cognitive operations indicated that the word could not have occurred in that source (see also Dodson & Schacter, 2001; Hunt, 2003; Johnson et al., 1981). Similarly, Brown, Lewis, and Monk (1977) found that subjects were unlikely to falsely recognize their own name as having occurred on a list of names. These authors proposed that subjects expected their name to be very memorable, and this allowed them to use metacognitive strategies similar to the distinctiveness heuristic to avoid false recognition (see also Groninger, 1976). In a final example, Pesta, Murphy, and Sanders (2001) found that subjects were less likely to falsely recognize emotionally valenced words than neutral words, potentially due to the expected levels of memorability of emotional stimuli (see also Kensinger & Corkin, in press). Not all "strength" effects on false recognition are necessarily due to monitoring processes based on expected memorability (see Wixted,

1992, for a rejection of expected memorability as an explanation of word frequency effects on false recognition), but given the present results, and given other results reviewed here, monitoring processes based on expected recollections must be considered a viable explanation to many of these effects.

Finally, we do not wish to leave the reader with the impression that distinctive or complex events will never be falsely remembered, as there is ample evidence that they can be (e.g., Garry, Manning, Loftus, & Sherman, 1996; Lindsay, Hagen, Read, Wade, & Garry, 2004; Neisser & Harsch, 1992). Numerous processes contribute to the creation of false memories, such as illusory recollection, imagination, and cognitive constraints at the time of retrieval, all of which could potentially thwart monitoring processes. Further, exactly what qualifies as a "distinctive" event no doubt depends on a complex interaction between the types of events under scrutiny and the context in which they occur (e.g., Hunt, 2003). The present results instead imply that, all other factors being equal, questionable events that are expected to elicit more perceptually detailed recollections are less likely to be falsely remembered. More generally, this study adds to a growing body of research that indicates that recollective expectations play a significant role in memory decisions, in addition to the classic notions of recollection, familiarity, and familiarity-based criterion shifts.

References

Balota, D. A., Burgess, G. C., Cortese, M. J., & Adams, D. R. (2002). The word-frequency mirror effect in young, old, and early-stage Alzheimer's disease: Evidence for two processes in episodic recognition performance. *Journal of Memory and Language, 46*, 199–226.

Banks, W. P. (2000). Recognition and source memory as multivariate decision processes. *Psychological Science, 11*, 267–273.

Bink, M. L., Marsh, R. L., & Hicks, J. L. (1999). An alternative conceptualization to memory "strength" in reality monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 804–809.

Bower, G. H. (1967). A muli-component theory of the memory trace. In K. W. Spencer & J. T. Spence (Eds.), *The psychology of learning and motivation* (Vol. I, pp. 229–325). New York: Academic Press.

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.

Brown, J., Lewis, V. J., & Monk, A. F. (1977). Memorability, word frequency and negative recognition. *Quarterly Journal of Experimental Psychology, 29*, 461–473.

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology, 58*, 17–22.

Dobbins, I. G., Kroll, N. E. A., Yonelinas, A. P., & Liu, Q. (1998). Distinctiveness in recognition and free recall: The role of recollection in rejection of the familiar. *Journal of Memory and Language, 38*, 381–400.

Dodson, C. S., & Johnson, M. K. (1993). Rate of false source attributions depends on how questions are asked. *American Journal of Psychology, 106*, 541–557.

Dodson, C. S., & Schacter, D. L. (2001). "If I had said it I would have remembered it": Reducing false memories with a distinctiveness heuristic. *Psychonomic Bulletin & Review, 8*, 155–161.

Dodson, C. S., & Schacter, D. L. (2002a). Aging and strategic retrieval processes: Reducing false memories with a distinctiveness heuristic. *Psychology and Aging, 17*, 405–415.

Dodson, C. S., & Schacter, D. L. (2002b). When false recognition meets metacognition: The distinctiveness heuristic. *Journal of Memory and Language, 46*, 782–803.

Donaldson, W. (1996). The role of decision processes in remembering and knowing. *Memory & Cognition, 24*, 523–533.

Gallo, D. A. (2004). Using recall to reduce false recognition: Diagnostic and disqualifying monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 120–128.

Gallo, D. A., & Roediger, H. L. (2003). The effects of associations and aging on illusory recollection. *Memory & Cognition, 31*, 1036–1044.

Gallo, D. A., Roediger, H. L., & McDermott, K. B. (2001). Associative false recognition occurs without strategic criterion shifts. *Psychonomic Bulletin & Review, 8*, 579–586.

Gardiner, J. M., & Conway, M. A. (1999). Levels of awareness and varieties of experience. In B. H. Challis & B. M. Velichovsky (Eds.), *Stratification in cognition and consciousness* (pp. 237–254). Amsterdam: John Benjamin Publishing.

Garry, M., Manning, C. G., Loftus, E. F., & Sherman, S. J. (1996). Imagination inflation: Imagining a childhood event inflates confidence that it occurred. *Psychonomic Bulletin & Review, 3*, 208–214.

Ghetti, S. (2003). Memory for nonoccurences: The role of metacognition. *Journal of Memory and Language, 48*, 722–739.

Glanzer, M. A., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory & Cognition, 13*, 8–20.

Groninger, L. D. (1976). Predicting recognition during storage: The capacity of the memory system to evaluate itself. *Bulletin of the Psychonomic Society, 7*, 425–428.

Herron, J. E., & Rugg, M. D. (2003). Retrieval orientation and the control of recollection. *Journal of Cognitive Neuroscience, 15*, 843–854.

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, and Cognition, 27*, 375–383.

Hintzman, D. L., & Curran, T. (1994). Retrieval dynamics of recognition and frequency judgments: Evidence for separate processes of familiarity and recall. *Journal of Memory and Language, 33*, 1–18.

Hintzman, D. L., Curran, T., & Caulton, D. A. (1995). Scaling the episodic familiarities of pictures and words. *Psychological Science, 6*, 308–313.

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

Hirshman, E., Lanning, K., Master, S., & Henzler, A. (2002). Signal-detection models as tools for interpreting judgments of recollections. *Applied Cognitive Psychology, 16*, 151–156.

Hunt, R. R. (2003). Two contributions of distinctive processing to accurate memory. *Journal of Memory and Language, 48*, 811–825.

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

Jacoby, L. L. (1999). Ironic effects of repetition: Measuring age-related differences in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 3–22.

Jacoby, L. L., Kelley, C. M., & Dywan, J. (1989). Memory attributions. In H. L. Roediger & F. I. M. Craik (Eds.), *Varieties of memory and consciousness: Essays in honour of Endel Tulving* (pp. 391–422). Hillsdale, NJ: Erlbaum.

Jacoby, L. L., Kelley, C. M., & McElree, B. D. (1999). The role of cognitive control: Early selection versus late correction. In S. Chaiken & Y. Trope (Eds.), *Dual process theories in social psychology* (pp. 383–400). New York: Guilford.

Jacoby, L. L., Jones, T. C., & Dolan, P. O. (1998). Two effects of repetition: Support for a dual-process model of knowledge judgments and exclusion errors. *Psychonomic Bulletin & Review, 5*, 705–709.

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin, 114*, 3–28.

Johnson, M. K., & Raye, C. L. (1981). Reality monitoring. *Psychological Review, 88*, 67–85.

Johnson, M. K., Raye, C. L., Foley, H. J., & Foley, M. A. (1981). Cognitive operations and decision bias in reality monitoring. *American Journal of Psychology, 94*, 37–64.

Kensinger, E. A., & Corkin, S. (in press). The effects of emotional content and aging on false memories. *Cognitive, Affective, and Behavioral Neuroscience*.

Koriat, A., Goldsmith, M., & Pansky, A. (2000). Toward a psychology of memory accuracy. *Annual Review of Psychology, 51*, 481–537.

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

Lindsay, D. S., Hagen, L., Read, J. D., Wade, K. A., & Garry, M. (2004). True photographs and false memories. *Psychological Science, 15*, 149–154.

Lindsay, D. S., & Johnson, M. K. (1989). The eyewitness suggestibility effect and memory for source. *Memory & Cognition, 17*, 349–358.

Marsh, R. L., & Hicks, J. L. (1998). Test formats change source-monitoring decision processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*, 1137–1151.

McElree, B., Dolan, P. O., & Jacoby, L. L. (1999). Isolating the contributions of familiarity and source information to item recognition: A time course analysis. *Journal of Experimental Psychology: Learning, Memory and Cognition, 25*, 563–582.

Miller, M. B., & Wolford, G. L. (1999). Theoretical commentary: The role of criterion shift in false memory. *Psychological Review, 106*, 398–405.

Mintzer, M. Z., & Snodgrass, J. G. (1999). The picture superiority effect: Support of the distinctiveness model. *American Journal of Psychology, 112*, 113–146.

Mitchell, K. J., & Johnson, M. K. (2000). Source monitoring: Attributing mental experiences. In E. Tulving & F. I. M. Craik (Eds.), *The Oxford handbook of memory* (pp. 179–195).

Morrell, H. E. R., Gaitan, S., & Wixted, J. T. (2002). On the nature of the decision axis in signal-detection-based models of recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*, 1095–1110.

Multhaup, K. S. (1995). Aging, source, and decision criteria: When false fame errors do and do not occur. *Psychology and Aging, 10*, 492–497.

Neisser, U., & Harsch, N. (1992). Phantom flashbulbs: False recollections of hearing the news about Challenger. In E. Winograd & U. Neisser (Eds.), *Affect and accuracy in recall: Studies of "flashbulb" memories* (pp. 9–31). Cambridge: Cambridge University Press.

Nelson, D. L. (1979). Remembering pictures and words: Appearance, significance, and name. In L. S. Cermak & F. I. M. Craik (Eds.), *Levels of processing in human memory* (pp. 45–76). Hillsdale, NJ: Erlbaum.

Paivio, A. (1971). *Imagery and verbal processes*. New York: Holt, Rinehart, & Winston.

Pesta, B. J., Murphy, M. D., & Sanders, R. E. (2001). Are emotionally charged lures immune to false memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 328–338.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 803–814.

Roediger, H. L., & McDermott, K. B. (1999). False alarms about false memories. *Psychological Review, 106*, 406–410.

Roediger, H. L., Watson, J. M., McDermott, K. B., & Gallo, D. A. (2001). Factors that determine false recall: A multiple regression analysis. *Psychonomic Bulletin & Review, 8*, 385–407.

Rotello, C. M., Macmillan, N. A., & Van Tassel, G. (2000). Recall-to-reject in recognition: Evidence from ROC curves. *Journal of Memory and Language, 43*, 67–88.

Schacter, D. L., Cendan, D. L., Dodson, C. S., & Clifford, E. R. (2001). Retrieval conditions and false recognition: Testing the distinctiveness heuristic. *Psychonomic Bulletin & Review, 8*, 827–833.

Schacter, D. L., Israel, L., & Racine, C. (1999). Suppressing false recognition in younger and older adults: The distinctiveness heuristic. *Journal of Memory and Language, 40*, 1–24.

Schacter, D. L., Norman, K. A., & Koutstaal, W. (1998). The cognitive neuroscience of constructive memory. *Annual Review of Psychology, 49*, 289–318.

Schacter, D. L., & Wiseman, A. L. (in press). Reducing memory errors: The distinctiveness heuristic. In R. R. Hunt & J. Worthen (Eds.) *Distinctiveness and memory*. New York: Oxford University Press.

Strack, F., & Bless, H. (1994). Memory for nonoccurrences: Metacognitive and presuppositional strategies. *Journal of Memory and Language, 33*, 203–217.

Stretch, V., & Wixted, J. T. (1998). On the difference between strength-based and frequency-based mirror effects in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*, 1379–1396.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychologist, 26*, 1–12.

Underwood, B. J. (1969). Attributes of memory. *Psychological Review, 76*, 559–573.

Whittlesea, B. W. A., & Williams, L. D. (1998). Why do strangers feel familiar, but friends don't? A discrepancy-attribution hypothesis of feelings of familiarity. *Acta Psychologia, 98*, 141–165.

Wickens, T. D., & Hirshman, E. (2000). False memories and statistical decision theory: Comment on Miller and Wolford (1999) and Roediger and McDermott (1999). *Psychological Review, 107*, 377–383.

Wixted, J. T. (1992). Subjective memorability and the mirror effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*, 681–690.

Wixted, J. T., & Stretch, V. (2000). The case against a criterion shift account of false memory. *Psychological Review, 107*, 368–376.

Yonelinas, A. P. (1997). Recognition memory ROCs for item and associative information: The contribution of recollection and familiarity. *Memory & Cognition, 25*, 747–763.

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

Yonelinas, A. P., & Jacoby, L. L. (1996). Noncriterial recollection: Familiarity as automatic, irrelevant recollection. *Consciousness and Cognition, 5*, 131–141.