Available online at www.sciencedirect.com

ScienceDirect

Journal of Memory and Language 58 (2008) 1095–1111

Journal of Memory and Language

www.elsevier.com/locate/jml

Deep levels of processing elicit a distinctiveness heuristic: Evidence from the criterial recollection task ☆

David A. Gallo [a,b,*], Nathaniel G. Meadow [c], Elizabeth L. Johnson [a], Katherine T. Foster [a]

[a] *University of Chicago, Department of Psychology, USA*
[b] *University of Chicago, Center for Cognitive and Social Neuroscience, USA*
[c] *Washington University, Department of Psychology, USA*

Received 14 August 2007; revision received 30 November 2007
Available online 24 January 2008

Abstract

Thinking about the meaning of studied words (deep processing) enhances memory on typical recognition tests, relative to focusing on perceptual features (shallow processing). One explanation for this levels-of-processing effect is that deep processing leads to the encoding of more distinctive representations (i.e., more unique semantic or conceptual features that can be recollected to differentiate the words). This recollective distinctiveness hypothesis predicts that deep processing should reduce false recognition errors, because expecting more distinctive recollections can facilitate retrieval monitoring accuracy (i.e., a distinctiveness heuristic). We report several experiments confirming this prediction, while ruling out explanations based on familiarity or overall memory strength. Additional support for the distinctiveness hypothesis was that a manipulation designed to selectively enhance the distinctiveness of words in the shallow condition eliminated the levels-of-processing effect on false recognition. These findings suggest that conceptual processing can elicit the distinctiveness heuristic, and that recollective distinctiveness drives levels-of-processing effects.

Keywords: Levels of processing; False recognition; Distinctiveness; Metacognition

The levels-of-processing effect refers to the finding that memory for a list of words is better when the meaning or semantics of the words is encoded (deep processing), relative to focusing on more superficial aspects of the words (shallow processing) such as their perceptual, phonological, or orthographic characteristics (e.g., Craik & Tulving, 1975). For example, deciding whether each study word is "pleasant" leads to better recall on a subsequent test compared to deciding whether it contains the letter "e" (Hyde & Jenkins, 1969). A common explanation of this effect is that deep processing activates more relevant knowledge than shallow processing, and this activated information becomes associated with the word to form a more elaborate memory trace. The processing framework that this effect inspired has had a major impact on cognitive psychology (Craik & Lockhart, 1972). Despite this popularity, several results qualify the levels-of-processing effect,

☆ Ian McDonough provided useful suggestions on an earlier draft.

* Corresponding author. Address: Department of Psychology, University of Chicago, 5848 S. University Avenue, Chicago, IL 60637, USA. Fax: +1 773 702 0886.
E-mail address: dgallo@uchicago.edu (D.A. Gallo).

0749-596X/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jml.2007.12.001

and the theoretical implications have been called into question numerous times (see Roediger & Gallo, 2001).

Much of this debate has focused on the importance of retrieval factors, in addition to encoding factors. For instance, the levels-of-processing effect can be eliminated or reversed with memory tests that are sensitive to more shallow aspects of prior processing (e.g., Jacoby & Dallas, 1981; Morris, Bransford, & Franks, 1977; Stein, 1978). These and other findings led to the development of the transfer-appropriate-processing framework (e.g., Morris et al., 1977; Roediger, Gallo, & Geraci, 2002). This framework proposes that the degree of overlap between the cognitive procedures engaged during the study and test phase is the most critical factor in determining memory performance, as opposed to the depth of processing during encoding. By this view, semantic processing leads to better retention on typical recall and recognition tests because these tests happen to be mostly sensitive to the retrieval of semantic or conceptual information (e.g., Roediger, Weldon, & Challis, 1989). Because subjects are not given deep or shallow instructions on typical recall or recognition tests, their default tendency is probably to process test words at the semantic level, given that words are used to convey meaning in normal language usage.

The importance of retrieval factors also gains support from the levels-of-processing effect on false recognition. Several recognition studies have found an interaction between levels-of-processing at study and the type of nonstudied lures used at test (e.g., Davies & Cubbage, 1976; Elias & Perfetti, 1973; see also Coltheart, 1977; Parkin, 1983; Wallace, 1968; Wright et al., 1977). Although the methodology and findings were not always consistent across these studies, the general pattern was that semantic encoding of studied words led to greater false recognition of semantically related lures compared to surface encoding (often phonological or orthographic processing), whereas surface encoding led to greater false recognition of phonologically or orthographically related lures than did semantic encoding. These findings recently were replicated and extended by Chan, McDermott, Watson, and Gallo (2005) using the DRM task (e.g., Roediger & McDermott, 1995). These studies support the idea that the type of encoding process (shallow or deep) influences the types of features stored in memory (surface or semantic), and these features are most likely to influence test performance when they match the retrieval cues. Phonological lures are most likely to cue the retrieval of similar phonological features in memory, thereby enhancing familiarity-based false recognition, whereas semantic lures activate semantic features. These findings are consistent with the encoding specificity principle (e.g., Tulving, 1979), which like transfer-appropriate-processing, focuses on encoding and retrieval interactions.

Levels-of-processing and distinctiveness

Although the aforementioned studies highlight the importance of encoding/retrieval interactions, other findings indicate that semantic processing of words still has unique advantages over surface levels of processing. First, many of the encoding/retrieval studies showed that that the levels-of-processing effect on standard or semantically oriented tests was larger than the effects obtained on more surface-oriented tests (e.g., Fisher & Craik, 1977; Morris et al., 1977; Stein, 1978). Semantic encoding led to greater memory than surface encoding even when comparing test conditions that were thought to be transfer-appropriate. Second, this main effect of semantic processing also was evident in the false recognition data of Chan et al. (2005). False recognition of semantic lures following a semantic encoding task was greater than false recognition of phonological lures following a phonological encoding task. These false recognition findings potentially reflect the superior retention of semantically encoded information, analogous to the main effect of semantic processing on true memory. Finally, several studies have shown that some semantic judgments or encoding tasks could yield better retention than other semantic judgments or tasks, even though the same test was used in all conditions (e.g., Craik & Tulving, 1975; Seamon & Virostek, 1978). These findings again point towards the importance of various semantic encoding factors in determining performance.

In order to explain these sorts of effects, the idea of distinctiveness was incorporated into the levels-of-processing framework (e.g., Fisher & Craik, 1977; Jacoby & Craik, 1979; Lockhart & Craik, 1990; Moscovitch & Craik, 1976). Although the term “distinctiveness” has various meanings, the core idea embraced here is that semantic processing allows subjects to encode more unique features from each word relative to surface processing. These additional conceptual or semantic features help to differentiate the studied words from each other, making these memories less susceptible to interference and/or providing more features that can be cued on a typical recall or recognition memory test. As alluded to by Moscovitch and Craik (1976) and others, lists of words share a relatively limited number of surface features (e.g., phonemes), but convey a variety of meanings. Thus, meaningful processing differentiates words more than surface levels of processing. Other modifications of the levels-of-processing framework have appealed to the degree of richness, spread, or elaboration within a given level of processing, but to the degree that these processes lead to the encoding of additional unique features, they ultimately have their effects via distinctiveness. The idea of distinctiveness can explain why deep processing leads to superior retrieval even under transfer-appropriate conditions, and also why some semantic judgments and tasks afford better memory than others (i.e., they differ in the number of semantic features encoded).

The distinctiveness explanation of the levels-of-processing effect makes a straightforward prediction for false recognition, or the likelihood of falsely deciding that a nonstudied word was studied. If deep levels of processing make memories more distinctive, then false recognition should be reduced when subjects are tested for deeply processed words, relative to shallowly processed words. This prediction stems from work on the distinctiveness heuristic, which has predominantly used comparisons between pictures and words (e.g., Dodson & Schacter, 2002; Schacter, Israel, & Racine, 1999). It has been shown across several tasks that subjects are less susceptible to false recognition if they had originally studied pictures, relative to words (for review see Schacter & Wiseman, 2006). According to the distinctiveness heuristic hypothesis, subjects expect more distinctive recollections for pictures than words, and the failure of nonstudied items to elicit these recollections helps subjects to decide that the item is new. The distinctiveness heuristic can be considered a special type of source monitoring process, analogous to memorability heuristics demonstrated in source memory tasks (e.g., Johnson, Hashtroudi, & Lindsay, 1993; Marsh & Hicks, 1998). In the context of picture effects, distinctiveness has been defined as the complexity and uniqueness of the perceptual features in the stimulus (e.g., Gallo, Weiss, & Schacter, 2004; for related ideas see Nelson, 1979). These distinctive features potentially help subjects to differentiate picture memories, and also provide more features that can subsequently be recollected. In a similar fashion, distinctiveness for deep levels of processing has been described as the unique number of semantic or conceptual features that are processed. If such conceptual distinctiveness influences recollective expectations, then a distinctiveness heuristic should be elicited by deep levels of processing, relative to shallow levels.

The hypothesis that deep levels of processing enhance the distinctiveness of studied words is not supported from the current false recognition literature.[1] As described above, semantic encoding has been found to elevate false recognition of semantic lures relative to surface processing in the DRM task (e.g., Chan et al., 2005; Thapar & McDermott, 2001). These findings are opposite to those predicted by the distinctiveness heuristic hypothesis. However, the DRM task is not ideal to investigate this hypothesis, because deep levels of processing can enhance semantic or associative processing, elevating false recognition of semantically related lures independent of any effects on retrieval monitoring accuracy (see Gallo, 2006). Further, both of these DRM studies manipulated levels-of-processing using a within-subjects design. According to the distinctiveness heuristic hypothesis, subjects should be most effective at using recollective expectations to reduce false recognition when they expect distinctive recollections from all of the test items, a situation unlikely to apply to these types of designs (e.g., Dodson & Schacter, 2001; Schacter et al., 1999). Consistent with this interpretation, several of the earlier studies that manipulated levels of processing between subjects found greater overall levels of false recognition in shallow conditions compared to deep conditions (e.g., Davies & Cubbage, 1976; Elias & Perfetti, 1973; Parkin, 1983; see Smith & Hunt, 1998, for related results in false recall). Similarly, a within-subjects study by Jacoby, Shimizu, Daniels, and Rhodes (2005) demonstrated that having subjects selectively search their memory for deeply processed targets in one test block led to reduced false alarms relative to the block where they searched memory for shallowly processed targets. All of these latter findings are consistent with the distinctiveness heuristic hypothesis, but as described next, they also are consistent with an alternative explanation.

The limitation with all of these prior levels-of-processing studies is that different levels of false recognition, on their own, are not definitive evidence for a recollection-based distinctiveness heuristic. According to dual process models, recognition tests can be based on both recollection and familiarity, and deep processing enhances both recollection and familiarity relative to shallow processing (e.g., Toth, 1996; for review see Yonelinas, 2002). As a result, subjects who deeply encoded words could reduce false recognition by using a more conservative familiarity-based criterion, as opposed (or in addition) to expecting more distinctive recollections. As discussed by Gallo et al. (2004), criterion shifts along a familiarity dimension often have been used to explain false recognition differences across many conditions where stimuli vary in memorability, as in studies of the mirror effect (for a signal detection perspecitive, see Hirshman, 1995; Stretch & Wixted, 1998; Verde & Rotello, 2007; for related ideas in the source memory literature, see Bink, Marsh, & Hicks, 1999; Johnson et al., 1993). Similarly, familiarity-based criterion shifts could explain why deep levels-of-processing reduced false recognition relative to shallow processing, at least in those studies that reported this effect. Given that such familiarity-based processes could cause

[1] There is evidence that generating words at study (e.g., solving anagrams for studied words) can suppress false recognition or source errors across several tasks, relative to simply reading them (e.g., Gunter, Bodner, & Azad, 2007; Johnson, Raye, Foley, & Foley, 1981; Marsh & Hicks, 1998). These effects can be explained via the distinctiveness heuristic hypothesis. However, the benefits of generative processing on memory typically are attributed to additional cognitive operations required to generate items (e.g., Marsh & Hicks, 1998; also see Bertsch, Pesta, Wiscott, & McDaniel, 2007), and so do not speak directly to levels-of-processing effects.

differences in false recognition across conditions, it is theoretically critical to control for familiarity effects when attempting to isolate a recollection-based distinctiveness heuristic.

In order to more directly test the distinctiveness heuristic hypothesis, one needs a task that focuses subjects on the different types of recollection at retrieval (as in Jacoby et al., 2005; or Marsh & Hicks, 1998), and one also needs to separate recollection processes from the potential influences of familiarity (and familiarity-based criterion shifts). We are not aware of any levels-of-processing studies that have satisfied both of these criteria, nor are we aware of any source memory tests that have evaluated whether subjects make fewer source confusions for a deep relative to a shallow source.

Criterial recollection task

The purpose of the present study was to provide a more definitive test of the distinctiveness hypothesis for levels of processing effects, using the criterial recollection task. This is a special type of source memory task, originally designed to test the hypothesis that studying pictures elicits a recollection-based distinctiveness heuristic (Gallo et al., 2004). Subjects in those experiments studied a list of words in black font that were paired with a perceptually distinctive stimulus (pictures) or with a less perceptually distinctive stimulus (the same word in red font), manipulated within-subjects. At test, subjects received words in black font as retrieval cues, with different retrieval instructions across test blocks. On the standard test, subjects were instructed to respond "yes" to studied items and "no" to nonstudied items, providing an overall estimate of memory strength (i.e., the combined influence of recollection and familiarity). On the picture test, they were to responded "yes" if they recollected that the test word had been studied with a picture, whereas on the red word test they were to respond "yes" if the test word had been studied in red font. Importantly, some items were studied in both formats so that red words and pictures were not mutually exclusive. In this way, subjects could not use a recall-to-reject strategy to avoid memory confusions, as is the case in exclusion tasks (e.g., using the recall of a picture to reject an item on the red word test, see Gallo, Cotel, Moore, & Schacter, 2007 for relevant evidence). Instead, subjects had to selectively search their memory for the to-be-recollected format on these criterial recollection tests (e.g., red font on the red word test, and pictures on the picture test).

The criterial recollection task provides the most direct test of the distinctiveness heuristic hypothesis, because subjects are explicitly required to search memory for one type of recollection on one test and another type on the other test. By directly comparing performance across these tests, one can determine how the different recollective expectations influenced false recognition errors. Using this logic, Gallo et al. (2004) found that subjects made fewer false recognition errors on the picture test than on the red word test. These findings suggest that subjects expected more distinctive recollections when searching memory for pictures, and these expectations facilitated retrieval monitoring accuracy (i.e., the distinctiveness heuristic). Importantly, these false recognition effects were dissociated from the effects of study format on true recognition, so that false recognition was lower on the picture test regardless of whether picture hits were greater than red word hits (the typical picture-superiority effect) or whether red word hits were made greater than picture hits (by repeating the red words at study and enhancing their familiarity). This dissociation between true and false recognition effects cannot be easily explained by criterion-shifts based on overall memory strength or familiarity of the targets. Instead, it is more consistent with the idea that pictures elicited more distinctive recollections than red words in all conditions, and subjects consistently used these expectations to suppress false recognition.[2]

The overall design of the current experiments was similar to that used in previous studies with this task, with the exception that the two presentation formats under scrutiny were words encoded with a deep task (deciding whether the word is pleasant) or with a shallow task (deciding whether the word contains an "e"). On the criterial recollection tests, subjects had to decide whether items had originally been given a pleasantness judgment (deep test) or an e-check judgment (shallow test). If deep processing leads to more distinctive recollections than shallow processing, then subjects should take advantage of this difference and false recognition should be reduced when subjects search their memory for pleasantness judgments compared to e-check judgments. That is, subjects should use a distinctiveness heuristic to reduce false recognition. The first three experiments tested this hypothesis, as well as the extent that these results could be explained by the competing

[2] For simplicity, we consider all types of false alarms in this task as "false recognition" errors, even though the cause of these errors is different (e.g., false alarms to nonstudied lures are driven by idiosyncratic processes, whereas false alarms to studied lures are additionally driven by familiarity from prior presentation, or source confusions). This simplified use of the term is theoretically justified here because, from a source monitoring perspective, the distinctiveness heuristic should apply to any lure that fails to elicit a distinctive recollection, irrespective of the source of the familiarity. Thus, evidence for the distinctiveness heuristic has been found for related and unrelated lures in the DRM task, studied and nonstudied lures in the criterial recollection task, and nonstudied lures in more standard source tests (see Gallo et al., 2004).

strength or familiarity-based criterion shift hypothesis. The final two experiments provided an additional test of the distinctiveness hypothesis, using a manipulation designed to enhance the distinctiveness of items in the shallow processing condition.

Experiment 1

Subjects

Twelve University of Chicago undergraduates participated for course credit or $10.

Materials and design

Two hundred and forty words were drawn from the online MRC database (Coltheart, 1981; Wilson, 1988), ranging from 5 to 7 letters in length, written frequency >1, and relatively high concreteness and imagability (5–7 on a 1–7 scale). These words were divided into 12 sets of 20, to be counterbalanced across the four study conditions (e-check judgment, pleasantness judgment, both judgments, nonstudied) and the three tests (standard test, shallow test, deep test) across subjects. Half of the words in each set contained a single letter "e," the other half did not contain this letter.

Procedure

Subjects were instructed that they would study two lists of words for a subsequent memory test. For one list they decided whether or not the words contained the letter "e" (shallow judgment), for the other list they decided whether the words were pleasant (deep judgment). Half of the subjects received the shallow list first; the other half received the deep list first. Words were presented visually in the center of the computer screen, and to further ensure that subjects made the appropriate response, a visual prompt indicating the upcoming judgment ("have e?" or "pleasant?") preceded each word by 500 ms. Words and prompts remained on the screen until subjects made a yes/no response on the keyboard, and the next prompt was presented 500 ms after each response. Out of the 180 study words, 120 were only presented once (60 in the shallow list, 60 in the deep list), and 60 words were presented twice: once in the shallow list, and once in the deep list (i.e., items for which subjects made both judgments). Subjects were told that some items would appear in both lists, and within each list, words were freshly randomized for each subject.

Following the study phase subjects were given test instructions. They were told that they would be presented with test words on the computer screen, and that some of these words were studied (i.e., they had made an e-check judgment, a pleasantness judgment, or both judgments), and others were not studied. There were three test blocks, each containing 20 items from each of the four study conditions (e-check judgment, pleasantness judgment, both judgments, or nonstudied). The items within each test were freshly randomized for each subject. On the standard test, they were told to press the "yes" key for all words that were presented at study, regardless of whether they had made an e-check or a pleasantness judgment. On the shallow test, they were told to press "yes" if they recollected making an e-check judgment for the word at study, and "no" if they did not recollect making this judgment (regardless of whether they had made a pleasantness judgment). On the deep test, they were told to press "yes" if they recollected making a pleasantness judgment during the study phase, and "no" if they did not (regardless of whether they had made an e-check judgment). It was reiterated that they had made both of the study judgments for some of the test items, and thus, if they recollected making one judgment this would not help them to decide whether or not they also had made the other judgment. Instead, they had to selectively search their memory for e-check judgments on the shallow test, and for pleasantness judgments on the deep test. The order of the three test blocks was counterbalanced across subjects. For simplicity, we referred to the deep test as the "pleasantness test" and the shallow test as the "e-judgment test" in our instructions to subjects, although we use the terms "deep" and "shallow" here given their theoretical implications.

Results and discussion

Results from Experiment 1 are in the first column of Table 1. Unless specified otherwise, all differences reported in this paper were significant at the conventional level ($p < .05$, two-tailed). Results from the standard test demonstrated the typical levels-of-processing effect on true memory. Hits to words studied only in the shallow list (shallow items, mean = .46) were lower than hits to words studied only in the deep list (deep items = .94, $t[11] = 5.65$, $SEM = .084$, $d = 2.23$). This effect also was obtained when comparing hits to these items across the criterial recollection tests (.47 and .85, $t[11] = 5.90$, $SEM = .065$, $d = 2.18$), and also when comparing hits to items that were studied in both encoding lists across the criterial recollection tests (.55 and .88, $t[11] = 6.53$, $SEM = .050$, $d = 2.37$). Note that both encoding judgments (deep and shallow) had been made for the latter items (i.e., they were the same types of items). Thus, the levels-of-processing effect for these "both items" reflects a retrieval orientation effect, with subjects searching memory for the e-check judgment on the shallow test, and the pleasantness judgment on the deep test.

In order to test the distinctiveness heuristic hypothesis we compared false recognition across the criterial rec-

Table 1
Mean recognition of each item type as a function of test type in Experiments 1–3a

	Experiment 1 P1, E1 Standard cuing	Experiment 2 P1, E3 Standard cuing	Experiment 3a P1, E3 Reverse cuing
Standard test			
Both hits	.94 (.02)	.93 (.03)	.95 (.01)
Shallow hits	.46 (.08)	.63 (.05)	.90 (.02)
Deep hits	.94 (.02)	.88 (.04)	.81 (.02)
New FAs	.21 (.05)	.15 (.03)	.12 (.02)
Shallow test			
Both hits	.55 (.05)	.73 (.03)	.85 (.03)
Shallow hits	.47 (.06)	.60 (.06)	.80 (.03)
Deep FAs	.43 (.05)	.47 (.06)	.46 (.03)
New FAs	.23 (.05)	.32 (.04)	.11 (.02)
Deep test			
Both hits	.88 (.03)	.87 (.04)	.77 (.03)
Deep hits	.85 (.03)	.85 (.03)	.59 (.04)
Shallow FAs	.25 (.04)	.30 (.05)	.38 (.05)
New FAs	.13 (.03)	.13 (.04)	.05 (.01)

Notes. Standard errors of each mean are in parenthesis. P1, pleasantness judgments were made once per study word; E1, e-check judgments were made once per study word; E3, e-check judgments were made three times per study word; FAs, false alarms.

ollection tests. A 2 (lure type: studied lure and nonstudied lure) × 2 (test type: shallow test and deep test) ANOVA revealed an effect of lure, $F(1,11) = 15.43$, $MSE = .019$, $\eta_p^2 = .584$, an effect of test, $F(1,11) = 12.82$, $MSE = .018$, $\eta_p^2 = .538$, and no interaction. As expected, false recognition of studied lures was greater than nonstudied lures. This effect indicates that subjects were not perfect on this task, but made significant number of source confusions (or familiarity-based errors). More importantly, subjects were less likely to make false recognition errors when searching memory for pleasantness judgments, compared to e-check judgments, suggesting that the former elicited more distinctive recollections and so facilitated the use of the distinctiveness heuristic.

We also analyzed the item effects on each test (inequality signs represent statistically significant differences, $p < .05$). On the standard test, both hits (.94) = deep hits (.94) > shallow hits (.46) > new false alarms (.21). In other experiments with this task, hits to items studied in both conditions are sometimes greater than hits to other items, because both items are presented twice, but this effect is not always obtained (e.g., Gallo et al., 2004). Either way these both items were included only to prevent a recall-to-reject strategy and so are not of central concern. On the shallow test, both hits (.55) = shallow hits (.47) = deep false alarms (.43) > new false alarms (.23). The failure to discriminate between shallow and deep items on this test demonstrates that it was very difficult to recollect e-check judgments. Nevertheless, the fact that the very large levels-of-processing effect observed on the standard test (+48%) was not observed for these same types of items on the shallow test (−4%) indicates that subjects were not simply relying on overall memory strength on this test, but instead were attempting to recollect the e-check judgments (as instructed). Finally, on the deep test, both hits (.88) = deep hits (.85) > shallow false alarms (.25) > new false alarms (.13). In contrast to the shallow test, subjects were able to discriminate shallow and deep items when searching memory for pleasantness judgments, even though these were the same types of items in each case. This difference again supports the assumption that subjects based their decisions on different types of recollections across these tests, and pleasantness judgments were easier to recollect.

Experiment 2

The results of Experiment 1 were consistent with the distinctiveness heuristic hypothesis, but an alternative explanation also is possible. Although subjects were instructed to use recollection on the criterial recollection tests, they were not perfectly able to do so, as they also made significant source confusions (relative to false recognition of nonstudied lures). These findings suggest that performance was influenced by familiarity from prior presentation, and so the obtained reduction of false recognition on the deep test might have been caused by familiarity or strength-based criterion shifts as opposed to a recollection-based distinctiveness heuristic. Because deep items were stronger in memory than shallow items (e.g., greater hits on the standard test), subjects may have used a more conservative strength based response criterion on the deep test, thereby reducing false recognition relative to the shallow test.[3] Experiment 2 was designed to test between these two alternative explanations by strengthening memory of the shallow items.

Method

Twelve University of Chicago undergraduates participated for course credit or $10. The method and procedures were identical to Experiment 1, with the

[3] If the deep items were more familiar than shallow items, then this difference alone could explain the false recognition pattern for studied lures. However, this explanation does not apply to the effects observed on nonstudied lures, which should have been equally familiar across conditions. The criterion shift hypothesis can explain false recognition effects for both types of lures, and thus is the major competing hypothesis.

exception that every item that was presented in the shallow list was repeated three times (nonconsecutively). It was assumed that repetition would increase the strength of the shallow items (e.g., enhance familiarity, and potentially the frequency that these items were recollected). However, repetition should not change the qualitative nature of the encoding decisions, and thus would not enhance the unique number of distinctive features stored in memory for each shallow item. Pleasantness judgments would continue to yield more distinctive recollection than e-check judgments. If false recognition differences across tests were based on strength-based criterion shifts, then elevating the memory strength of words in the shallow list should reduce or reverse these false recognition effects. In contrast, if recollective distinctiveness was the critical factor, then the previously obtained false recognition effects should be replicated.

Results and discussion

Results from Experiment 2 are in the second column of Table 1. As in Experiment 1, a levels-of-processing effect was found on true recognition on the standard test (shallow hits = .63, deep hits = .88, $t[11] = 4.15$, $SEM = .060$, $d = 1.69$), and also across the criterial recollection tests for the items encoded with one judgment (.60 and .85, $t[11] = 4.37$, $SEM = .056$, $d = 1.48$) or both judgments (.73 and .87, $t[11] = 3.44$, $SEM = .040$, $d = 1.05$). Also replicating Experiment 1, false recognition was greater on the shallow test than on the deep test. A 2 (lure type) × 2 (test type) ANOVA revealed an effect of lure, $F(1,11) = 34.78$, $MSE = .008$, $\eta_p^2 = .760$, an effect of test, $F(1,11) = 11.44$, $MSE = .033$, $\eta_p^2 = .510$, and no interaction. The effect of test is consistent with the use of a distinctiveness heuristic on the deep test, even though it is clear from Table 1 that the size of the levels-of-processing effect on true recognition was smaller in this experiment, owing to the repetition of the shallow items.

Within-test comparisons for Experiment 2 revealed that, on the standard test, both hits (.93) = deep hits (.88) > shallow hits (.63) > new false alarms (.15). On the shallow test, both hits (.73) > shallow hits (.60) > deep false alarms (.47) > new false alarms (.32). Subjects were better at discriminating shallow items from deep items on this test, compared to Experiment 1, because shallow items were repeated (although this effect was only marginally significant, $p = .05$, two-tailed). On the deep test, both hits (.87) = deep hits (.85) > shallow false alarms (.30) > new false alarms (.13). With the exception of the predicted elevation in memory for shallow judgments, the overall patterns in these results were quite similar to those of Experiment 1.

Experiment 3a

Experiment 2 replicated the false recognition results of Experiment 1, even though the hit rate to shallow items was increased by study repetition. These effects are consistent with a recollection based distinctiveness heuristic, but because deep judgments still led to better memory for studied words than shallow judgments, a strength or familiarity-based criterion shift still might have contributed to these effects. Experiment 3a was designed to provide a more definitive test between these two accounts, by reversing the levels-of-processing effect on true recognition. If true recognition were greater following the shallow judgment, relative to the deep judgment, then a strength or familiarity-based criterion shift would predict that false recognition would be lower on the shallow test than on the deep test. Reversing the strength or familiarity of these classes of items also should reverse the direction of a criterion shift that is based on these types of information. In contrast, deep judgments should still elicit more distinctive recollections than shallow judgments, so that the distinctiveness heuristic hypothesis predicts lower false recognition on the deep test than the shallow test.

The key to this experiment was finding a manipulation that would reverse the typical levels-of-processing effect on true recognition (e.g., by elevating the familiarity of the shallow words), but not change the qualitative nature of the encoding tasks, thereby leaving potential differences in recollective distinctiveness intact (i.e., deep > shallow). Study repetition by itself did not achieve this aim, so we adopted the reverse-cuing procedure used by Craik (1977). In that study the encoding tasks were randomly mixed within the same study list, and subjects were presented with each study word before they were given a prompt to make either a deep or a shallow encoding judgment on that word. Because subjects could not know which judgment would be made for each word, they had to read each word and hold it in mind until the judgment prompt. Craik (1977) found that the levels-of-processing effect was smaller with this procedure compared to the typical procedure, and this reduction was caused by a disproportionate increase in memory for the shallowly processed words. These findings suggest that the reverse-cuing procedure ensured a minimal amount of semantic processing for all of the words, which was especially beneficial for words in the shallow condition (although not sufficient to overcome the benefit of explicitly making deep judgments).

In the current experiment, we assumed that repeating the shallow words three times with this reversed-cuing procedure might enhance their familiarity more than deep words, potentially achieving the desired reversal of the levels-of-processing effect on true recognition. However, this reverse-cuing procedure could not change the qualitatively different nature of the shallow or deep encoding judgments made after each item's presentation. Pleasantness judgments require a degree of effortful, item-specific semantic processing that is not likely to be engaged by the reverse-cuing procedure alone, and so they still should afford more distinctive recollections than e-check judgments. As a result, we predicted that subjects again would suppress false recognition on the deep test, relative to the shallow test, via a distinctiveness heuristic. These predictions were tested using the criterial recollection task in the current experiment, and our assumptions about differences in recollection and familiarity across the encoding conditions were tested (and confirmed) in Experiment 3b.

Method

Thirty-six University of Chicago undergraduates participated for course credit or $10.[4] We modified the design of Experiment 2 so that the two encoding tasks were randomly mixed in a single study list, and we reversed the ordering of the study word and the levels-of-processing prompt. Each study word first appeared on the computer screen for 500 ms, was removed for 1000 ms, and then a prompt indicated which encoding judgment to make (i.e., "have e?" or "pleasant?") until subjects made their response. All other methods and procedures were identical to Experiment 2.

Results and discussion

From the third column of Table 1 it is clear that we were successful at reversing the typical levels-of-processing effect on true recognition in this experiment. On the standard test, shallow hits (.90) were greater than deep hits (.81, $t[35] = 3.87$, $SEM = .025$, $d = .72$), and this reversal was replicated across these same items on the criterial recollection tests (.80 and .59, $t[35] = 5.63$, $SEM = .038$, $d = .96$) and also across items that were studied with both encoding tasks (.85 and .77, $t[35] = 3.44$, $SEM = .026$, $d = .52$). Despite this reversal in overall memory strength, false recognition again was greater on the shallow test than on the deep test. A 2 (lure type) × 2 (test type) ANOVA revealed an effect of lure, $F(1,35) = 166.45$, $MSE = .025$, $\eta_p^2 = .826$, an effect of test, $F(1,35) = 5.98$, $MSE = .028$, $\eta_p^2 = .146$, and no interaction. Given that the shallow targets were stronger than the deep targets, it is remarkable that false recognition of shallow items on the deep test (mean = .38) still was lower than false recognition of deep items on the shallow test (.46). This finding, coupled with a similar finding for nonstudied lures, again suggests that subjects used a distinctiveness heuristic on the deep test.

Within-test comparisons for Experiment 3a revealed that, on the standard test, both hits (.95) > shallow hits (.90) > deep hits (.81) > new false alarms (.12). On the shallow test, both hits (.85) = shallow hits (.80) > deep false alarms (.46) > new false alarms (.11). On the deep test, both hits (.77) > deep hits (.59) > shallow false alarms (.38) > new false alarms (.05). With the exception of the planned reversal of the levels-of-processing effect on true recognition, the overall pattern of effects was similar to that obtained in Experiment 2.

Experiment 3b

The reverse-cuing procedure of Experiment 3a was successful in reversing the typical levels-of-processing effect on true recognition. We assumed that this reversal was due to enhanced familiarity of shallow items, relative to deep items, but that pleasantness items would still elicit more distinctive recollections than shallow items (thereby eliciting a distinctiveness heuristic). Experiment 3b was designed as a manipulation check to verify these assumptions. A new group of subjects underwent the same study procedures as in Experiment 3a, and were tested on the same subsets of items across the three test blocks. However, they received different instructions and made different judgments for these items across the three memory test blocks. These new test blocks were designed to measure the relative levels of recollection and familiarity of the test items.

The first test was a speeded yes/no recognition test. Subjects had to decide whether each test item was studied, irrespective of the judgment that had been made during study, but responses were to be made 700 ms after the onset of the test item. Prior work indicates that responding at such a fast speed relies more on familiarity than on recollection, and so this test provided a better estimate of familiarity than a self-paced yes/no recognition test (cf. Yonelinas, 2002). The second test was a self-paced recognition test, using an adaptation of the "remember"/"know" procedure (Tulving, 1985) to estimate the levels of recollection and familiarity for the test item (i.e., the recollect/familiar test). The third test also was a self-paced test, but this time we did not have subjects make a recognition judgment. Instead, they rated each item on a graded (0–7) scale according to the quality of their recollection for that item (cf. Higham & Vokey, 2004), making one judgment for overall "recollection strength" and a second judgment for the amount of "recollected detail" (i.e., a recollection quality test). Of course, it is difficult to gauge the validity of subjective

[4] Our effects tended to be more variable in this experiment, potentially because the reverse-cuing technique increased variability in the way subjects studied the words. As a result, we tested more subjects to increase reliability and power. Data from one subject was replaced because they did not follow instructions (i.e., average response latency on the final test (e-check) was 262 ms, compared to the group mean of 1527 ms).

judgments, and it was unclear to us whether subjects would comprehend the distinction between recollection "strength" and "detail." However, in the event that they did understand this distinction, we felt it important to separate the two judgments. Our reasoning was that the second question should be more important for the distinctiveness heuristic hypothesis, because it pertained to the amount of unique and distinctive details contained in the recollection. Thus, to minimize the possibility that subjects would confuse this judgment with an overall memory "strength" judgment, we first asked them to make a strength judgment. (As it turns out, the two judgments led to the same conclusion.)

Method

Twenty-four University of Chicago undergraduates participated for course credit or $10. Subjects were presented with the same materials and study procedures as in Experiment 3a, but they received different instructions across the three memory test blocks.[5] The first test block was a speeded recognition test. Subjects made yes/no recognition memory decisions as on the standard recognition test of the other experiments, but they were required to respond within a given time-window using a tempo procedure (Balota, Burgess, Cortese, & Adams, 2002). Subjects initiated each trial via the spacebar, and then were prompted with a series of visual arrows that appeared every 700 ms. On the third beat the test item appeared on the computer screen, and subjects were to make their recognition response on the fourth beat. If their response was between 600 and 750 ms, then a "good" signal appeared and they were allowed to press the spacebar to initiate the next item. Otherwise, a "too fast" or "too slow" prompt appeared on the screen for 3 s, accompanied by an error sound over the headphones. Subjects were instructed to be as accurate as possible and to respond on time.

The second test block was self-paced, and subjects were again told that they would see studied and nonstudied items. For each item they made an "actual recollect," "very familiar," or "new" decision. These instructions were adapted from typical "remember"/ "know" instructions, with the important exception that "actually recollect" judgments were to be based on whether they recollected the specific judgment (e-check or pleasantness) made at study, because this is the aspect of recollection that is most critical for the criterial recollection tests. Specific instructions read to subjects were as follows:

"Respond "Actually Recollect" ("AR") if you can recall, or bring to mind, a specific memory of the test word's presentation (i.e., you actually remember something specific about making an e-check judgment or a pleasantness judgment for the word). Respond "Very Familiar" ("VF") if the test word is familiar and you think that it was presented, but you cannot actually recollect any details about the item's presentation (i.e., you don't remember the specific judgment you made at study). You "just know" it was studied. Finally, respond "New" if you do not think the test word was studied at all. Note that the difference between Actually Recollect and Very Familiar is not confidence; you might be very confident that an item was studied because it is very familiar, but you cannot recollect any details about the presentation of the item. Alternatively, you might be very confident that the item was studied because you recollect very vivid details of the presentation experience. In order to make sure that you understand this task, please repeat back to me what you will do during this test".

The third test block also was self-paced, and was designed to provide a more fine-grained measure of the subjective experience of recollection. This test was administered after the "actual recollect"/"very familiar" test, so that subjects would already understand the distinction between recollection and familiarity. For each item on this final test, subjects made two consecutive ratings on a 0–7 scale. The first question was "How strong is your actual recollection of this item?" and the second was "How detailed is your actual recollection of this item?" Exact instructions for these ratings were as follows:

"This time, we are not interested in vague or general feelings of familiarity at all. Instead, your job is to make more specific ratings for your actual recollections from the study phase, using two different judgments for each test item. The first judgment is on recollection strength, and the second is on recollection details.

On the first judgment we would like you to rate how strong, or vivid, your actual recollection is for the item presented. For example, you may have a very strong or vivid memory of making an "e" judgment because you recollect a vivid memory of the "e" prompt appearing above the item. Similarly, you may have a very strong or vivid memory of making a pleasantness judgment because the item reminded you of a pleasant experience. Alternatively, you may have a weak or vague recollection of the item being associated with an "e" judgment or a pleasantness judgment, or you might simply think the item is familiar but not recollect much about the experience. Please rate your recollection strength on a 0–7 scale where a

[5] Data from one subject was replaced because she removed the headphones during the speeded test to avoid the negative feedback. This subject had the slowest response time on the speeded test (mean = 950 ms). Also, recollection ratings from the first subject were not analyzed because the wrong prompts were presented on screen.

rating of 0 would indicate that you do not recollect the item at all, a rating of 1 would indicate that your recollection for the item is weak, or vague, and a rating of 7 would indicate that your actual recollection is very strong, or vivid.

On the second judgment, which is on recollection detail or distinctiveness, we would like you to rate the number of unique details that you can recollect that are different from other items in the study phase, independent of how strong or vivid those recollections may be. Please rate the number of unique details on a 0–7 scale where a rating of 0 would indicate that you do not recollect any details at all, a rating of 1 would indicate that you recollect only a few unique details, and a rating of 7 would indicate that you recollect many unique details for the item. For example, when presented with an item, you may recollect some details that are unique to the item presented. You may remember several different reasons or associations that you made to determine that this item was pleasant, or you may remember looking at several unique aspects of the word to make an "e" judgment; this would be given a higher rating because the details are unique or distinctive to the item. Alternatively, you may not recollect many details, or you may recollect several details that are not unique to the item presented. Or, to give another example, you may have a general recollection of the pleasantness or e-judgment, but because you have these same sorts of general memories for many of the other items, these recollections would be given a lower rating because the details are not very unique to the specific item presented. Remember that these judgments are about your memories for distinctive details that you encountered during the study phase, not whether you can think of new unique details during the test. In order to make sure that you understand this task, please repeat back to me what you will do during this test".

Results and discussion

On average, subjects on the speeded test were very good at responding within the 600–750 ms response window (mean response time across all test items = 699 ms, range = 642–758 ms). Results from the speeded recognition test are summarized in the left column of Table 2. As expected, relative to the nonspeeded standard recognition test of Experiment 3a, speeding subjects reduced hits to studied items and increased false alarms to nonstudied items (all p's < .01). More importantly, the reverse levels-of-processing effect was replicated in this experiment, as shallow hits (.76) were greater than deep hits (.68, t[23] = 2.42, SEM = .034, d = .41). Note that, in order to avoid item-selection artifacts, this analysis included all recognition responses irrespective of response time. We also conducted a very strict cut-off analysis, selectively analyzing only those responses that were 700 ms or faster (eliminating 40% of the responses, on average). On this reduced dataset, the reverse levels-of-processing effect still was significant (shallow hits = .48, deep hits = .40, t[23] = 3.13, SEM = .028, d = .52). Assuming that speeded responding relied predominantly on familiarity, these results confirm that the reverse-cuing procedure had made shallow items more familiar than deep items.

Subjective judgments made on the recollect/familiar test are summarized in the middle columns of Table 2. Subjects were most likely to claim to "actually recollect" deep items (.71) than shallow items (.54, t(23) = 3.80, SEM = .043, d = .75). This difference is consistent with the idea that recollections for deep items were more distinctive than those for shallow items, although it also is possible that this subjective judgment reflected the quantitative "strength" of the recollection as opposed (or in addition) to qualitative differences in recollection (cf. Gallo et al., 2004), a point that we address below with the recollection quality ratings. In contrast to the recollection judgments, subjects were more likely to claim that shallow items were "very familiar" (.36) compared to deep items (.19, t(23) = 4.04, SEM = .043, d = .89). This pattern suggests that shallow items were more familiar than deep items, but of course, absolute comparisons of "very familiar" judgments can be misleading because these judgments were only made if subjects did not make an "actually recollect" judgment. In fact, the "very familiar" judgments to "both" items (.16) were

Table 2
Speeded recognition and subjective responses on the three test blocks of Experiment 3b

Item type	Speeded test	Recollect/familiar test			Recollection quality test	
	p "yes"	p "AR"	p "VF"	IRK	Strength	Details
Both	.83 (.03)	.81 (.04)	.16 (.03)	.84	6.08 (.17)	4.20 (.25)
Shallow	.76 (.03)	.54 (.05)	.36 (.05)	.78	4.89 (.19)	2.82 (.22)
Deep	.68 (.04)	.71 (.04)	.19 (.03)	.66	5.28 (.20)	3.63 (.29)
New	.26 (.03)	.04 (.02)	.16 (.02)	.17	1.01 (.26)	0.54 (.17)

Notes. Standard errors of each mean are in parenthesis. Ratings for the recollection quality judgments were on a 0–7 scale. AR, actually recollect; VF, very familiar; IRK, familiarity estimate from independent-recollection-familiarity adjustment.

significantly lower than those to shallow items (.36), even though the "both" items should have been at least as familiar as the other items. We therefore used Yonelinas' adjustment ($F = p$"VF"/[$1 - p$"AR"]) to estimate familiarity, which assumes that recollection and familiarity are independent (i.e., the independent-remember-know [IRK] procedure, see Yonelinas, 2002). Familiarity estimates from this estimate are listed in Table 2 (in the IRK column), and they were in the same direction (and quite similar) to the probability of responding "yes" on the speeded recognition test. The estimate of familiarity for shallow items (.78) was greater than that for the deep items (.66), and this effect was significant using a directional t-test ($t[22] = 1.80$, $SEM = .063$, one-tailed $p = .04$, $d = .49$).[6] Thus, whether one compared the raw levels of "very familiar" judgments or those that were adjusted with the independence assumption, our subjective results suggest that shallow items were more familiar than deep items.

Ratings from the recollection quality test are summarized in the final columns of Table 2. As expected, "both" items elicited the highest ratings and nonstudied items elicited the lowest ratings. The most important finding was that deep items were rated more highly on either dimension than were shallow items (5.28 vs. 4.89 for recollective strength, $t[22] = 2.15$, $SEM = .180$, $d = .41$, and 3.63 vs. 2.82 for recollective details, $t[22] = 4.23$, $SEM = .191$, $d = .61$). As discussed, the recollective detail judgments are arguably most relevant to the distinctiveness heuristic hypothesis, and these results confirm our assumption that deep items provided qualitatively more distinctive recollections. Finally, note that judgments of recollection strength were overall greater than those of recollected details (all p's < .01). This pattern is consistent with the idea that these two ratings reflected different aspects of subjective experience, although it is possible that subjects used the same dimension on each judgment but were more conservative on the latter. Either way, both of these recollection-based judgments yielded patterns across items that were similar to those for the "actual recollection" judgment of our recollection/familiar test, and they were quite different from the patterns obtained by our familiarity estimates. These dissociative patterns suggest that the various estimates converged on two different aspects of subjective experience—recollection and familiarity.

In sum, this experiment confirmed that the reverse-cuing and repetition procedure made shallow items more familiar than deep items, but deep items still were recollected more frequently and with more distinctive or unique details than shallow items. These results bolster the conclusion that the false recognition effect obtained in Experiment 3a (shallow test > deep test) was due to greater recollective distinctiveness of the deep items. As in the other experiments, subjects had used a recollection-based distinctiveness heuristic to suppress false recognition on the deep test.

Experiments 4 and 5

The final two experiments provided an additional test of the distinctiveness heuristic hypothesis. In Experiments 2 and 3 we tested this hypothesis using a repetition manipulation that increased the quantitative memory strength of shallow items (e.g., familiarity), but did not make them qualitatively more distinctive in memory than deep judgments (e.g., recollective distinctiveness). In contrast, in Experiments 4 and 5 we used a manipulation that potentially would enhance the recollective distinctiveness of the studied words, by adding a qualitatively different type of processing. The manipulation was transcribing the study words onto paper, just prior to making the relevant encoding judgment.

Craik (1977) found that the levels-of-processing effect was smaller than usual when subjects transcribed study words, potentially because writing the words made them more distinctive and so minimized the distinctiveness differences caused by the levels-of-processing judgments. However, Craik (1977) did not report false recognition, so there was little evidence to support this distinctiveness interpretation. In a more recent study, Seamon et al. (2003) found that hearing and then transcribing each study word reduced subsequent false recognition of nonstudied words in the DRM task, relative to a condition where subjects simply heard the study words. These findings provide stronger evidence that transcribing each word made memories more distinctive, potentially because subjects encoded the additional cognitive and motor operations involved in transcribing, as well as the perceptual characteristics of the written words (see Gallo, 2006, for further discussion). Taken together, these studies suggest that study transcription increases recollective distinctiveness, thereby reducing levels-of-processing effects (Craik, 1977) and suppressing false recognition (Seamon et al., 2003) relative to conditions without transcription.

In the currents studies, if subjects had used the distinctiveness heuristic to avoid false recognition on the deep test, then making shallow judgments more distinctive should reduce the false recognition differences across these two tests. Experiment 4 tested this prediction. Methods were similar to Experiment 2, in that words were studied in a deep list, a shallow list, or both lists, and all words in the shallow list were repeated

[6] Data from one subject was not included in this comparison because they never responded "very familiar" to a pleasantness item.

three times. The only difference from Experiment 2 was that subjects transcribed each word in the shallow study list just prior making the e-check judgment. We predicted two patterns of results. First, true recognition of shallow items would be enhanced relative to Experiment 2, potentially eliminating or reversing the levels-of-processing effect on true recognition. (Note that this experiment did not use the reverse-cuing procedure of Experiment 3, so it was unclear whether transcription would be sufficient to reverse the levels-of-processing effect on true recognition.) Second, and more importantly, because subjects could expect to recollect transcribing all of the e-check words, the failure to recollect this information would suggest that the word had not been studied in the e-check list, allowing subjects to use a distinctiveness heuristic on the shallow test. In contrast, whether a word had been transcribed would be irrelevant on the deep test, because half of the targets on that test were studied only in the pleasantness list (without transcription), whereas the other half were presented in both lists (with transcription in the e-check list). Instead, subjects had to search their memory for pleasantness judgments on this test. Under these conditions the distinctiveness advantage of transcribing e-check words (on the shallow test) should counteract the distinctiveness advantage of pleasantness judgments (on the deep test), minimizing false recognition differences across the shallow and deep tests.

Experiment 5 was similar to Experiment 4, except subjects transcribed all of the study words, regardless of the encoding judgment. As in Experiment 4, this manipulation should minimize the levels-of-processing effect on true recognition by enhancing memory for the shallow words more than the deep words. More importantly, and unlike Experiment 4, the recollection of transcribing a word would be less diagnostic of the study judgment in this experiment, because subjects could recollect transcribing words from both the e-check study list and the pleasantness study list. Instead, as in the first three experiments, subjects would have to selectively search their memory for e-check judgments on the shallow test, and pleasantness judgments on the deep test. In this case, the relatively greater distinctiveness of pleasantness judgments should help subjects to reduce false recognition on the deep test, relative to the shallow test, providing a distinctiveness heuristic pattern.

Method

Twelve University of Chicago undergraduates participated in each experiment for course credit or $10.[7] The methods of these experiments was identical to Experiment 2, with the exception that subjects in Experiment 4 were instructed to transcribe each word in the e-check list prior to making their encoding judgment, and subjects in Experiment 5 were instructed to transcribe each word in both the e-check list and the pleasantness list prior to the encoding judgment. Subjects were given a single sheet of paper upon which to write each to-be-transcribed item, and the study phase was self paced. All other procedures were identical to Experiment 2.

Results and discussion

Results from Experiment 4 are presented in the left column of Table 3, where it is obvious that having subjects selectively transcribe the e-check words led to a very different pattern of results compared to prior experiments. First, there was no difference in true recognition for shallow items and deep items (.86 and .84 on the standard test, .75 and .77 on the criterial recollection tests, and .78 for items studied with both encoding tasks, all t's < 1). Selectively transcribing the shallow items at study, in addition to repeating them, eliminated the levels-of-processing effect on true recognition. Second, and more important, the false recognition differences observed across the criterial recollection tests in the first three experiments were not observed in this experiment. A 2 (lure type) × 2 (test type) ANOVA revealed an effect of lure, $F(1,11) = 81.36$, $MSE = .006$, $\eta_p^2 = .881$, but no effect of test or interaction (both F's < 1). This pattern suggests that selectively transcribing items in the shallow list at study enhanced their recollective distinctiveness to the level of items in the deep list, and in fact, performance on the two criterial recollection tests was quite similar. Unlike the first three experiments, recollection-based monitoring was equally effective on the two criterial recollection tests in this experiment.

The results from Experiment 5 are presented in the right column of Table 3. Having subjects transcribe all of the studied items (and repeating the shallow items) eliminated the levels-of-processing effect on true recognition (hits to shallow items and deep items were .89 and .85 on the standard test, .85 and .85 on the criterial recollection tests, and .89 and .90 for items studied in both encoding tasks, all t's < 1). However, there was one major difference between this experiment and Experiment 4—the false recognition pattern predicted by the distinctiveness heuristic hypothesis was restored in Experiment 5. A 2 (lure type) × 2 (test type) ANOVA revealed an effect of lure, $F(1,11) = 21.17$, $MSE = .031$, $\eta_p^2 = .658$, an effect of test, $F(1,11) = 7.44$, $MSE = .021$, $\eta_p^2 = .403$, and a marginal interaction, $F(1,11) = 4.84$, $MSE = .012$, $\eta_p^2 = .306$, $p = .05$. False recognition was significantly greater on the shallow test than on the deep test for studied lures (.45 and .27, $t[11] = 2.92$, $SEM = .063$, $p < .05$, $d = .66$), but this effect was not significant for nonstudied

[7] Data from one subject in Experiment 4 was replaced due to an error in the experimental script that miscoded the test items.

Table 3
Mean recognition of each item type as a function of test type in Experiments 4 and 5

	Experiment 4 Transcribe shallow	Experiment 5 Transcribe all
Standard test		
Both hits	.98 (.01)	.98 (.01)
Shallow hits	.86 (.03)	.89 (.03)
Deep hits	.84 (.03)	.85 (.04)
New FAs	.15 (.04)	.16 (.06)
Shallow test		
Both hits	.78 (.05)	.89 (.02)
Shallow hits	.75 (.05)	.85 (.04)
Deep FAs	.28 (.05)	.45 (.09)
New FAs	.09 (.03)	.15 (.05)
Deep test		
Both hits	.78 (.07)	.90 (.03)
Deep hits	.77 (.07)	.85 (.04)
Shallow FAs	.30 (.05)	.27 (.07)
New FAs	.08 (.02)	.10 (.05)

Notes. Pleasantness judgments were made once per study word, whereas e-check judgments were made three times (standard cuing procedure). Standard errors of each mean are in parenthesis. FAs, false alarms.

lures (.15 and .10, $t[11] = 1.17$, $SEM = .039$, $p = .27$, $d = .28$), although the effect was in the predicted direction.

The false recognition difference across tests (shallow > deep) cannot be attributed to overall memory strength, because hits to shallow items were equated with hits to deep items. Instead, this effect suggests that subjects took advantage of enhanced distinctiveness of pleasantness judgments, relative to e-check judgments, to avoid false recognition on the deep test. In contrast to the situation in Experiment 4, transcribing all of the study items in Experiment 5 (i.e., those in the shallow list and those in the deep list) made the recollection of this transcription less informative as to whether one had performed a shallow judgment or a deep judgment. Instead, subjects had to search their memory for e-check judgments on the shallow test and for pleasantness judgments on the deep test, and as in the first three experiments, evidence for a distinctiveness heuristic on the deep test was obtained.

Finally, we report within-test comparisons for these two experiments. Consider Experiment 4 first. On the standard test, both hits (.98) > shallow hits (.86) = deep hits (.84) > new false alarms (.15). On the shallow test, both hits (.78) = shallow hits (.75) > deep false alarms (.28) > new false alarms (.09). On the deep test, both hits (.78) = deep hits (.77) > shallow false alarms (.30) > new false alarms (.08). Consider next Experiment 5. On the standard test, both hits (.98) > shallow hits (.89) = deep hits (.85) > new false alarms (.16). On the shallow test, both hits (.89) = shallow hits (.85) > deep false alarms (.45) > new false alarms (.15). On the deep test, both hits (.90) = deep hits (.85) > shallow false alarms (.27) > new false alarms (.10). With the exception of the critical differences in the levels-of-processing effects on false recognition, these patterns were orderly and similar to those observed in the other experiments.

General discussion

Deep levels of processing can suppress false recognition, relative to shallow levels of processing, by enhancing recollective distinctiveness. Fig. 1 summarizes the false recognition results observed in the four experiments reported here that were predicted to show these effects. In all four cases false recognition was less likely on the deep test, compared to the shallow test, and these effects were found for false recognition of both studied and nonstudied lures. The only difference across these tests was in the instructions, with subjects searching memory for recollections of pleasantness judgments on the deep tests and e-check judgments on the shallow test. As a result, differences in false recognition across these tests can be attributed to differences in retrieval expectations, with subjects expecting more distinctive recollections on the deep test. The only exception to this pattern was Experiment 4, and this exception also was predicted by the distinctiveness heuristic hypothesis because we selectively enhanced the distinctiveness of e-check words via study transcription.

The distinctiveness heuristic hypothesis gains further support from the fact that, across experiments, we dissociated the effect of levels-of-processing on hit rates from the effect of retrieval orientation on false recognition. In Experiment 1 a large levels-of-processing effect was found on hits to studied items (deep > shallow), in Experiment 2 this effect was attenuated, in Experiment 3a this effect was reversed (shallow > deep), and in Experiment 5 the hit rates were equated across the conditions (deep = shallow). Further, Experiment 3b confirmed that the study procedures used in Experiment 3a made shallow items more familiar than deep items. Despite these varying patterns of true recognition, false recognition always was lower on the deep test than on the shallow test. This dissociation is theoretically important, because it shows that the false recognition patterns reported here cannot be explained by differences in the relative strength or familiarity of the test lures. In the same vein, explanations of false recognition differences that appeal to strength or familiarity-based criterion shifts have difficulty explaining these patterns, because the false recognition effects were consistent even though the relative levels of memory strength and familiarity for the different studied items varied widely across experiments.

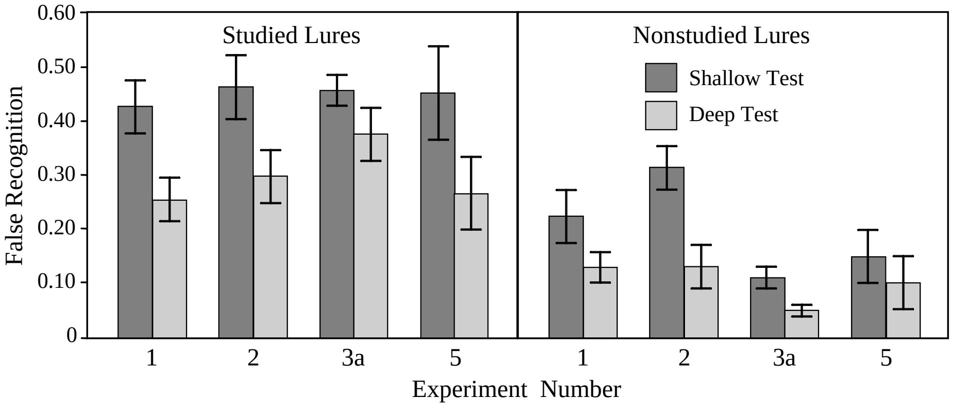

Fig. 1. False recognition in the four experiments where deep judgments were predicted to be more distinctive than shallow judgments. True recognition was greater for deep judgments in Experiments 1 and 2, for shallow judgments in Experiment 3a, and equated across these two judgments in Experiment 5.

Retrieval monitoring implications

The current findings make two important theoretical contributions. The first has implications for our understanding of retrieval monitoring processes. Our results suggest that a recollection-based distinctiveness heuristic can be based on a manipulation of conceptual distinctiveness (e.g., a deep level of processing). Prior studies of the distinctiveness heuristic have focused only on perceptual distinctiveness (e.g., pictures versus words in Schacter et al., 1999; saying words aloud versus simply reading them in Dodson & Schacter, 2001), or on distinctiveness arising from cognitive operations (see Footnote 1). However, according to the source-monitoring framework (Johnson et al., 1993), the accuracy of retrieval monitoring will benefit to the degree that any type of information helps to distinguish one source from another in memory (e.g., perceptual features, conceptual features, cognitive operations, emotional reactions, etc.). The current results are consistent with this framework, and show that conceptual distinctiveness, much like perceptual distinctiveness or cognitive operations, can facilitate retrieval monitoring accuracy. Further, by dissociating the effects of levels-of-processing on true and false recognition, particularly in Experiment 3b, we were able to eliminate strength and familiarity-based criterion shifts as an explanation for the false recognition effects observed here. Of course, this is not to say that such criterion shifts never influence false recognition, only that models based on a concept of overall memory strength or familiarity, without incorporating the idea of recollective distinctiveness, cannot explain the current results.

The conclusion that deep processing enhanced retrieval monitoring in our experiments also helps to reconcile previously inconsistent results. As described in the Introduction, some studies have demonstrated patterns that are consistent with the current results (e.g., Davies & Cubbage, 1976; Elias & Perfetti, 1973; Jacoby et al., 2005) but others have shown the opposite pattern (e.g., Chan et al., 2005; Thapar & McDermott, 2001). The current study points to the importance of task design in obtaining distinctiveness effects on false recognition. As described in the Introduction, studies that found the opposite effect manipulated levels-of-processing within-subjects and used a single recognition test at retrieval. These are not ideal conditions to test the distinctiveness heuristic hypothesis (and indeed, they were not designed to do so). Further, these studies used the DRM task, which is particularly sensitive to the influence of associative activation on false recognition, potentially obscuring the influence of monitoring processes (see Gallo, 2006, for discussion and review). The criterial recollection task used here was specifically designed to isolate recollection-based monitoring processes, controlling other factors that can influence false recognition, and we found more definitive evidence for the use of the distinctiveness heuristic.

Experiments 4 and 5 provided additional evidence that the current effects were due to the distinctiveness heuristic. Prior work by Seamon et al. (2003) indicated that transcribing heard words during study could suppress subsequent false recognition, compared to a condition where subjects simply listened to the words. These findings suggest that transcribing words was a distinctive form of processing, thereby facilitating the distinctiveness heuristic. In Experiment 4 we had subjects transcribe all of the words in the shallow list, but none of the words in the deep list. It was predicted that this manipulation would enhance the distinctiveness of the words in the shallow list, thereby counteracting the distinctiveness benefits afforded by the pleasantness judgments. Consistent with this idea, this was the only experiment in our series in which there were no false recognition differences across the shallow and the deep tests. Instead, false recognition was equally low in both conditions, suggesting that retrieval monitoring was equally effective. It was further predicted that having subjects transcribe all of the study words in Experiment 5 would minimize the benefits of this manipulation, because under these conditions the recollection of study transcription was less diagnostic of presentation in either the shallow list or the deep list. Consistent with this idea, the effect of levels of processing on false recognition (shallow > deep) re-emerged in Experiment 5. It is difficult to explain our transcription effects as a strength-based criterion shift, because true recognition was equated across the conditions of these experiments. These results provide more evidence for the use of a recollection-based distinctiveness heuristic, and that this heuristic is sensitive to the diagnosticity of the information that subjects can expect to recollect from different sources.

Levels-of-processing implications

The second contribution of the current study is to provide new evidence for the idea that the typical levels-of-processing effect on true memory is based on recollective distinctiveness. As discussed in the Introduction, the idea of distinctiveness can explain why various semantic encoding factors tend to have a larger influence on typical recall or recognition tests than surface levels of processing. Other evidence comes from studies showing that the uniqueness of the semantic encoding conditions or the uniqueness of the retrieval cues is an important determinant of recall (e.g., Craik & Tulving, 1975; Moscovitch & Craik, 1976). These latter findings indicate that uniqueness or distinctiveness within a level of processing can influence performance, consistent with many other studies highlighting the importance of distinctiveness for recall and recognition (for recent reviews see Hunt, 2006; McDaniel & Geraci, 2006). However, showing that distinctiveness can influence performance is not the same as showing that distinctiveness is the critical factor that differentiates deep and shallow encoding tasks, and indeed, depth and distinctiveness are sometimes considered as separate factors (cf. Lockhart & Craik, 1990). When viewed in this way, the false recognition data reported here provide stronger evidence that deep levels of processing, per se, lead to more distinctive memories of words than do shallow levels of processing (i.e., the encoding of more unique features for each item). Of course, deep processing also might enhance familiarity (e.g., Yonelinas, 2002), but our point here is that familiarity differences alone cannot explain all levels-of-processing effects, and that recollective distinctiveness appears to play a very significant role.

Even though we have adopted a feature-based definition of distinctiveness, it is still a relative concept, and the factors that make an item distinctive will differ across situations. Distinctiveness can be influenced by the uniqueness of semantic or conceptual features in memory, holding the perceptual input relatively constant, as suggested by the levels-of-processing effects in the current experiments. Distinctiveness also can be influenced by the complexity and uniqueness of perceptual features in memory, holding conceptual processing relatively constant, as described in studies that compared words to pictures (e.g., Gallo et al., 2004; Schacter & Wiseman, 2006). The fact that both perceptual and conceptual manipulations can reduce false recognition in the criterial recollection task, with all other factors being equal, adds credence to the idea that both conceptual and perceptual features can contribute to recollective distinctiveness.

The idea that deep processing of words leads to more distinctive recollections than shallow processing can explain many findings, but one remaining question is why deep processing is more distinctive. One explanation is that, by definition, words exist to convey an infinite number of unique meanings while using a very limited set of surface features (e.g., finite orthography and phonology). In the current task, as in most levels-of-processing experiments, the words were presented using homogenous perceptual parameters, making surface features a relatively impoverished means of differentiating them relative to the vastly different conceptual associations subjects could have made to each word during the pleasantness judgments (cf. Moscovitch & Craik, 1976). When viewed in this way, the levels-of-processing effect on true memory does not reflect a general semantic superiority effect in memory. Instead, the most effective way to encode items for subsequent recall or recognition is to associate each item with information from preexisting knowledge that can later provide a large number of unique features to retrieve. Semantic information is the domain of knowledge that best differentiates lists of unrelated words, but other domains will be more important for other stimuli in other contexts (e.g., perceptual features for pictures or faces, Intraub & Nicklos, 1985). As described by Lockhart and Craik (1990), deep levels of processing lead to better retention, but exactly what qualifies as "deep" depends on the domain of knowledge involved.

In conclusion, we emphasize the importance of qualitative over quantitative aspects of memories during the retrieval monitoring process. From a source memory or multidimensional perspective, memories can differ qualitatively (i.e., the "type" of information stored in memory, with some recollections being more distinctive than others) as well as quantitatively (i.e., the relative "strength" of memories of a given type, with some being more familiar or being recollected more frequently than others). In the current study, repeating the shallow items made them quantitatively stronger in memory, but it was not thought to alter the quality of processing, and it did not eliminate our false recognition effects. Instead, the addition of a qualitatively new encoding task (transcription) was needed to eliminate these false recognition effects. Analogous dissociations between memory "strength" and false recognition effects have been observed in studies comparing pictures and words, using the current task (e.g., Gallo et al., 2004, 2007) as well as other tasks that used more typical recognition tests (e.g., Dodson & Schacter, 2002; Schacter et al., 1999). The fact that false recognition was predominantly influenced by recollective distinctiveness in all of these studies suggests that subjects tend to weigh qualitative information more heavily than quantitative information when making their memory decisions. This is not to say that quantitative differences are unimportant, but it is to say that subjects seem to prefer to rely on qualitative differences when possible. This conclusion is in the spirit of the original levels-of-processing framework, which emphasized that the quantity of rehearsals at encoding were not as important as the qualitative nature of the processing. Here we extend that notion to retrieval, emphasizing the importance of qualitative memory differences over quantitative differences in retrieval monitoring processes.

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.

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & Cognition, 35*, 201–210.

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.

Chan, J. C. K., McDermott, K. B., Watson, J. M., & Gallo, D. A. (2005). The importance of material-processing interactions in inducing false memories. *Memory & Cognition, 33*, 389–395.

Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology, 33A*, 497–505.

Coltheart, V. (1977). Recognition errors after incidental learning as a function of different levels of processing. *Journal of Experimental Psychology: Human Learning and Memory, 3*, 437–444.

Craik, F. I. M. (1977). Depth of processing in recall and recognition. In S. Dormic & P. M. A. Rabbitt (Eds.), *Attention and performance VI* (pp. 679–697). Hillsdale, NJ: Erlbaum.

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*, 671–684.

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General, 104*, 268–294.

Davies, G., & Cubbage, A. (1976). Attribute coding at different levels of processing. *Quarterly Journal of Experimental Psychology, 28*, 653–660.

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. (2002). When false recognition meets metacognition: The distinctiveness heuristic. *Journal of Memory and Language, 46*, 782–803.

Elias, C. S., & Perfetti, C. A. (1973). Encoding task and recognition memory: The importance of semantic encoding. *Journal of Experimental Psychology, 99*, 151–156.

Fisher, R. P., & Craik, F. I. M. (1977). Interaction between encoding and retrieval operations in cued recall. *Journal of Experimental Psychology: Human Learning and Memory, 3*, 701–711.

Gallo, D. A. (2006). *Associative illusions of memory: False memory research in DRM and related tasks*. New York: Psychology Press.

Gallo, D. A., Cotel, S. C., Moore, C. D., & Schacter, D. L. (2007). Aging can spare recollection-based retrieval monitoring: The importance of event distinctiveness. *Psychology and Aging, 22*, 209–213.

Gallo, D. A., Weiss, J. A., & Schacter, D. L. (2004). Reducing false recognition with criterial recollection tests: Distinctiveness heuristic versus criterion shifts. *Journal of Memory and Language, 51*, 473–493.

Gunter, R. W., Bodner, G. E., & Azad, T. (2007). Generation and mnemonic encoding induce a mirror effect in the DRM paradigm. *Memory & Cognition, 35*, 1083–1092.

Higham, P. A., & Vokey, J. R. (2004). Illusory recollection and dual-process models of recognition memory. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 57A*, 714–744.

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.

Hunt, R. R. (2006). The concept of distinctiveness in memory research. In R. R. Hunt & J. B. Worthen (Eds.), *Distinctiveness and memory* (pp. 3–26). Oxford: Oxford University Press.

Hyde, T. S., & Jenkins, J. J. (1969). Differential effects of incidental tasks on the organization of recall of a list of highly associated words. *Journal of Experimental Psychology, 82*, 472–481.

Intraub, H., & Nicklos, S. (1985). Levels of processing and picture memory: The physical superiority effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*, 284–298.

Jacoby, L. L., & Craik, F. I. M. (1979). Effects of elaboration of processing at encoding and retrieval: Trace distinctiveness and recovery of initial context. In L. S. Cermak & F. I. M. Craik (Eds.), *Levels of processing in human memory* (pp. 1–21). Hillsdale, NJ: Erlbaum.

Jacoby, L. L., & Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General, 110*, 306–340.

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

Johnson, M. J., Hashtroudi, S., & Lindsay, D. S. (1993). 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.

Lockhart, R. S., & Craik, F. I. M. (1990). Levels of processing: A retrospective commentary on a framework for memory research. *Canadian Journal of Psychology, 44*, 87–112.

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.

McDaniel, M. A., & Geraci, L. (2006). Encoding and retrieval processes in distinctiveness effects: Towards an integrative framework. In R. R. Hunt & J. B. Worthen (Eds.), *Distinctiveness and memory* (pp. 65–88). Oxford: Oxford University Press.

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior, 16*, 519–533.

Moscovitch, M., & Craik, F. I. M. (1976). Depth of processing, retrieval cues, and uniqueness of encoding as factors in recall. *Journal of Verbal Learning and Verbal Behavior, 15*, 447–458.

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.

Parkin, A. J. (1983). The relationship between orienting tasks and the structure of memory traces—Evidence from false recognition. *British Journal of Psychology, 74*, 61–69.

Roediger, H. L., III., & Gallo, D. A. (2001). Levels of processing: Some unanswered questions. In M. Naveh-Benjamin, M. Moscovitch, & H. L. Roediger (Eds.), *Perspectives on human memory and cognitive aging: Essays in honour of Fergus Craik* (pp. 28–47). New York: Psychology Press.

Roediger, H. L., III., Gallo, D. A., & Geraci, L. (2002). Processing approaches to cognition: The impetus from the levels-of-processing framework. *Memory, 10*, 319–332.

Roediger, H. L., III., Weldon, M. S., & Challis, B. H. (1989). Explaining dissociations between implicit and explicit measures of retention: A processing account. In H. L. Roediger & F. I. M. Craik (Eds.), *Varieties of memory and consciousness: Essays in honour of Endel Tulving* (pp. 3–41). Hillsdale, NJ: Lawrence Erlbaum.

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., & Wiseman, A. L. (2006). Reducing memory errors: The distinctiveness heuristic. In R. R. Hunt & J. Worthen (Eds.), *Distinctiveness and memory* (pp. 89–107). New York: Oxford University Press.

Seamon, J. G., Goodkind, M. S., Dumey, A. D., Dick, E., Aufseeser, M. S., Strickland, S. E., et al. (2003). "If I didn't write it, why would I remember it?" Effects of encoding, attention, and practice on accurate and false memory. *Memory & Cognition, 31*, 445–457.

Seamon, J. G., & Virostek, S. (1978). Memory performance and subject-defined depth of processing. *Memory & Cognition, 6*, 283–287.

Smith, R. E., & Hunt, R. R. (1998). Presentation modality affects false memory. *Psychonomic Bulletin & Review, 5*, 710–715.

Stein, B. S. (1978). Depth of processing reexamined: The effects of the precision of encoding and test appropriateness. *Journal of Verbal Learning and Verbal Behavior, 17*, 165–174.

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.

Thapar, A., & McDermott, K. B. (2001). False recall and false recognition induced by presentation of associated words: Effects of retention interval and level of processing. *Memory & Cognition, 29*, 424–432.

Toth, J. P. (1996). Conceptual automaticity in recognition memory: Levels-of-processing effects on familiarity. *Canadian Journal of Experimental Psychology, 50*, 123–138.

Tulving, E. (1979). Relation between encoding specificity and levels of processing. In L. S. Cermak & F. I. M. Craik (Eds.), *Levels of processing in human memory* (pp. 405–428). Hillsdale, NJ: Erlbaum.

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

Verde, M. F., & Rotello, C. M. (2007). Memory strength and the decision process in recognition memory. *Memory & Cognition, 35*, 254–262.

Wallace, W. P. (1968). Incidental learning: The influence of associative similarity and formal similarity in producing false recognition. *Journal of Verbal Learning and Verbal Behavior, 7*, 50–54.

Wilson, M. D. (1988). The MRC Psycholinguistic Database: Machine Readable Dictionary, Version 2. *Behavioural Research Methods, Instruments and Computers, 20*, 6–11.

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.