Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Memory and Language 52 (2005) 493–504

Journal of Memory and Language

www.elsevier.com/locate/jml

Age differences in depth of retrieval: Memory for foils ☆

Larry L. Jacoby *, Yujiro Shimizu, Katerina Velanova, Matthew G. Rhodes

Department of Psychology, Washington University, St. Louis, MO 63130, USA

Received 10 September 2004, revision received 11 January 2005
Available online 23 February 2005

Abstract

Control over memory can be achieved in two ways: by constraining retrieval such that only sought after information comes to mind or, alternatively, by means of post-access monitoring. We used a *memory-for-foils* paradigm to gain evidence of differences in retrieval constraints. In this paradigm, participants studied words under deep or shallow encoding conditions and were given a memory test that required them to discriminate between new items (foils) and either deep or shallow targets. A final recognition test was used to examine memory for the foils. For young adults, foil memory was superior when participants attempted to retrieve deep, rather than shallow, targets on the earlier test. In contrast, older adults showed no difference in memory for foils from the two types of tests. We discuss the importance of differences in depth of retrieval processes for theories of metacognition and for understanding age-related differences in memory performance.

Keywords: Aging; Cognitive control; Memory

Cognitive control of memory can be viewed as analogous to *quality* control in manufacturing (Jacoby, Shimizu, Daniels, & Rhodes, in press). An obvious means of handling deficiencies in quality control is to increase the number of inspectors that monitor manufactured goods, rejecting those that do not meet standards. However, a more efficient method is to increase the precision of production techniques so as to meet standards more reliably, thereby reducing the need for inspectors. The two means of quality control are in fact interrelated. That is, increasing the precision of production techniques can produce *qualitative* differences in the form of evaluation carried out (e.g., a change in those aspects of manufactured goods that are most closely monitored) and reduce the number of inspectors needed. In turn, imperfections revealed by inspectors can guide the improvement of production techniques.

The quality-control analogy is illustrative of models of executive function (e.g., Burgess & Shallice, 1996) and metacognition (e.g., Koriat & Goldsmith, 1996; see Fernandez-Duque, Baird, & Posner, 2000, for a discussion of the close relationship between the two types of models). For example, memory monitoring processes proposed in models of metacognition (e.g., Koriat & Goldsmith, 1996; Nelson, 1996) serve a role similar to that of quality-control inspectors by assessing the adequacy of a candidate response after it is retrieved, as in the use of confidence judgments to control the output of responses. Conversely, production techniques correspond to retrieval processes that are instrumental for bringing potential responses to mind. Imposing

☆ This research was supported by National Institute on Aging Grant AG13845. We are grateful to Carole Jacoby for her data-collection assistance.
* Corresponding author. Fax: +1 314 935 7588.
E-mail address: lljacoby@artsci.wustl.edu (L.L. Jacoby).

0749-596X/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.jml.2005.01.007

constraints on retrieval provides an alternative to monitoring as a means of controlling accuracy. Thus, the contrast is between gaining control by increasing constraints on memory retrieval compared to relying on a post-retrieval monitoring process.

Most prior investigations of metacognition and cognitive control have examined the post-retrieval monitoring of *quantitative* differences in memory, such as those revealed by confidence judgments (e.g., Koriat & Goldsmith, 1996) or judgments of learning that are used to guide study (e.g., Dunlosky & Nelson, 1992). Instead, we focus on *qualitative* differences in memory that reflect differences in constraints on memory access and examine age-related differences in the ability to constrain retrieval processes. We do so by examining memory for foils (new words) presented in a test of recognition memory (Jacoby et al., in press).

Recognition memory is traditionally described as a judgment relying on unidimensional trace strength or familiarity. For example, global activation models of memory (e.g., Gillund & Shiffrin, 1984) suggest that recognition is accomplished by assessing a memory probe's strength (familiarity) against a decision criterion. If the probe's value exceeds criterion, it is accepted as "old"; otherwise, it is rejected as "new." The greater the match between a memory probe and traces in memory, including the match between study and test contexts, the greater the level of familiarity. Thus, the emphasis lies in the *quantitative* relationship between the familiarity of a probe and a decision criterion. In contrast, we emphasize the *kind* of memory that is sought. Specifically, we propose that processes implemented during study are sometimes re-implemented during retrieval by the rememberer. This approach parallels that underlying the notion of transfer-appropriate processing (e.g., Morris, Bransford, & Franks, 1977), in that it emphasizes the similarity of processes implemented during study and test. The important difference is that whereas transfer-appropriate processing focuses on self-initiated processes during study that may or may not match the form of test, we focus on cue elaboration at the time of test that serves to recapitulate study processing and, thereby, constrain memory retrieval (e.g., Jacoby, Kelly, & Dywan, 1989).

The experiments reported in this article used memory for foils from a recognition-memory test to both gain evidence of qualitative differences in the bases for recognition memory (Jacoby et al., in press; Shimizu & Jacoby, in press; Velanova et al., 2003) and to examine age-related differences in constraining memory retrieval. Because of a deficit in the type of self-initiated processing required to engage in recollection, older adults rely more heavily on familiarity as a basis for recognition memory than do young adults (e.g., Jacoby, 1999; Koriat, Ben-Zur, & Sheffer, 1988). As will be described, older adults' greater reliance on familiarity, a less constrained form of memory access, was expected to result in a reduction in their depth of retrieval processing.

Experiment 1 was designed to show that source-constrained retrieval can produce a *qualitative* change in the type of information used by young adults for recognition memory judgments, which is not evident for older adults. Specifically, during the first phase of Experiment 1, participants either judged the pleasantness of words in a list (deep-processing group) or made judgments about the vowels of words in a list (shallow-processing group). In the second phase of the experiment, both groups were administered a recognition-memory test for words studied in the first phase. We expected results from the recognition-memory test given in Phase 2 to replicate the results of prior experiments (e.g., Craik & Lockhart, 1972). That is, judging the pleasantness of words (deep processing) should produce better recognition-memory performance than making vowel judgments (shallow processing).

More important than quantitative differences in recognition performance, we expected qualitative differences in the bases for recognition memory in Phase 2. This follows because recognition is held to be accomplished by constraining retrieval processing in a way that recapitulates study processing. Consequently, when attempting to recognize pleasantness-judged old words, participants would likely process the meaning of *both* targets and foils, perhaps considering each test word's pleasantness to determine whether they had made a similar judgment previously. In contrast, attempting to recognize vowel-judged, old words would likely rely less on the processing of meaning. The deeper processing of foils when pleasantness-judged words were targets was expected to result in higher subsequent recognition memory for those foils.

To gain evidence of differences in retrieval processing, a third phase of the experiment tested recognition memory for *new items* (*foils*) that appeared on the recognition-memory tests given in Phase 2. For young adults, we expected depth of retrieval to be constrained by the depth of study processing of old words in the recognition-memory test, with the result that subsequent memory for foils would be greater in the deep than in the shallow processing condition. Conversely, older adults might rely more heavily on familiarity than young adults for recognition of old words in both types of test lists (cf. Jacoby, 1999), without further constraining their retrieval processing when pleasantness-judged words were targets. If this were the case, older adults would not exhibit differences in memory for foils from the two types of test lists.

Such predictions based on depth of retrieval processing differ considerably from what would be expected on the basis of global familiarity. That is, if participants simply assess global familiarity when making recognition judgments (e.g., Gillund & Shiffrin, 1984), there would be no reason to expect differential processing of the foils depending on the study processing of targets, and, thus, no reason to expect differences in subsequent memory for foils. In fact, any predictions made by a familiarity account would likely suggest that because recognition decisions would be more difficult for shallowly processed items, increasing their exposure duration, foils from such lists would enjoy a recognition advantage. In contrast, better memory for foils when targets were deeply processed would provide direct evidence of differences in *retrieval depth* that reflects constrained memory retrieval.

Experiments 2 and 3 further examined age differences in memory for foils. To anticipate, in each of the experiments reported, differences in memory for foils that reflected differences in depth of retrieval processes were found for younger adults, but were absent for older adults. In the General discussion, we consider the importance of constraining memory access as a means of cognitive control for theories of metacognition. We also discuss the implications of our results for theories of recognition memory and for understanding age-related differences in memory performance.

Experiment 1

Method

Participants

Thirty-two young adults and 32 older adults participated in the experiment. The young adults were Washington University undergraduates who received course credit for their participation. Their mean age was 19.6 years (range 18–26 years) and their mean score on the Shipley Vocabulary test was 34.34 (range 26–37). The older adults were recruited from the Washington University older adult subject pool. Their mean age was 75.8 years (range 61–87 years), and their mean score on the Shipley Vocabulary test was 35.03 (range 26–40). Educational attainment for the older adults ranged from a high school diploma ($n = 9$), to a college degree ($n = 11$), to a post-graduate degree ($n = 12$). All participants spoke fluent English and were tested individually.

Materials and design

Stimuli were 272 words (216 critical and 56 buffers) matched in frequency (1–104 per million, $M = 15.67$, Kucera & Francis, 1967), length (4–8 letters, $M = 6.67$), syllables (1–4, $M = 2.15$), and the presence of an O or U. Assignment of words to each condition was fully counterbalanced across participants. Age and Levels of Processing were between-participant factors with Item status (target, foil) manipulated within-participants.

Procedure

In Phase 1, participants were given one of two orienting tasks to perform on a list of 88 words. Of these words, 8 served as practice items, 8 served as buffers (with 4 of these at the beginning of the list and 4 at the end), and 72 served as study items. Participants in the *Deep* condition made pleasantness judgments for each word, whereas participants in the *Shallow* condition made vowel judgments, indicating whether each word included an O or U. Words were presented at a 1.5-s rate in the center of the screen and judgments were made as self-paced key presses.

In Phase 2, participants were given a recognition-memory test for words that had been presented in Phase 1. That test included the 72 critical study items intermixed with an equal number of new items (72 foils). An additional 10 items (5 old and 5 new) served as primacy and recency buffers.

In Phase 3, we tested memory for words that had served as foils in the recognition-memory test given in Phase 2. The 72 foils from that test were intermixed with 72 new words. Participants were instructed to judge a word as "old" if it had been presented earlier during *any phase* of the experiment, and to respond "new" only if the word had not been presented earlier. We emphasized that "old" words included all the words that were foils in the earlier recognition test. All recognition tests were self-paced with responses made as key presses.

Unless otherwise noted, the alpha level for all statistical tests reported in Experiment 1 and subsequent experiments was set to .05.

Results and discussion

Recognition memory

A summary of recognition performance is displayed in Table 1. As expected, both young and older adults exhibited more hits and fewer false alarms for the deep than for the shallow processing condition, $F(1, 60) = 303.27$, $Mse = 1.93$. In addition, compared to young adults, older adults were less likely to correctly recognize old words and were more likely to incorrectly call new words "old," $F(1, 60) = 32.00$, $Mse = .20$. The three-way interaction of level of processing by age group and old/new status was not significant, $p > .20$.

Table 1
Probability of responding "old" for the recognition-memory test of Experiment 1

Group	Item type			
	Deep-old	Deep-new	Shallow-old	Shallow-new
Young	.93	.03	.60	.15
Elderly	.87	.09	.50	.26

Final test performance: Memory-for-foils

Results from the test of memory for foils (Table 2) showed that young adults exhibited better recognition of foils and fewer false alarms following deep, as compared to shallow, study processing. Older adults did not demonstrate an effect of study processing on correct recognition of foils but did show an effect on false alarms. The interaction of study processing, age group, and old/new status was not significant, $F(1, 60) = 2.09$, $Mse = .01$, $p < .16$. However, because of our a priori predictions, we conducted separate analyses for the two age groups.

For young adults, the interaction of study processing and old/new status was significant, $F(1, 30) = 12.38$, $Mse = .11$. This is equivalent to showing that recognition corrected for bias by subtracting false alarms from hits was higher after deep study processing (.78) than after shallow study processing (.61). Simple effects tests revealed that the probability of correctly calling an item "old" was marginally higher for the deep condition than for the shallow condition (.87 vs. .79), $F(1, 30) = 3.03$, $Mse = .05$, $p < .10$, whereas the probability of incorrectly calling a new item "old" was significantly lower for the deep condition than for the shallow condition (.09 vs. .18), $F(1, 30) = 8.56$, $Mse = .05$.

For older adults, the interaction of study processing with old/new status was not significant, $F(1, 30) = 1.67$, $Mse = .01$, $p > .20$. This indicates that recognition corrected for bias by subtracting false alarms from hits did not differ statistically between the deep (.54) and the shallow study conditions (.48). Simple effects tests confirmed that the probability of correctly calling an item "old" was statistically equivalent for the deep and shallow conditions (.73 vs. 74), $F < 1$. The probability of a false alarm, though numerically different, did not differ statistically between the two conditions, $F(1, 30) = 2.56$, $Mse = .03$, $p > .11$.

The effect of depth of processing on memory for foils demonstrated by young adults replicates results reported by Jacoby et al. (in press). However, the advantage in memory for foils after deep, as compared to shallow, study processing is substantially larger in the current experiment (.78 vs. .61) than in that reported by Jacoby et al. (.84 vs. .76). The reason for this difference is likely the use of a between-participants manipulation of depth in the current experiment compared to the within-participants manipulation used by Jacoby et al. Intermixing deep and shallow old foils in the test of memory for foils, as done by Jacoby et al., likely reduces the degree to which participants can constrain their retrieval processing in a way that is appropriate for recognizing a particular type of old foil. More importantly, the between-participants design allows the basis for rejection of new foils to vary with the depth of processing of old foils, and much of the difference in memory for foils in the current experiment was because of a difference in false alarm rates. It may be easier to reject a new foil because its pleasantness was not considered on an earlier test than it is to reject a new foil because its vowels were not earlier considered.

Table 2
Probability of responding "old" for the foil recognition test of Experiment 1

Group	Item type			
	Deep-foil	Deep-new	Shallow-foil	Shallow-new
Young	.87	.09	.79	.18
Elderly	.73	.19	.74	.26

Differences in memory for foils provide evidence that retrieval processing was deeper when target words had been deeply processed during study. Consistent with results reported by Jacoby et al. (in press), further analyses ruled out the possibility that this advantage in memory for foils resulted from increased difficulty of the earlier recognition-memory test. Specifically, foil memory for the young was better after deep study processing, which produced fewer false alarms on the initial recognition test (.03), than after shallow study processing (.15). Moreover, rejection time did not differ between the deep (1081 ms) and shallow conditions (1093 ms) for the young ($F < 1$). In sum, it was the depth of retrieval as constrained by the depth of study processing that was responsible for the difference in memory for foils, rather than a difference in the difficulty or amount of time to reject the foils on the prior test.

Although the effect of depth of processing on memory for foils was smaller for older adults, they did show a tendency toward better memory for foils from tests of deeply processed items. The finding of an effect of level of processing for the recognition-memory tests given in Phase 2 also suggests that older participants used meaning for judgments of recognition. Indeed, the levels effect for older adults was similar to that for young adults although overall recognition was lower for older adults. The difference in overall recognition produces difficulties for interpreting results, making it unclear whether any differences in memory for foils are the result of age-related differences in memory or, rather, the result of the lower recognition-memory performance. This problem was addressed in Experiment 2.

Experiments 2a and 2b

In Experiment 2, rather than varying the orienting task, we examined depth of retrieval using a manipulation of materials, varying the form of similarity shared by words in a study list. Shimizu and Jacoby (in press) have demonstrated that for young adults, manipulating similarity among words studied in a short list influences depth of retrieval, as indexed by memory for foils. As in their procedure, in the first phase of Experiment 2 participants were shown a series of four-word lists with a test of recognition memory following each of the lists. The study sets consisted of either semantically related (e.g., “BED, REST, WAKE, and DREAM”) or orthographically similar (e.g., “TRUCK, TRAIN, TREND, and TRAMP”) words. Following each study list, participants received a test probe that was either an old word (one of the four words just presented) or a foil. Importantly, the foils were always *unrelated* in meaning and appearance to the words in the study set. We expected the form of the relationship among words in a study list to create a *processing set* that would constrain the depth of retrieval processing of test probes, both targets and foils. Consequently, foils that followed the semantically related lists were expected to be rejected on the basis of their meaning, whereas foils following the orthographically related lists were expected to be rejected on the basis of their appearance.

Both young and older adults were administered a test of memory for foils taken from the orthographically and semantically related lists. For young adults, foils following semantically related lists were expected to be more deeply processed and better remembered than were foils that followed orthographically related lists, replicating results reported by Shimizu and Jacoby (in press). In contrast, we predicted that older adults might be less flexible in their retrieval processing and rely on appearance for recognition judgments rather than meaning, regardless of the relationship among words in a list. This would eliminate differences in memory for foils from tests for the two types of list.

Recognition performance on test probes following short lists was expected to be near perfect for young and older adults, equating their recognition-memory performance on the initial tests. Of course, age equivalence on the initial tests might simply reflect a ceiling effect, with a memory advantage for younger adults being hidden by the near-perfect performance of both age groups. A more useful measure of age equivalence can be gained by examining memory for foils that followed orthographically related lists. Both young and older adults were expected to reject those foils on the basis of their appearance (orthography). Consequently, we did not expect memory for foils that followed orthographically related lists to differ across age groups. Age equivalence in foils that followed orthographically related lists would suggest that any age difference in memory for foils that followed semantically related lists was *not* because of overall level of memory performance but, rather, reflected an age difference in depth of retrieval processing.

If age differences in memory for foils were found, those differences might reflect a functional difference in study time for young and older adults with younger adults perhaps requiring less study time to create a deep-processing set than older adults. In Experiment 2a, the presentation rate of words in the short lists was the same for young and older adults. To examine effects of time for study processing, Experiment 2b included only young adults, and used a much faster presentation rate for study words than was used in Experiment 2a.

Method

Participants

Sixteen young adults and 16 older adults participated in Experiment 2a. The young adults were Washington University undergraduates who received credit for participating in the experiment. Their mean age was 19.1 years (range 18–22 years) and their mean score on the Shipley Vocabulary test was 34.12 (range 29–37). The older adults were recruited from the Washington University older adult subject pool. Their mean age was 71.9 years (range 65–79 years), and their mean score on the Shipley Vocabulary test was 36.12 (range 32–39). Educational attainment for the older adults ranged from a high school diploma ($n = 2$), to a college degree ($n = 8$), to a post-graduate degree ($n = 6$). Sixteen additional young adults, Washington University undergraduates, participated for course credit in Experiment 2b. All participants spoke fluent English and were tested individually.

Materials and design

Stimuli that served as items in the study set were of two types: semantic and orthographic. The semantic study sets were constructed by selecting 136 words from the materials used by McDermott and Watson (2001). Each set contained four semantically related words (e.g., BED, REST, WAKE, and DREAM). The orthographic study sets were constructed from a set of 136 words. These words were selected with the restriction that each word in a set contained the same number of letters and began with the same two letters (e.g., TRUCK, TRAIN, TREND, and TRAMP).

Stimuli that served as foils were 84 words (80 critical, 4 buffers), 4 to 7 letters in length, that were semantically and orthographically *unrelated* to the study sets. Critical items were rotated through 3 conditions: 20 *Deep foils* (foils in a semantic context), 20 *Shallow foils* (foils in an orthographic context), and 40 *New foils* (words that were not presented prior to the test of memory for foils). The assignment of words to conditions was fully counterbalanced across participants. The remaining four words were presented as foils on primacy and recency buffer trials.

Procedure

In Phase 1, a series of study/test trials was presented to the participants as an immediate memory task. Each trial began with a fixation cross in the center of the screen for 1.5 s followed by a prompt in the center of the screen for 1.5 s. This prompt was either the word "meaning" or the word "appearance." Participants were told that this prompt would indicate the nature of the similarity among the study items. Following the prompt, four study words (the study set) were presented one at a time in the center of the screen. Each study word was presented for 1000 ms with a 500 ms inter-stimulus interval (ISI). Immediately after the presentation of the study list, a single word (the test probe) was presented in the center of the screen accompanied by the prompt "old or new?" This word remained on the screen until a response was recorded. Participants were asked to respond "old" if the test item was a word presented in the immediately preceding study set and "new" if it was not. Participants made their responses as key presses. A total of 68 study/test trials was presented during the initial study/test phase (4 primacy, 4 recency, and 60 critical). Of the 60 critical study/test trials, 30 trials consisted of semantic study sets and 30 trials consisted of orthographic study sets. Participants were presented with an old test item on 1/3 of these tests and with a foil on 2/3 of these tests. This design resulted in participants receiving a foil as the test item on 40 of the 60 critical trials (20 in a semantic study context and 20 in an orthographic study context). The presentation of the semantic and orthographic study sets was intermixed and the presentation order was re-randomized for each participant. To avoid primacy and recency effects, four study/test trials were presented at the beginning and end of the initial study/test phase. Among these buffer trials, half of the test items were words presented in the immediately preceding study set (old words), and the other half were foils (new words). All foils were unrelated in meaning and appearance to the study sets.

The final foil recognition test phase included 20 *Deep foils* and 20 *Shallow foils* intermixed with 40 *New foils.* Words were presented one at a time in the center of the screen. Each word remained on the screen until a response was recorded. Participants were told that for this phase, *any* word that had been presented earlier should be considered as an old word, *including all the words that were previously presented as new test words.* Participants responded by pressing one of four colored keys on the computer keyboard, corresponding to "definitely old" (dark red), "probably old" (red), "probably new" (green), and "definitely new" responses (dark green). Participants were instructed to respond: (1) definitely old—only if they were certain that the word had been presented earlier; (2) probably old—if they were unsure, but would guess that the word had been presented earlier; (3) probably new—if they were unsure, but would guess that the word had not been presented earlier; (4) definitely new—if they were certain that the word had not been presented earlier. Participants were able to refer to a response legend presented at the bottom of the computer screen throughout the task. There was no time limit for this test.

Experiment 2b was the same as Experiment 2a except for the following changes: (1) the presentation rate of each study word was reduced from 1500 ms (1000 ms presentation and a 500 ms ISI) to 300 ms (250 ms presentation and a 50 ms ISI), (2) only two response options were accepted in the final phase (old vs. new) instead of four, and (3) only young adults participated.

Results and discussion

Immediate recognition memory

As expected, recognition memory on the immediate tests, corrected for guessing by subtracting false alarms from hits, was near perfect for the deep (Young = .99, Elderly = .98) and the shallow (Young = .98, Elderly = .98) conditions (all F's < 1). Consequently, subsequent differences in memory for foils cannot be attributed to differences in performance during the initial tests or to false alarms, which were rare (<.02).

Final test performance: Memory-for-foils

A summary of performance on the foil recognition test is displayed in Table 3. Because the conditions share the same false alarm rate (.24 and .26 for the young and elderly, respectively), analyses were conducted on the hits for each condition. Examination of overall "old" claims, collapsed across "definitely old" and "probably old" responses, revealed a significant processing type by age interaction, $F(1, 30) = 5.62$, $Mse = .07$. Follow-up tests showed that young adults were significantly more likely to recognize deep foils than shallow foils (.73 vs. .63), $t(15) = 2.72$, $SE = .03$. By contrast, older adults did not show a memory advantage for deep foils, as they demonstrated slightly *poorer* recognition performance for deep foils than for shallow foils (.58 vs. 62), $t < 1$.

A similar pattern emerged when examining only "definitely old" claims, with a significant processing type by age group interaction evident, $F(1, 30) = 4.55$, $Mse = .05$. Follow-up tests revealed that this interaction reflected a significant effect of processing type on memory for foils for the young, $t(15) = 2.28$, $SE = .03$, with the deep foils better recognized than the shallow foils (.51 vs. .42). In contrast, older adults showed poorer recognition performance for deep foils than for shallow foils, (.34 vs. 37), $t < 1$. The false alarm rate for "definitely old" claims was .07 for the young and .11 for the elderly.

Table 3
Probability of responding "old" for the foil recognition test of Experiment 2a

Group	Item type					
	Deep foils			Shallow foils		
	Definite	Probably	Total	Definite	Probably	Total
Young	.51	.22	.73	.42	.21	.63
Elderly	.34	.24	.58	.37	.25	.62

Taken together, age differences in memory for foils from semantically similar lists are notable given that older (.62) and younger (.63) adults did not differ in memory for foils that followed orthographically similar lists. This suggests that differences in memory for foils from the deep lists were not the result of age differences in memory per se but, rather, reflect an age difference in the extent to which retrieval processing was constrained. Young adults engaged in deep retrieval processing when study words were semantically related whereas older adults did not do so.

Experiment 2b

It is unlikely that the age difference in memory for foils was because of a functionally faster rate of study presentation for older adults. The results of Experiment 2b, which used a faster rate of study presentation for young adults, closely replicated those of the young participants in Experiment 2a. Performance was again near ceiling for the immediate recognition-memory test (corrected recognition = .99 for both conditions), $F < 1$. An analysis of recognition memory for foils revealed a significant effect of processing type, $F(1, 15) = 9.60$, $Mse = .08$, with recognition memory for deep foils higher than that for shallow foils (.72 vs. 62). The false alarm rate was .27.

As in Experiment 1, better memory for deep foils for young participants cannot be attributed to differences in rejection time or study time for the foils. Specifically, young participants' rejection time did not differ between the two types of foils in Experiment 2a (deep foils = 748 ms, shallow foils = 730 ms) $p > .30$, or in Experiment 2b (deep foils = 992 ms, shallow foils = 938 ms), $p > .40$. Rejection time for older adults also did not differ between types of foils in Experiment 2a (deep foils = 1743 ms, shallow foils = 1562 ms), $p > .12$.

Experiment 3

Data from Experiments 1 and 2 indicate that older adults are less likely to constrain retrieval in a manner that recapitulates prior, deep processing of targets. As noted previously, this may reflect older adults' dependency on familiarity-based responding for recognition judgments (e.g., Jacoby, 1999). Thus, the distinction between constrained and relatively unconstrained processing at test may correspond to a distinction between recollection and familiarity (cf. Jacoby, Debner, & Hay, 2001). This issue was examined in Experiment 3. Specifically, encoding was varied such that recognition memory for one list of words would depend primarily on familiarity, with recognition memory for a second list largely based on more constrained, deeper retrieval processing that recapitulated processes engaged during study. For the *familiarity* list, a single list of words was read aloud at a fast rate and followed by a test of recognition memory, with this procedure repeated five times. The repeated old words were mixed with different foils (new words) for each of the five tests. Thus, the repeated presentations and tests in combination with new foils were expected to encourage the use of familiarity as a basis for recognition memory. In contrast, words in a second list were presented only once and were not tested in the first phase of the experiment. However, for that single presentation, deep processing (i.e., Craik & Lockhart, 1972) was encouraged by requiring participants to judge the pleasantness of each word during its presentation.

In the second phase of the experiment, participants were given additional recognition tests. For one test, participants were correctly informed that all old words were ones for which they had made a pleasantness judgment whereas for another test they were correctly informed that all old words had been repeatedly presented and tested. Both recognition tests contained foils that were new (i.e., the foils had not been presented earlier in the experiment). Performance for both young and older adults was expected to be near perfect for the familiarity test list in Phase 2, with older adults demonstrating poorer recognition memory for pleasantness-judged, old words than young adults.

Finally, a third phase of the experiment tested recognition memory for *new items* (*foils*) that appeared on the recognition-memory tests given in Phase 2. The deeper processing of foils when pleasantness-judged words were targets was expected to result in better recognition memory for those foils compared to foils from the test list for which recognition judgments were based on familiarity. In contrast to young adults, older adults might rely more heavily on familiarity for recognition of old words in both types of test lists, without further constraining their retrieval processing when pleasantness-judged words were targets. If this were the case, older adults would not exhibit differences in memory for foils from the two types of test lists.

Method

Participants

Twenty-four young (mean age 22.4, range 20–37 years) and 24 older adults (mean age 75.8, range 66–86 years) participated in the experiment. Young adults participated in exchange for course credit. Older adults were recruited from the Washington University community, and received \$10 for their participation. Educational attainment for the older adults ranged from a high school diploma ($n = 12$), to a college degree ($n = 7$), to a post-graduate degree ($n = 5$). All participants spoke fluent English and were tested individually.

Materials

A set of 330 words was used in the experiment. These words were 3–10 letters in length (mean = 5.6) and ranged in frequency from 2 to 25 per million (mean = 12.4) as indexed in the Kucera and Francis (1967) norms. The set of 330 words was divided into 11 matched lists of 30 words each that were rotated through conditions. One of the 11 lists was studied and tested 5 times, with 5 other lists serving as foils for each of the 5 tests. Another list of words was presented only once in the context of a pleasantness judgment task. For the recognition-memory tests given in Phase 2 of the experiment, one of the remaining four lists served as foils for the test of memory for words that had been studied and tested five times, and another list served as foils for the test of memory for pleasantness-judged words. The remaining two lists of words appeared in Phase 3 of the experiment and served as new foils for the test of memory for old foils from the tests given in Phase 2.

Procedure and design

The experiment was conducted in three phases. Throughout, participants sat in front of a computer monitor and viewed words presented via a Power Macintosh computer, making key-press responses on the computer keyboard.

In the first phase of the experiment, participants repeatedly studied and were tested on one set of 30 words. Words were presented at a 1 s rate followed by a 500 ms ISI, and were presented in a different random order for each study presentation. Participants were instructed to read the words aloud in preparation for a memory test. Each study session was immediately followed by a recognition test. During these tests, the 30 words presented for study were randomly intermixed with 30 new foils, that differed for each test. All test stimuli were presented for 2 s, followed by a 500 ms ISI. Participants responded "old" or "new" to each test word by using the index finger of each hand. Mapping of hand (right vs. left) to response was counterbalanced across participants. Participants were instructed to make their old/new judgments as quickly and accurately as possible. This study-test procedure was repeated five times to allow participants to become well-practiced at recognizing the studied words.

Next, participants performed a pleasantness judgment task in which 30 words (not presented elsewhere in the experiment) were presented at a 1-s rate. Participants read each word aloud and then made a pleasantness judgment by stating aloud "pleasant" or "unpleasant." The experimenter keyed in participants' judgments by pressing one of two keys on the computer keyboard. Each key-press initiated the next word-presentation trial. Phase 2 began immediately following the deep encoding task.

In Phase 2 of the experiment, participants were given two recognition-memory tests requiring discrimination between "old" target words and "new" foils (not seen elsewhere in the experiment). Across tests, the nature of old target words was manipulated. For one test, target words were those that had been repeatedly studied and tested. For the other test, target words were those presented in the deep encoding task. Participants were informed about the source of the old words prior to each test. The order of tests was counterbalanced across participants and test items were presented in the same way as for the previously described tests.

In the final phase of the experiment, participants were given a surprise test for foils presented on the recognition tests in Phase 2. During this test, 120 words were presented at participant-controlled durations. These words consisted of 30 foils presented in the test of repeatedly presented words, the 30 foils presented in the test of deeply encoded words, and 60 new foils. As in Experiment 2, participants were given four response options: "definitely old," "probably old," "probably new," and "definitely new." Participants made their responses as key presses.

Results and discussion

In the initial phase of the experiment, participants studied one list of words five times under intentional encoding instructions and were tested following each list presentation. Mean probabilities that participants responded "old" to targets and foils for the five Phase 1 tests are presented in Table 4.

Critical test recognition-memory performance

As shown in Table 5, both young and older participants were more likely to recognize items that had been presented and tested five times (the familiarity condition) than they were to recognize items presented once for pleasantness judgments (the deep condition). However, elderly adults were less likely to correctly recognize old items, particularly those presented in the pleasantness judgment task, and exhibited more false alarms than young adults. This pattern of performance produced a

Table 4
Mean probabilities that participants responded "old" to targets and foils for the five Phase 1 tests of Experiment 3

Test	Probability of responding "old"			
	Young adults		Old adults	
	Item type		Item type	
	Old	New	Old	New
1	.83	.05	.74	.10
2	.97	.02	.92	.07
3	.99	0	.96	.03
4	.98	.01	.98	.03
5	1	0	.97	.02

Table 5
Probability of responding "old" for the critical tests of Experiment 3

Group	Item type			
	Deep		Familiarity	
	Old	New	Old	New
Young	.96	.02	1.00	.01
Elderly	.82	.09	.93	.07

significant three-way interaction of test type, item status, and age group, $F(1, 46) = 4.46$, $Mse = .004$, modifying main effects of test type, $F(1, 46) = 10.66$, $Mse = .003$, and item status, $F(1, 46) = 4552.42$, $Mse = .008$. Two-way interactions of item status and age group, $F(1, 46) = 41.93$, $Mse = .008$, and test type and item status, $F(1, 46) = 27.06$, $Mse = .004$, were also evident.

Final test performance: Memory-for-foils

Because conditions share the same false alarm rate (.19 and .16 for the young and elderly, respectively), analyses were conducted on the proportion of correctly recognized foils from each critical test (Table 6). An analysis of overall "old" claims collapsed across "probably old" and "definitely old" responses resulted in a significant two-way interaction of test type and age group, $F(1, 46) = 12.34$, $Mse = .011$. Specifically, young adults exhibited better memory for foils from the test of pleasantness-judged items than foils from the test of items that had been repeatedly studied and tested, $t(23) = 3.64$. In contrast, older adults' memory for old foils from the two test lists did not differ, $t(23) < 1$.

A similar pattern emerged when only "definitely old" claims were examined. Specifically, a significant two-way interaction of test-type and age group was observed, $F(1, 46) = 24.04$, $Mse = .014$. Follow-up tests indicated that young adults' memory for foils from the test of pleasantness-judged old words was significantly better than their memory for foils from the test of old words that had been repeatedly studied and tested (.76 vs. .57), $t(23) = 4.91$. Conversely, older adults exhibited a trend toward *poorer* memory for foils from the test of pleasantness-judged, old words (.34 vs. .39), $t(23) = -1.67$, $p = .11$. The false alarm rate for "definitely old" claims was .13 for the young, and .11 for the elderly.

Overall, consistent with prior experiments, younger, but not older adults, demonstrated better memory for foils from the test of deeply encoded words. The difference in memory for foils provides evidence that young adults constrained their retrieval processing differently based on the prior processing of target words. When attempting to recognize pleasantness-judged, old words, foils were processed more deeply than when recognition of old words was based on their familiarity. Older adults did not constrain their retrieval processing differently based on the prior processing of target words. Rather, they relied primarily on familiarity as a basis for recognition, regardless of the type of prior processing of the targets.

In contrast to results from Experiments 1 and 2, recognition memory for foils in Experiment 3 was best in the condition that resulted in slower rejection of foils for the initial test (i.e., the deep processing condition). However, differences in memory for foils cannot be explained in terms of differences in rejection time, which corresponds to study time. In particular, the difference in rejection time was as large for older adults (deep foils = 1118 ms, familiarity foils = 967 ms) as for young adults (deep foils = 880 ms, familiarity foils = 738 ms) but only young adults showed an advantage in memory for foils from the pleasantness-judged test. This suggests that, consistent with results from prior experiments, it was an age difference in the depth of retrieval as constrained by the depth of study processing that was responsible for the difference in memory for foils, rather than a difference in the difficulty or amount of time to reject the foils on the prior test.

General discussion

Young adults constrained their retrieval processing differently based on the prior processing of target items, deeply processing both targets and foils at test when target words had been deeply processed at study. Evidence of such deep retrieval processing is revealed by their later memory for foils. Specifically, memory for foils was higher when pleasantness-judged old words were targets in the earlier test rather than words that had been shallowly processed in a vowel-judgment task (Experiment 1) or repeatedly studied and tested (Experiment 3). Manipulating the similarity among words in study lists

Table 6
Probability of responding "old" for the final test of Experiment 3

Group	Item type					
	Deep			Familiarity		
	Definite	Probably	Total	Definite	Probably	Total
Young	.76	.13	.89	.57	.20	.77
Elderly	.34	.26	.60	.39	.24	.63

(Experiment 2) also produced effects of depth of retrieval processes on memory for foils. For older adults, none of the experiments revealed a significant effect of depth of processing of targets on later recognition memory of foils. Such effects on memory for foils illustrate qualitative differences in bases for recognition memory that are consistent with encoding specificity and transfer-appropriate processing (e.g., Morris et al., 1977; Tulving & Thomson, 1973). That is, cue elaboration at test served to constrain memory retrieval, and influenced the test encoding of both old and new words on the recognition test, resulting in better memory for foils when targets had been deeply processed.

Constraining retrieval and models of memory

Similar to arguments made here, Jacoby et al., (1989; see also Kelly and Rhodes, 2002) discussed the notion of constrained memory retrieval in the context of their description of relative fluency as a basis for memory performance. They likened the rememberer to an intuitive scientist who seeks transfer from a prior event as evidence that the event occurred. While transfer from one experience to another could provide a cue that an event was part of past experience, the transfer must be specific to be diagnostic of past experience. A failure to elaborate on the cues provided at the time of test would restrict the opportunity for transfer to be experienced on various levels—conceptual as well as perceptual. Thus, remembering is a process of using cues at increasing levels of exactness or constraint (cf. Burgess & Shallice, 1996) in order to limit irrelevant sources of information on performance, and allow the more accurate attribution of familiarity. In this sense, familiarity can arise at various levels of constrained memory retrieval or specification.

Such an idea of increasing levels of constraint can serve to clarify the distinction between recollection and familiarity as bases for recognition memory. Dual-process models of recognition memory (e.g., Jacoby, 1991; Mandler, 1980) hold that recollection and familiarity serve as alternative bases for recognition memory (for a review, see Yonelinas et al., 2002). Recollection can be described as relying on a relatively effortful, attention-demanding form of memory access that is constrained by the goal to retrieve a particular episode and recapitulate study processing. Familiarity is a less constrained, more automatic basis for recognition. This is consistent with findings from the current study demonstrating that older adults, who rely more heavily on familiarity than younger adults (e.g., Jacoby, 1999), were less likely to engage in deep retrieval processing. Thus, the distinction between recollection and familiarity may be conceptualized as a distinction between degrees of constraint at retrieval, constraint that older adults have difficulty achieving (see Jacoby et al., 2001, for a similar account of age differences in cued-recall performance).

Source memory

Older adults' deficits in source identification (for a review, see Spencer & Raz, 1995) may reflect a lack of flexibility in constraining retrieval processing. Source identification is typically tested by asking participants to decide whether a test item originated from one of several sources (e.g., read vs. heard) or was not presented previously (for a review, see Johnson, Hashtroudi, & Lindsay, 1993), and is generally assumed to follow recognition memory, which is presumably based on familiarity (e.g., Bayen, Murname, & Erdfelder, 1996). By that view, age differences in source identification result from older adults' lessened ability to *monitor* the source of recognized items. In contrast, our results suggest that differences in source memory might reflect qualitative differences between young and older adults in their basis for recognition memory per se. That is, rather than source memory *following* recognition, these data demonstrate that young adults constrain their retrieval processing in a manner that is consistent with the source (i.e., the prior processing) of a target item (see also Jacoby et al., in press), whereas older adults are less likely to do so.

Global memory models

In global activation models of memory (e.g., Gillund & Shiffrin, 1984), recognition is accomplished by comparing a memory probe's strength (familiarity) against a decision criterion. If the probe's value exceeds criterion, it is accepted as "old"; otherwise, it is rejected as "new." However, if participants simply assess global familiarity when making recognition judgments (e.g., Gillund & Shiffrin, 1984), there would be no reason to expect differential processing of the foils depending on the study processing of targets, and, consequently, no reason to expect the differences in subsequent memory for foils that we observed.

Humphreys et al. (2003) have shown the importance of test instructions for memory access and have argued that memory decisions are based on a match between a reinstated context, including processing context, and a context that is retrieved using the probe as a cue (also see Marsh & Hicks, 1998). Global memory models also retain information about context that can be used to influence recognition-memory performance (for a review, see Clark & Gronlund, 1996). However, Hockley and Cristi (1996) have provided impressive results demonstrating that current global memory models are unable to account for people's ability to constrain memory access for item and associative information to events from a specified source.

Data of this sort highlight what, by our view, comprises the primary weakness of current theorizing about recognition memory. That is, recognition is construed as a matching process without sufficient importance being given to retrieval processing—cue elaboration of a sort that constrains memory access. Describing context as a tag attached to a memory (Anderson & Bower, 1973) or as an entry in a vector (e.g., Hintzman, 1988) does not capture the recapitulation of study processing that is responsible for the effects on memory for foils that we report. Whereas global memory models emphasize *quantitative* differences in strength or familiarity, we emphasize *qualitative* differences in the form of information upon which recognition is based, differences that constrain retrieval processing and are evident in subsequent memory for foils.

Neural evidence of differences in bases for recognition

Consistent with our findings of differences in depth of retrieval, Rugg, Allan, and Birch (2000) found differences in event-related potentials for new words that depended on whether old words in a recognition test had been studied in a deep or shallow encoding task. They described their results as arising from differences in retrieval orientation that reflect the nature of retrieval cues used or the kinds of memory representations that are accessed.

In event-related fMRI studies, we (Velanova et al., 2003; Velanova, Lustig, Jacoby, & Buckner, 2004) used familiarity and deep-processing conditions, similar to those used in Experiment 3 here, to explore the neural correlates of controlled retrieval in young and older adults. Results revealed activations in frontal control regions that suggest that young adults engage controlled processes early on to constrain retrieval. In contrast, older adults engage regions implicated in control later in the processing stream, suggesting a fundamental shift in the use of controlled processes across the lifespan. Older adults might be less likely to gain control of memory by means of constraining memory access than are young adults and, instead, more often rely on post-access monitoring, a less efficient means of cognitive control.

Constrained memory retrieval and metacognition

An emphasis on qualitative differences in memory retrieval may be useful for understanding metacognition. By our view, current theorizing overemphasizes the importance of post-access monitoring processes. Undoubtedly, monitoring processes of the sort involved in using confidence judgments to control the accuracy of memory performance (e.g., Koriat & Goldsmith, 1996) do serve a role. However, we suggest that controlling memory access by constraining retrieval processes is a more efficient and, perhaps a more common means of cognitive control. Even when differences in monitoring are found, the important difference might be in *what* is monitored (cf. Kelley & Sahakyan, 2003), resulting from qualitative differences in memory access, rather than differences in monitoring processes per se. What is needed are methods of separating the contributions of memory access and post-access monitoring as means of gaining cognitive control.

Our strategy of examining memory for foils provides a way to gain evidence that differences in goal-constrained memory access contribute to age differences in cognitive control. The contrast between goal-constrained retrieval (early selection), and post-retrieval monitoring (late correction) is a general one that likely applies to forms of cognitive control in social settings as well as to the control of memory (Jacoby, Kelley, & McElree, 1999). Our interest in separating the contributions of different means of cognitive control is in the service of a larger aim: the development of procedures to enhance the memory performance of older adults. Just as attempts to improve quality control require questioning whether the primary problem is one of production procedures or of quality-control inspectors, attempts to repair memory performance best begin by devising ways to distinguish among different deficits in cognitive control. A deficit in the ability to constrain memory access would likely require rehabilitation procedures that are very different from those required to rehabilitate a deficit in post-access monitoring.

References

Anderson, J. R., & Bower, G. H. (1973). *Human associative memory*. Oxford, England: V.H. Winston & Sons.

Bayen, U. J., Murname, K., & Erdfelder, E. (1996). Source discrimination, item detection, and multinomial models of source monitoring. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 22*, 197–215.

Burgess, P. W., & Shallice, T. (1996). Confabulation and the control of recollection. *Memory, 4*, 359–411.

Clark, S. E., & Gronlund, S. D. (1996). Global matching models of recognition memory: How the models match the data. *Psychonomic Bulletin & Review, 3*, 37–60.

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.

Dunlosky, J., & Nelson, T. O. (1992). Importance of the kind of cue for judgments of learning (JOL) and the delayed-JOL effect. *Memory & Cognition, 20*, 373–380.

Fernandez-Duque, D., Baird, J. A., & Posner, M. I. (2000). Executive attention and metacognitive regulation. *Consciousness and Cognition, 9*, 288–307.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91*(1), 1–67.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace model. *Psychological Review, 93*, 411–428.

Hockley, W. E., & Cristi, C. (1996). Tests of the separate retrieval of item and associative information using a frequency-judgment task. *Memory & Cognition, 24*, 796–811.

Humphreys, M. S., Dennis, S., Maguire, A. M., Reynolds, K., Bolland, S. W., & Hughes, J. D. (2003). What you get out of memory depends on the question you ask. *Journal of Experimental Psychology: Learning, Memory & Cognition, 29*, 797–812.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language, 30*, 513–541.

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

Jacoby, L. L., Debner, J. A., & Hay, J. F. (2001). Proactive interference, accessibility bias, and process dissociations: Valid subjective reports of memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 686–700.

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

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

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

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

Kelley, C. M., & Rhodes, M. G. (2002). Making sense and nonsense of experience: Attributions in memory and judgment. In B. Ross (Ed.), *The psychology of learning and motivation* (pp. 293–320). New York: Academic Press.

Kelley, C. M., & Sahakyan, L. (2003). Memory, monitoring, and control in the attainment of memory accuracy. *Journal of Memory and Language, 48*, 704–721.

Koriat, A., Ben-Zur, H., & Sheffer, D. (1988). Telling the same story twice: Output monitoring and age. *Journal of Memory and Language, 27*, 23–39.

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review, 103*, 490–517.

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

Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review, 87*, 252–271.

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

McDermott, K. B., & Watson, J. M. (2001). The rise and fall of false recall: The impact of presentation duration. *Journal of Memory and Language, 45*, 160–176.

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning & Verbal Behavior, 16*(5), 519–533.

Nelson, T. O. (1996). Consciousness and metacognition. *American Psychologist, 51*, 102–116.

Rugg, M. D., Allan, K., & Birch, C. S. (2000). Electrophysiological evidence for the modulation of retrieval orientation by depth of study processing. *Journal of Cognitive Neuroscience, 12*, 664–678.

Shimizu, Y., & Jacoby, L.L. (in press). Similarity-guided, depth of retrieval: Constraining at the front end. *Canadian Journal of Experimental Psychology*.

Spencer, W. D., & Raz, N. (1995). Differential effects of aging on memory for content and context: A meta-analysis. *Psychology & Aging, 10*, 527–539.

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review, 80*, 352–373.

Velanova, K., Jacoby, L. L., Wheeler, M. E., McAvoy, M. P., Peterson, S. E., & Buckner, R. L. (2003). Functional-anatomic correlates of sustained and transient processing components engaged during controlled retrieval. *Journal of Neuroscience, 17*, 8560.

Velanova, K., Lustig, C., Jacoby, L. L., & Buckner, R. L. (2004). Evidence for frontal processes: Processing difference during controlled memory retrieval in older adults, submitted.

Yonelinas, A. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of memory and Language, 46*(1), 441–517.