The effect of forced recall on illusory recollection in younger and older adults

MICHELLE L. MEADE
University of Illinois at Urbana–Champaign

HENRY L. ROEDIGER III
Washington University at St. Louis

The effect of an initial forced recall test on later recall and recognition tests was examined in younger and older adults. Subjects were presented with categorized word lists and given an initial test under standard cued recall instructions (with a warning against guessing) or forced recall instructions (that required guessing); subjects were later given a cued recall test for the original list items. In 2 experiments, initial forced recall resulted in higher levels of illusory memories on subsequent tests (relative to initial cued recall), especially for older adults. Older adults were more likely to say they remembered rather than knew that forced guesses had occurred in the original study episode. The effect persisted despite a strong warning against making errors in Experiment 2. When a source monitoring test was given, older adults had more difficulty than younger adults in identifying the source of items they had originally produced as guesses. If conditions encourage subjects to guess on a first memory test, they are likely to recollect these guesses as actual memories on later tests. This effect is exaggerated in older adults, probably because of their greater source monitoring difficulties. Both dual process and source monitoring theories provide insight into these findings.

The experiments reported in this article examine a situation that often occurs when one repeatedly recalls the same events. If a person witnesses a series of events and then recalls them with encouragement to guess and to produce as many responses as possible, what effect does this have on later recall? Might people recall at a later time items that had been produced as guesses on the earlier test? Furthermore, will they confidently attribute the items they produced as guesses as having actually been witnessed? If the answer to these three questions is "yes," we can further ask whether older adults may be more powerfully affected by guessing than younger adults. That is, will older adults be even more likely to remember their own erroneous responses from a first test as confident memories on a later test? The two experiments reported here show that the answers to all of these questions are positive. We account for the results in terms of Johnson's

AMERICAN JOURNAL OF PSYCHOLOGY
Fall 2006, Vol. 119, No. 3, pp. 433–462

source monitoring theory (Johnson, Hashtroudi, & Lindsay, 1993) and Jacoby's attributional/bias theory (Jacoby, Dolan, & Marsh, 2001).

The effects of guessing can be brought under experimental control through forced recall. Forced recall requires subjects to produce a given number of responses on a memory test beyond what they can actually recall (Erdelyi & Becker, 1974). Erdelyi and Becker used forced recall to equate response criterion across repeated tests, so subjects were required to produce the same large number of items on each test. They were interested to see whether hypermnesia (increased recall with repeated testing) could be obtained when the total number of responses was held constant across tests. According to generate and recognize theories of recall (Anderson & Bower, 1972, among many), forcing people to guess is expected to increase the number of items correctly produced, even if at the expense of intrusions. However, when Roediger and Payne (1985) compared forced recall to free recall with a warning against guessing after subjects studied lists of unrelated words, they found no difference in the number of items correctly produced. They concluded that guessing with these types of materials does not necessarily lead to production of more target items. This finding has generally been replicated, although results indicate that forced recall can increase accuracy over standard free recall instructions when the materials tested are from an easily guessable set (see Erdelyi, Finks, & Feigin-Pfau, 1989; Ritter & Buschke, 1974) and when the data are not corrected for guessing (Roediger, Srinivas, & Waddill, 1989). However, with unrelated materials, assessing memory with forced rather than free recall typically does not change the number of list items produced, even though it does increase the number of intrusions.

The aforementioned studies all used the criterion that if the subject wrote down an item that had been presented on the list, it was counted as correct recall. This practice seems reasonable under some types of retrieval instructions (e.g., free recall), but when subjects are forced to guess, it is questionable because items that are counted as correct by the experimenter might have been, from the subject's point of view, sheer guesses. Roediger, Wheeler, and Rajaram (1993) included subjects' own assessments of the accuracy on free and forced recall tests by having them, after a test of each type, rate the produced items on a scale ranging from 1 (*sure the item was not on the list*) to 6 (*sure the item was on the list*), with intervening values representing intermediate levels of confidence for positive (4, 5, or 6) or negative (3, 2, or 1) judgments. Replicating the results of earlier research, Roediger et al. (1993) found that the number of list items produced did not differ between free and forced recall tests, even though many more intrusions occurred on the forced recall test. However, when they analyzed subjects' own judgments of what they remembered based on the confidence ratings, a different pattern of results was obtained.

Relative to those in free recall, forced recall subjects were much more likely to produce list items and not realize they were correct on the later confidence test. Therefore, recognition failure of recallable words was obtained under conditions in which there were no external retrieval cues, as in the standard paradigm for studying this phenomenon (Tulving & Thomson, 1973). More interestingly for present purposes, when subjects produced items under conditions of forced recall, they made many false recognition errors, attributing the guessed items to having actually been studied in the list. These false memories produced after forced recall represent the point of departure for the present research.

The research by Roediger et al. (1993) questions the accuracy of subjects' recall when they are encouraged or even forced to guess, especially on a later test. Of course, this condition often prevails in memory testing outside the laboratory, as when a police officer asks a witness to try to recollect everything about a possible crime scene, guessing if necessary, or when a therapist asks someone struggling to recall childhood events to let his or her mind roam freely and to produce related information. Retrieval under hypnosis typically uses instructions that encourage people to guess; it is no surprise that courtroom testimony from previously hypnotized witnesses is banned (American Medical Association, 1986). The findings from the Roediger et al. (1993) study suggest that hypnosis is not critical to the act of guessing contaminating later recall; guessing in a normal state of consciousness has the similar effect of undermining later recollections.

Research from other paradigms confirms that forcing people to produce intrusions on a forced recall test inflates the incidence of false memories on later tests. For example, Ackil and Zaragoza (1998) presented younger children and college-aged adults with a short video. Subjects were later forced to produce answers to questions about events that never happened in the video. After a 1-week delay, subjects were tested again on their memory of the video. Forcing subjects to produce answers to questions about events that did not happen increased the probability that subjects would later report that they remembered that the event had actually occurred. The effect was especially pronounced in younger children. Later research suggested that the false memories resulting from the errors produced on a forced recall test were greater still when subjects were given confirmatory feedback by the experimenter on the initial test (Zaragoza, Payment, Ackil, & Beck, 2001).

The effect of forced recall on false memory has been examined in other paradigms as well. Roediger, Jacoby, and McDermott (1996) found that subjects in a standard Loftus eyewitness situation (Loftus, Miller, & Burns, 1978) were more likely to misremember events from a slide sequence under conditions in which they had been encouraged to produce incorrect information on a previous test. Prior production of erroneous information also increased subjects' reports that they remembered the erroneous details as having been presented (also see Schooler, Foster, & Loftus, 1988). Henkel (2004) demonstrated that subjects were more likely to misattribute the source of previously seen or previously imagined pictures when they were tested on a forced recall test than on a free recall test. Finally, similar results with forced recall have been shown using the Deese–Roediger–McDermott (DRM; Deese, 1959; Roediger & McDermott, 1995) associative memory paradigm by McKelvie (1999, 2001).

In the literature reviewed, forced recall has consistently increased the level of false recall relative to free recall, whereas the level of veridical recall remains unchanged. The current study examined forced recall and false memories on categorized lists using a paradigm similar to that of Smith, Ward, Tindell, Sifonis, and Wilkenfeld (2000). This paradigm permits testing of more critical items than in the standard DRM paradigm. We predicted that forced recall would produce errors on a recall test given later and that source monitoring tests would show that items erroneously produced during forced recall would later be misattributed to the study presentation. We were also interested in the effects of aging on forced recall because older adults have been shown to have greater source monitoring difficulties than younger adults (McIntyre & Craik, 1987) and therefore are more likely show elevated levels of false recall in some paradigms (e.g., Norman & Schacter, 1997; Balota et al., 1999). Previous demonstrations that older adults' source monitoring deficits are exaggerated with highly similar items (Henkel, Johnson, & De Leonardis, 1998) lend further support to our hypothesis that the current categorized list paradigm would result in heightened levels of false memory for older adults. We first review recent work on effects of varying response criteria on recall and relevant research about aging before describing our own experiments.

Kelley and Sahakyan (2003) provided evidence that older adults show a lower correspondence between confidence ratings (subjective judgments) and accuracy of items (objective correctness) than younger adults (also see Koriat & Goldsmith, 1996; Mitchell, Johnson, & Mather, 2003). They argued that older adults' deficits in monitoring processes stem from their inability to recollect specific details of studied episodes, an idea consistent with Jacoby's dual process theory (Jacoby, 1991). Kelley and Sahakyan further suggested that older adults may be more likely than younger adults to base their memory decisions on the plausibility that an item occurred or its familiarity (Koustaal & Schacter, 1997; Reder, Wible, & Martin, 1986) or on the general ease with which an item is generated (Jacoby & Hollingshead, 1990; see Kelley & Rhodes, 2002, for a review). Judgments of plausibility and fluency may be especially important in recall of categorized lists, as in the current research, because common exemplars can be generated easily. Jacoby (1999) has shown that older adults are more likely than younger adults to misattribute highly accessible items (those easily brought to mind) to memory. Because producing an item on a forced recall test increases the item's accessibility, we predicted that older adults would be more likely than younger adults to misattribute guesses on a forced recall test to items having been previously presented.

Furthermore, the issue of forced recall for older adults is interesting because of older adults' possible deficit in source monitoring abilities (e.g., Hashtroudi, Johnson, & Chrosniak, 1989; McIntyre & Craik, 1987). Source monitoring is subjects' ability to determine where or how they acquired information (Johnson et al., 1993). Source monitoring deficits may occur when subjects remember information but cannot recollect the source of the information on later tests. After tests of forced recall, older adults should have more difficulty than younger adults in distinguishing the source of items as having been produced on a forced recall test or as having occurred in the study list.

Past research supports the hypothesis that forcing subjects to respond may have a greater effect on older adults than on younger adults (Hashtroudi et al., 1989; Jacoby, 1999; McIntyre & Craik, 1987). The current research is intended to test this prediction in forced recall of categorized lists in which subjects must generate an entire list of possible items that might have been presented at study. We ask whether forced recall effects are different for younger and older adults and, in addition, whether the recollection of items erroneously produced on a forced recall test will differ for younger and older adults on later memory tests. As previously discussed, younger adults have been shown to misattribute items produced during forced recall to having actually occurred in the study episode (Roediger et al., 1993), and we hypothesize that this tendency will be even greater in older adults because of their impaired monitoring abilities (Kelley & Sahakyan, 2003; McIntyre & Craik, 1987) and greater tendency to show accessibility bias (Jacoby, 1999). Two experiments were designed to determine whether false memory effects would occur in a categorized list paradigm under free or forced recall instructions and whether there would be an age difference in the effects of forced recall.

We also included a final source monitoring test to determine whether subjects could accurately identify the source of their guesses and memories. We predicted older adults would have more difficulty distinguishing items produced as guesses from items produced as memories after a forced recall test because of their poorer source monitoring abilities.

EXPERIMENT 1

Experiment 1 asked whether forced recall would produce higher levels of false memories in older than younger adults. Subjects recalled lists on a first test under cued recall or forced recall. We assessed false recall on a second cued recall test in which subjects attempted to recall the initial list, with instructions to be as accurate as possible and not to guess. Finally, subjects received a source monitoring recognition test.

METHOD

Subjects

Younger adult subjects were 20 Washington University undergraduates who participated in the experiment for partial fulfillment of a class requirement or for $10 cash (age range 18–25, mean 19). The older adults were 20 people (age range 67–88, mean 77) recruited from the Washington University Older Adult Subject Pool. Older adults received $15 cash for their participation. Older adults were paid more than younger adults because they were slower at the task and therefore took longer (the rate was $10 per hour). Although Shipley Vocabulary scores were not collected for the subjects of Experiment 1, both younger and older adult subjects were drawn from the same population of subjects tested in other experiments in the lab where the mean Shipley for younger adults is 33 (range 27–37) and the comparable score for older adults is 35 (range 26–39).

Design

The experiment consisted of a 2×2 between-subject design. Retrieval condition (forced recall or cued recall) on the first test and age (younger or older adults) were both manipulated between subjects. The primary dependent variables were correct and false recall on the second test and correct and false recognition of the critical items on the source monitoring test.

Materials

Materials included six lists constructed from the Battig and Montague (1969) category norms (birds, human body parts, vegetables, four-footed animals, articles of clothing, and flowers). The first 22 exemplars in each category were selected, and then each list was constructed to contain exemplars numbered 6–22, for a total of 17 items (outdated or repetitive words in the norms were removed and replaced with the next word). The first five exemplars of each category were excluded from the lists and used later as the critical items on recall tests and the final recognition source monitoring test. These items are the most typical members of the category and hence most likely to be intruded. The recognition source monitoring test consisted of 90 items (the five critical items from each of the six study lists, five randomly selected studied items from each study list, and 30 unrelated filler items). The six 17-item study lists and the five critical missing items appear in the Appendix.

Procedure

Each subject was tested alone. The six categorized word lists were presented visually on a computer screen in the following order: birds, body parts, vegetables, four-footed animals, articles of clothing, and flowers. Subjects were instructed to pay careful attention to each item and told they would later be tested on their memory for each list. List items were presented for 1.5 s each in the center of the screen. At the completion of each list, the computer prompted the subject to press the "Enter" key to continue. The screen then displayed the title "Next List," and the experimenter verbally labeled the upcoming list (e.g., the bird list, the human body part list). After seeing all six lists, subjects were given a filler task (visuospatial puzzles) before recall to eliminate short-term memory effects.

Subjects then participated in a test phase in which they recalled aloud to the experimenter items from each list. The experimenter provided the verbal label for each list, and the subjects then recalled items from that list. Subjects in the cued recall condition were asked to recall only the items from the lists and were told not to guess. When subjects in the cued recall condition had recalled all items they thought possible, they were allowed to move on to the next list, regardless of the number of items produced. Subjects in the forced recall condition were required to produce 20 items from each list (only 17 had actually been presented). One young subject in the forced recall condition commented that she had counted only 17 words and would not produce 20, so her data were replaced with those of an additional subject. No other subject mentioned the disparity in number of items.

After this initial recall, subjects were asked to complete a written recall test. All subjects were instructed to be accurate on this recall test and were told not to guess. The category names were again presented in the same order on individual sheets, and subjects had 3 min per category to write down as many items as possible. In addition to writing the items, subjects were asked to make "remember" and "know" judgments for each of their responses. The instructions given for these judgments were based on those provided by Tulving (1985), Gardiner (1988), and Rajaram (1993). "Remember" responses indicated a specific recollection about the item, whereas "know" responses were given when subjects had no specific recollection of the item's occurrence but believed that the item had been in the list. The exact instructions read as follows: "After each word, please put an "R" or a "K". *R* stands for *remember* and *K* stands for *know*. Remembering something implies you recollect that the item actually occurred in the lists. You remember having thought about it or noted something about the way it looked—you actually remember having seen the item in the list. A *know* response means that you do not actually recollect the occurrence of the item (you remember nothing specific about having seen it), but you nonetheless know that it was there." The instructions were discussed until each subject indicated that he or she understood the distinction.

Finally, subjects were given the 90-item recognition source monitoring test. For each item on the test, subjects were asked to indicate whether the item had appeared on the study list (list only), whether they said it aloud during the initial recall (self only), whether it was in the list and they had said it (both list and self), or whether the item was not on the list and the subject had not said it (neither). Subjects had these labels on their recognition test and repeated back instructions to show they understood the task. Finally, subjects were thanked for their participation and fully debriefed.

RESULTS AND DISCUSSION

Recall Test 1

Table 1 shows the mean proportions of list items and critical items produced in the first recall test. Cued and forced recall were analyzed separately because the two tasks were fundamentally different: In cued recall subjects wrote what they remembered, and in forced recall subjects produced both remembered items and guesses. We turn first to the analyses done for the list items produced. Statistical significance is set at $p \leq .05$ unless otherwise noted.

An independent samples *t* test between younger and older adults computed for the proportion of list items produced on the cued recall test revealed no main effect of age on recall of list items, $t < 1.0$, with older adults correctly recalling a similar proportion of items as younger adults. This null outcome is not unusual in experiments with strong retrieval cues or external retrieval support (e.g., Craik, 1983; Craik & McDowd, 1987). A benefit of younger and older adults recalling similar numbers of list items is that it gave us an equal baseline from which to examine relative changes in memory performance on Test 2. In contrast to cued recall, the pattern of results obtained from a separate *t* test computed for forced recall revealed a significant age difference, $t(18) = 2.99$, $SEM = .02$, with younger adults producing significantly more list items than older adults.

Turning to production of the five critical items on each list, additional *t* tests were conducted separately for recall of critical items on the cued recall test and the forced recall test. Under cued recall instructions, we were most interested in whether older adults were more likely to produce critical items. Although the difference is not statistically significant, $t(18) = 1.78$, $SEM = .07$, there is a numerical trend in the direction of older adults recalling more critical items than younger adults.

Examining next the critical items produced on the forced recall test, we found no age difference in the proportion of critical items produced by younger and older adults in the forced recall condition, $t(18) < 1$. Of course, both younger and older adults were instructed to guess in this

Table 1. Mean proportion of items recalled by younger and older adults on a cued recall or forced recall test (Experiment 1, $N = 40$)

	Cued recall		Forced recall	
	Younger	Older	Younger	Older
List recall	.39	.39	.65	.59
Critical recall	.09	.22	.70	.69

condition, so equivalent production of critical items on the first recall test was not surprising and aids interpretation of data from later tests.

Recall Test 2

Table 2 presents the mean proportion of veridical and false recall on the second recall test, along with the proportion of items given "remember" and "know" responses. The second recall test was always taken under cued recall instructions, with subjects told to avoid guessing and to report only words they had seen on the computer screen. The labels on the table refer to the subjects' condition on the first recall test, and our interest is in seeing what effects carry over to the second test. We first discuss the proportion of list and critical items recalled and then turn to analyses of "remember" and "know" judgments.

A 2 (prior cued or prior forced recall) × 2 (younger or older adults) ANOVA on correctly recalled items showed that subjects who had recalled under forced recall conditions recalled approximately the same proportion of list items (.51) as those who had recalled under cued recall conditions (.46), $F(1, 36) = .2.33$, $MSE = .02$. Furthermore, there was no main effect of age, $F(1, 36) = .47$, $MSE = .02$, suggesting that older and younger adults were comparable in their cued recall of list items, as in the first test. However, this age-equivalent performance on list items contrasts with the age differences found in recall of critical items.

Consistent with our hypothesis that prior forced recall would increase error rates on subsequent tests, after forced recall subjects produced .46 critical items relative to .23 when the first test had been cued recall, $F(1, 36) = 13.54$, $MSE = .04$. In addition, older adults were more likely to produce the critical items, $M = .44$, than were younger adults, $M = .25$, as

Table 2. Mean proportion of items recalled and mean proportion of remember and know responses on subsequent individual cued recall test (Experiment 1, $N = 40$)

	Prior cued recall		Prior forced recall	
	Younger	Older	Younger	Older
	List recall			
Total	.48	.43	.51	.51
Remember	.39	.34	.40	.37
Know	.09	.09	.12	.14
	Critical recall			
Total	.17	.28	.32	.60
Remember	.06	.16	.04	.33
Know	.11	.12	.28	.27

revealed by a significant main effect of age, $F(1, 36) = 9.32$, $MSE = .04$. Again, we found a greater tendency for false recall in older than in younger adults. This age effect is especially interesting in light of the fact that younger and older adults produced similar proportions of critical items on the first recall test.

Stated differently, after an initial forced recall test, younger adults were better at reducing the level of false recall on the second recall test than were older adults. This outcome may suggest that younger adults were better able to monitor which of the previously recalled items had actually been presented in the list and which items they themselves had generated, an idea consistent with the source monitoring framework (Johnson et al., 1993).

"Remember" and "know" responses

Subjects gave "remember" and "know" judgments for each item recalled on the second recall test, with the mean proportions given in Table 2. The "remember" and "know" responses sum to the overall number of responses, with any inconsistency caused by rounding error.

We first consider "remember" responses for correctly recalled items. A 2 (younger or older adults) × 2 (prior cued or forced recall) ANOVA revealed no significant differences in "remember" responses for any of the factors, Fs < 1. "Know" responses for list items were analyzed in a similar manner, and again there were no differences for any of the factors, Fs < 2.4. Subjects were just as likely to say they remembered or they knew a list item regardless of experimental condition.

The situation was quite different for critical items. "Remember" and "know" responses were also analyzed for critical items using separate 2 (prior cued or forced recall) × 2 (younger or older adults) ANOVA. Interestingly, "remember" responses for the critical items in older adults were higher than for younger adults, indicating greater false remembering for older adults, $F(1, 36) = 13.32$, $MSE = .03$. No other effects were significant. Similar analyses on the "know" responses revealed no main effects of age, but we did find a significant main effect of retrieval condition in Test 1, $F(1, 36) = 10.10$, $MSE = .03$, indicating that subjects were more likely to give "know" responses for critical items when they had previously recalled under forced recall instructions than under cued recall instructions.

Recognition source monitoring test

Subjects' performance on the final recognition source monitoring test is presented in Table 3. The condition labels ("prior cued recall" and "prior forced recall") refer to the condition in which subjects recalled the list items in the first recall test. Of course, prior testing may confound recognition performance in the current experiment (Gallo & Roediger,

Table 3. Mean proportion of veridical recognition for list items and false recognition for critical items (Experiment 1, $N = 40$)

	Prior cued recall		Prior forced recall	
Source judgment	Younger	Older	Younger	Older
Veridical recognition of list items				
List only	.42	.37	.29	.29
Both list and self	.46	.45	.58	.62
Total correct recognition	.88	.82	.87	.91
Self	.02	.06	.06	.04
Neither list nor self	.10	.12	.07	.05
False recognition of critical items				
List only	.43	.35	.16	.17
Both list and self	.16	.28	.37	.66
Total false recognition	.59	.63	.53	.83
Self only	.01	.05	.33	.10
Neither list nor self	.40	.33	.13	.07

2002; Roediger et al., 1996), but the variable of interest is the effect of prior test conditions on later memory performance. Data in Table 3 are grouped according to the various response options for list items and for critical items. We define the total correct recognition as the proportion of occasions on which subjects correctly attributed an item as having occurred in the list ("list only" responses plus "both list and self" responses). One older adult did not understand the test instructions and therefore was eliminated from the analyses.

For correct recognition, a 2 (prior cued or forced recall) × 2 (younger or older adults) ANOVA revealed no main effect of retrieval condition of Test 1, $F(1, 36) = 1.14$, $MSE = .02$, indicating that subjects who had initially recalled the list items under forced recall instructions, $M = .89$ collapsing across age, correctly recognized approximately equal numbers of list items as did subjects who previously recalled the list under cued recall instructions, $M = .85$. Finally, no age differences were found for correct recognition, $F(1, 36) = .03$, $MSE = .02$, suggesting that older adults, $M = .87$ collapsing across retrieval condition, and younger adults, $M = .88$, correctly recognized approximately equal proportions of list items. This finding of equivalent recognition performance for younger and older adults is especially interesting in light of their differing performance for false recognition of critical items, which we discuss next.

False recognition is defined as recognition of items that were not present in the original list but may have been produced during previous recalls (or "list only" responses plus "both list and self" responses). We found that older adults, $M = .73$ collapsing across retrieval condition, falsely recognized significantly more critical items than did younger adults, *M* = .56, $F(1, 36) = 3.96$, *MSE* = .07. As suggested previously, this age difference in false recognition seems to be driven mostly by older adults' inability to distinguish between items produced by themselves to fulfill requirements of the forced recall task and items they produced that actually were on the list. Evidence for this hypothesis comes from the differing proportions of older and younger adults' responses in each of the source judgment categories. Specifically, younger adults who had previously recalled under forced recall instructions were able to correctly attribute 33% of items ("self only") as having been produced on the first test even though they had not appeared on the list. In contrast, older adults thought that only 10% of the items falsely produced in the forced recall condition had not actually been in the list. For older adults, if the item had been produced at all during the initial forced recall phase, they were more likely to say that it had also appeared in the list ("both list and self"; *M* = .66 for older adults and *M* = .37 for younger adults).

There was no main effect of prior retrieval condition on Test 1, $F(1, 36) = .86$, *MSE* = .07, with roughly equivalent levels of false recognition for subjects who had previously recalled the lists under forced recall instructions, *M* = .68, than cued recall instructions, *M* = .61.

EXPERIMENT 2

One of the most interesting findings of forced recall in Experiment 1 was its exaggerated effect on older adults' false recall and recognition. Results on the final source monitoring test indicated older adults' greater tendency to believe that most items they produced earlier on the forced recall test had occurred on the list when they in fact had not. One issue addressed in Experiment 2 was whether older and younger adults can reduce their false recollections from prior recall when they receive an explicit warning about the nature of the lists and the possible effects of forced recall.

Giving subjects an explicit warning against false memories may especially help to decrease older adults' false memory reports. Warnings to subjects have reduced the incidence of false recollections in a variety of paradigms without eliminating them entirely (Gallo, Roberts, & Seamon, 1997; Gallo, Roediger, & McDermott, 2001; McDermott & Roediger, 1998; Meade & Roediger, 2002; Wright, 1993), and our interest was in seeing whether similar effects could be obtained after forced recall. Also, the effect of warning may vary with age; the results of Experiment 1 suggest that younger adults are better able to determine the source of the items and therefore may be better able to heed the warning. On the other hand, some studies find that older adults can perform monitoring operations if instructed to do so (Multhaup, 1995).

An additional purpose of Experiment 2 was to clarify the responses of subjects on the initial forced recall test by measuring their confidence that items produced had occurred in the original study list. Experiment 1 revealed that subjects increase their output under forced recall instructions, but there was no way to determine which of the items subjects actually remembered and which ones were simply guesses, which confidence ratings will permit us to do. Confidence ratings also aid in determining whether subjects can differentiate between veridical memories and guesses when they initially produce the items, or whether this confusion occurs only on later recall tests. However, this change on the first test of Experiment 2 makes it difficult to compare the results of the delayed tests for the two experiments because the ratings of confidence during the first test may alter the pattern of responding during the delayed tests. That is, because confidence ratings on the first test will call attention to the status of the produced items, subjects may better monitor their initial recall attempts, and this monitoring may carry over to the second test and lead subjects to make fewer errors. However, this issue is an empirical question, and the comparison between experiments will prove useful in answering it.

A final change in Experiment 2 was to use a selected population of older adults. Other research in our lab has shown that when older adults are tested on neuropsychological measures designed to tap frontal lobe functioning, the approximately 25% of older adults who score high on these measures resemble younger adults when placed in several memory paradigms (Butler, McDaniel, Dornburg, Price, & Roediger, 2004). The 75% who score low on these measures show memory effects normally associated with aging. Therefore, in Experiment 2 we used older adults who scored low on these frontal tests (as defined by Glisky, Polster, & Rothieaux, 1995). Selecting subjects in this way permits a purer comparison between younger adults and older adults with typical impaired functioning (i.e., we have removed older adults who behave like younger adults in terms of performance on memory tests). Consistent with past studies assessing frontal lobe functioning for older adults only (e.g., Henkel et al., 1998; Mather, Johnson, & De Leonardis, 1999), young adults in this experiment were not screened on the frontal battery because we assume that the battery does not differentiate young adults.

METHOD

Subjects

The subjects were 60 Washington University undergraduates who participated in the experiment in partial fulfillment of a class requirement or for $10 cash (age range 18–25, $M = 20$; Shipley Vocabulary Test mean 33, range 27–37) and

60 older adults recruited through the Washington University Older Adult Subject Pool (age range 66–85, $M = 76$; Shipley Vocabulary Test mean 34, range 27–39). Older adults received $15 for their participation. All older adults were tested on a battery of neuropsychological measures (Modified Wisconsin Card Sort, controlled oral word association test, mental arithmetic, mental control from the Weschler Adult Intelligence Scale, and backward digit span) thought to reflect frontal lobe functioning (Glisky et al., 1995; Glisky, Rubin, & Davidson, 2001). Based on the results of these tests, subjects were categorized as having high or low frontal functioning. Only low frontal functioning older adults were called back to participate in the current experiment, for reasons discussed earlier. Older subjects across experimental conditions were matched on frontal scores as well as age; vocabulary scores were equated for younger and older adults.

Design

The design consisted of a 2 × 2 × 2 between-subject design. Age (younger or older adults) and retrieval condition (cued or forced recall on Test 1) and warning (standard or strong warning on Test 2) were all manipulated between subjects. The dependent variables were veridical and false recall and recognition.

Materials

The same materials used in Experiment 1 were used in Experiment 2. The only exception involved the final source monitoring recognition test, which was modified to be simpler for subjects. The items presented on the test were not changed, but the format of the test was revised to include the three options for source (lists, self, and neither). Subjects were asked to place a check in all the boxes that applied for a given word, so that they could check two boxes for each word (e.g., "list" and "self" if they believed that the word had been in the list and that they had produced it).

Procedure

The procedure of Experiment 2 was similar to that of Experiment 1. Subjects were presented with six categorized word lists via computer and asked to pay careful attention because they would later be tested on their memory for the lists. Subjects then completed a visual spatial filler task for 4 min, followed by the first recall test. The initial recall test was cued or forced, and subjects were asked to recall items from one category at a time. A second cued recall test followed the first test, and finally subjects completed the recognition source monitoring test.

Several changes were introduced in Experiment 2 to provide a different assessment of the effects of forced recall. One difference in the procedure of Experiment 2 was that the initial recall test was written rather than verbal. Based on the results of Gardiner, Passmore, Herriot, and Klee (1977), the results of the written recall tests used in Experiment 2 were not expected to vary significantly from the results of the verbal recall tests used in Experiment 1, either in initial recall or in their effects on later tests.

The use of written initial recall tests permitted testing of subjects in small groups. Written recall tests and the group testing sessions resulted in one more procedural change in Experiment 2: time limits. Subjects in the cued recall condition were given 2 min per category to recall items, and subjects in the forced recall condition were allotted 4 min per category to write down 20 items. At the end of each time limit, all subjects were asked if they needed more time to complete their recall. Subjects rarely asked for more time, and when they did, the total time limit never exceeded 2.5 min per category for cued recall and 6 min per category for forced recall. The difference in time between the two groups was meant to roughly equate the amount of time allotted for each item produced because subjects in the forced recall condition must produce a greater number of items than subjects in the cued recall condition.

Experiment 2 also included a confidence rating task during the first recall test. For each item produced on the initial recall test, subjects were asked to indicate how confident they were that the item occurred in the study lists using a scale from 1 to 4 (1 = *sure the item did not occur on the list,* 2 = *pretty sure the item did not occur on the list,* 3 = *pretty sure the item did occur on the list,* 4 = *sure the item did occur on the list*). Confidence ratings were made concurrently with recall, and the rating scale was visible throughout the test. Experiment 2 also provided confidence ratings (rather than the "remember" and "know" judgments of Experiment 1) on the second recall test.

One final manipulation of Experiment 2 involved the warning given to subjects before they completed the second recall test and the recognition source monitoring test. One group of subjects replicated the standard warning given to subjects in Experiment 1; another group of subjects received a stronger warning. Instructions against guessing were built into the instructions for the second recall test of Experiments 1 and 2, and the strong warning given to half the subjects in Experiment 2 therefore was meant to be a more specific and severe warning against false memories. Specifically, subjects in the standard warning condition were told, "Do not guess. Just write down the items that you are reasonably sure you saw on the screen." Subjects in the strong warning condition received the following instructions: "When recalling on the first recall test, you may have produced some items that weren't actually on the computer screen by generating exemplars of the category label or by guessing. On this test, please be especially careful to write down only those items you are sure you saw on the computer screen when you originally studied the material. Do not guess. There is no need to produce all items from each list. Only produce those items that you originally studied."

RESULTS AND DISCUSSION

Recall Test 1

Table 4 presents the mean proportion of list and critical items produced on the first recall test by younger and older adults under instructions of cued and forced recall. Analyses of list items are considered first, followed by analyses of critical items.

As in Experiment 1, cued and forced recall were analyzed separately for recall Test 1. Replicating the results of Experiment 1 for both cued and forced recall, we found no age difference for list items produced on the cued recall test, $t(58) < 1.0$, but we did find a significant age difference

Table 4. Mean proportion of items recalled and mean proportion of confidence ratings of 3 ("pretty sure the item was on the list") and 4 ("sure the item was on the list") produced by younger and older adults on a cued recall or forced recall test. Total recall also includes items rated 1 or 2, which are not shown (Experiment 2, $N = 120$)

	Cued recall		Forced recall	
	Younger	Older	Younger	Older
	List recall			
Total	.44	.42	.63	.56
Confidence rating 4	.39	.32	.43	.40
Confidence rating 3	.05	.08	.11	.09
	Critical recall			
Total	.20	.36	.67	.63
Confidence rating 4	.04	.13	.12	.28
Confidence rating 3	.15	.16	.21	.20

between list items produced by younger and older adults on the forced recall test, with younger adults producing more items than older adults, $t(58) = 3.67$, $SEM = .07$

To determine the effect of age on the proportion of critical intrusions produced on recall Test 1, separate analyses were conducted. On the cued recall test, older adults produced significantly more critical items than did younger adults, $t(58) = 3.62$, $SEM = .04$. This finding contrasts with those of Experiment 1, although the same numerical trend was evident in Experiment 1. For forced recall, there was no age difference in the proportion of critical items produced, $t(58) = 1.58$, $SEM = .03$, replicating Experiment 1. Of course, both younger and older subjects in this condition were forced to guess, so both groups were expected to produce critical items. The main interest is in examining the recall of those items on the second recall test with instructions to recall only list items, so the near equivalence of critical items produced by younger and older adults on Recall Test 1 permits a more clear-cut analysis in Recall Test 2.

High confidence ratings

In addition to producing items on recall Test 1, subjects also indicated their confidence (ranging from 1 to 4) that the item had been presented in the original study phase. The confidence ratings were examined to reveal the proportion of items subjects produced that they were sure had been presented in the lists. Table 4 presents the mean proportion of items produced by subjects that were given a confidence rating of 4 ("sure the item was on the list") and 3 ("pretty sure the item was on the list"). Analyses of confidence ratings of 3 and 4 are reported for forced recall only. (These analyses are not reported for cued recall because subjects were instructed not to guess on this test; therefore, nearly all items produced were given a confidence rating of 3 or 4).

Roediger et al.'s (1993) research showed that subjects under forced recall conditions sometimes believe that the items produced from guessing were actually studied. We sought to replicate this effect and see whether it would be even more pronounced in older adults. The confidence ratings revealed an interesting dissociation between age groups. Whereas younger adults, $M = .54$, and older adults, $M = .49$, were equally likely to say they were confident that list items had been presented, $t(58) < 1.5$, older adults were more confident in the critical items produced, $M = .48$, than were younger adults, $M = .33$, $t(58) = 3.26$, $SEM = .05$. Considering only the highest level of confidence, the effect is also impressive, .28 to .12 for older and younger adults, respectively. Thus, even though younger and older adults produced equivalent numbers of critical items on the forced recall test, older adults were more likely than younger adults to judge that the items had been previously presented. Monitoring the sources of information after forced recall is especially difficult for older adults (Kelley & Sahakyan, 2003).

Recall Test 2

The mean proportions of list and critical items produced on recall Test 2 are presented in Table 5. Before the second test, half the subjects had received a standard warning against guessing and half received the strong warning against guessing. The labels in the table ("prior cued recall" and "prior forced recall") refer to the conditions subjects were in on the first recall test.

Table 5. Mean proportion of items recalled and mean proportion of confidence ratings of 3 ("pretty sure the item was on the list") and 4 ("sure the item was on the list") produced by younger and older adults on a cued recall test collapsed across standard warning or strong warning. Total recall also includes items rated 1 or 2, which are not shown (Experiment 2, $N = 120$)

	Prior cued recall		Prior forced recall	
	Younger	Older	Younger	Older
	List recall			
Total	.43	.41	.48	.36
Confidence rating 4	.39	.35	.41	.32
Confidence rating 3	.04	.05	.07	.04
	Critical recall			
Total	.19	.37	.23	.36
Confidence rating 4	.09	.15	.11	.25
Confidence rating 3	.10	.17	.12	.10

Examining the mean proportion of list items produced, a 2 (prior cued or prior forced recall) × 2 (younger or older adults) × 2 (standard or strong warning) ANOVA revealed no main effect of warning, nor any significant interaction between warning and other variables, *F*s < 1.0. Thus, data in Table 5 are collapsed across warning. The ANOVA did reveal a significant main effect of age, $F(1, 112) = 10.24$, $MSE = .01$. Younger adults recalled a significantly greater proportion of list items, $M = .45$, than older adults, $M = .39$. More interesting is the significant interaction between age and retrieval condition, suggesting that the veridical recall of younger and older adults was differentially affected by prior retrieval condition, $F(1, 112) = 5.40$, $MSE = .01$. Guessing on a forced recall test increased younger adults' retention of list items but impaired that of older adults (probably due to source monitoring limitations).

A separate 2 (prior cued or prior forced recall) × 2 (younger or older adults) × 2 (standard or strong warning) ANOVA was conducted for recall of critical items produced on the second recall test. Again, we found no significant main effects or interactions with the warning variable, so data in the table are collapsed across warning. Supporting our original hypothesis that older adults would show greater memory errors than younger adults, there was a significant main effect of age, $F(1, 112) = 16.56$, $MSE = .04$, with older adults recalling significantly more critical items, $M = .36$ collapsing across retrieval condition, than did younger adults, $M = .21$. Interestingly, there was no main effect of prior retrieval condition (cued or forced recall) in Experiment 2 ($F < 1.0$), unlike the results of Experiment 1. Subjects in Experiment 2 were just as likely to produce the critical items when they had previously recalled under cued recall instructions, $M = .28$ collapsing across age and warning, as when they had previously recalled under forced recall instructions, $M = .30$, although there was a trend in the same direction as found in Experiment 1. Although we cannot be certain, the different findings between experiments probably resulted from the collection of confidence ratings on the first test in Experiment 2 but not Experiment 1. The requirement for confidence ratings on the first test probably increased monitoring and reduced false recall on the second test.

Although the confidence ratings used in Experiment 2 were intended as a means of assessing subjects' judgments of recollections during free and forced recall, the ratings may have forced attention to source information on the first test. Subjects in Experiment 1 did not give confidence ratings during the first test, so they may have had an even harder time discriminating on the second test between which items had been on the list and which had not, because an initial recall of a false item increases the likelihood of later falsely recalling it again (Ackil & Zaragoza, 1998; Roediger et al., 1996). If this reasoning is correct, the confidence ratings on the first test in Experiment 2 may have reduced the false memory effect in the forced recall paradigm through enhanced source monitoring during the test. Henkel (2004) provided evidence that forced recall tests with source monitoring instructions show fewer memory errors than forced recall tests with no source monitoring instructions. Furthermore, monitoring of source information has been shown to reduce false memories in other paradigms, such as the eyewitness misinformation paradigm (Lindsay & Johnson, 1989) and the false fame paradigm of Jacoby, Kelley, Brown, and Jasechko (1989; see Multhaup, 1995).

High confidence ratings

As in recall Test 1, subjects made confidence ratings when taking the second recall test under cued recall instructions. The mean proportions of items recalled with high confidence are also presented in Table 5. Because recall Test 2 was always a cued recall test with instructions against guessing, nearly all of the items produced were given confidence ratings of 3 or 4. Thus, to avoid redundancy with analyses of items produced, confidence ratings analyses are not reported, but the conclusions from these analyses were the same as those for the overall results.

Recognition source monitoring test

Subjects' performance on the final recognition source monitoring test is presented in Table 6. The recognition test was always taken individually, and the category labels ("prior cued recall" and "prior forced recall") refer to the condition in which subjects recalled the list items in the first recall

Table 6. Mean proportion of veridical recognition for list items and false recognition for critical items (Experiment 2, $N = 120$)

	Prior cued recall		Prior forced recall	
Source judgments	Younger	Older	Younger	Older
Veridical recognition of list items				
List only	.38	.35	.24	.31
Both list and self	.50	.47	.61	.49
Total correct recognition	.88	.82	.85	.80
Self only	.02	.02	.06	.03
Neither list nor self	.11	.16	.09	.17
False recognition of critical items				
List only	.40	.30	.17	.27
Both list and self	.25	.38	.39	.52
Total false recognition	.65	.68	.56	.79
Self only	.02	.04	.30	.08
Neither list nor self	.33	.28	.14	.13

tests. Data in the tables are grouped according to the various response options for list items and critical items.

As in Experiment 1, total correct recognition was defined as the proportion of occasions on which subjects correctly attributed an item as having occurred in the list ("list only" responses plus "both list and self" responses). False recognition is defined as the proportion of occasions on which subjects erroneously attributed an item as having occurred in the list ("list only" responses plus "both list and self" responses). We performed separate analyses on these two measures.

A 2 (prior cued or prior forced recall) × 2 (younger or older adults) × 2 (standard or strong warning) ANOVA on the proportion of items correctly recognized revealed no significant main effect of warning, nor any interaction between warning and other variables, so data again are collapsed across warning. Furthermore, no significant main effects were obtained for retrieval condition or age. As is evident in Table 6, older and younger adults (*M*s = .81 and .86, collapsed across retrieval condition) were about equally likely to correctly recognize items from the previously presented word lists, consistent with Experiment 1 results and with other research showing little or no age difference on recognition tests (see Balota, Dolan, & Duchek, 2000). The finding of no age differences in correct recognition is especially interesting when examined in relation to the age changes found in false recognition.

A separate 2 (prior cued or prior forced recall) × 2 (younger or older adults) × 2 (standard or strong warning) ANOVA conducted on the mean proportion of critical items falsely recognized again revealed no effects of warning, Fs < 1.0, so the data in Table 6 are collapsed across this variable. The effect of age on false recognition replicated nicely the findings of Experiment 1. Older adults falsely recognized a greater proportion of critical items, $M = .74$, than younger adults, $M = .60$, $F(1, 112) = 5.67$, $MSE = .07$. In Experiment 2, this age effect was further modulated by a significant age by retrieval condition interaction, $F(1, 112) = 5.66$, $MSE = .07$. There was an age difference in false recognition in the forced recall condition, $t(58) = 3.57$, $SEM = .06$, but no age difference in the cued recall condition. After the item has been produced on a forced recall test, older adults are especially likely to think it was studied. This outcome is especially interesting when considered in the context of older adults' equal veridical recognition with younger adults after forced recall. We assume that older adults' poorer source monitoring ability accounts for their especially poor performance on critical items after forced recall. Evidence in support of this hypothesis comes from the finding that after taking a forced recall test, older adults are much less likely than younger adults to attribute an item to the "self only" category. Rather, if an older adult produced the item, he or she is likely to attribute it to having occurred in the list ("list only" and "both list and self" categories), increasing the false recognition rate.

GENERAL DISCUSSION

To summarize our findings, in two experiments we showed that forced recall instructions increased production of list and critical items on an initial test relative to cued recall instructions. On later tests of cued recall and source monitoring, critical items previously produced on the forced recall test often were misattributed to prior study, and this effect was especially pronounced in older adults. Not only were older adults more likely to produce the errors, but they were also more likely to say that they remembered the erroneously recalled items from the list (Experiment 1) or that they were more confident in their erroneous responses (Experiment 2). The effects of forced recall persisted after both standard and strong (specific) warnings against guessing (Experiment 2). On the final recognition source monitoring test, older adults falsely recognized more critical items previously generated during forced recall than did younger adults.

Our results provide an interesting technique to create illusory recall and to study age differences in false recall and false recognition. Forced recall instructions cause subjects to generate their own "misleading information" by analogy to other paradigms in which misinformation is presented by an outside source (as in the eyewitness paradigm developed by Loftus et al., 1978). Prior research has established that forced recall can inflate later false memory levels for younger adults (Roediger et al., 1993; Henkel, 2004) and children (Ackil & Zaragoza, 1998), but the two experiments reported here are the first to show that forced recall can also have dramatic effects on older adults' recollections. The current results show that older adults are even more likely to misattribute items produced during forced recall as actual memories on a later test.

The memory illusion resulting from one's own generation of information in forced recall seems to be especially robust in that even a strong warning before the test did not diminish errors, even though similar warnings reduce false recollections in other paradigms (e.g., Gallo et al., 1997; Wright, 1993). The current results agree with those of Gallo et al. (2001) in showing that a warning just before retrieval can be ineffective in reducing false memories. Furthermore, the effect persisted in Experiment 2 even when subjects made confidence judgments while taking the first forced recall test, which called attention to the source of their recollections. Although these immediate confidence judgments drew attention to the source of the items during the initial test, they did not eliminate errors even when subjects were warned, as occurs in other paradigms (Lindsay & Johnson, 1989; Multhaup, 1995).

The present results provide support for the dual process theoretical analyses (Mandler, 1980; Jacoby, 1991). Kelley and Sahakyan (2003) explained changes in older adults' performance in terms of dual process theories, and we outline their reasoning here and apply it to our results. Dual process theory posits two bases of memory performance: recollection and familiarity. Much evidence suggests that older adults lose the ability to recollect specific details of events and instead rely on familiarity in their responses on memory tests. Although the definition of familiarity has changed somewhat over the years as the theory has developed (see Kelley & Rhodes, 2002), the core idea is that familiarity arises from the fluency or the ease with which an item comes to mind (Jacoby & Dallas, 1981; Jacoby, 1999). When making memory judgments and attributions about retrieved information, subjects may assess both the familiarity of the item (the ease with which it is brought to mind) and recollection (of specific details about the events). The former process is thought to be largely automatic, whereas the latter is more controlled, and the processes are hypothesized to be independent of one another (Jacoby, 1991; Yonelinas & Jacoby, 1994).

Dual process theory makes specific predictions about the effects of aging. Younger and older adults both have intact automatic processes; both can rely on familiarity to infer that an item was studied, and therefore on recognition tests, in which fluency plays a critical role, older adults sometimes perform as well as younger adults (at least on nonspeeded tests). However, older adults are less able to recollect specific details of events than are younger adults, and consequently they have more difficulty in correctly attributing familiarity to its proper context (Dywan & Jacoby, 1990; Jacoby 1999; Jennings & Jacoby, 1993). To explain the current results, the key process underlying older adults' greater tendency to memory errors would be difficulties in recollection. For subjects in both age groups, production of items on the forced recall test increased their familiarity and accessibility. However, on the later cued recall test, younger adults were better able than older adults to recollect which familiar items had been produced as guesses in forced recall and which had been recalled from the list. Older adults' relative inability to oppose familiarity with recollection, then, led to their increased levels of false recall and recognition. Older adults misattributed familiarity from generating items during forced recall to recollection of having seen critical items presented in the list, as shown by their errors on the source monitoring recognition tests.

Dual process theory also accounts well for the finding that false recall was higher in conditions in which subjects had previously experienced forced recall rather than cued recall in Experiment 1. Producing items on a forced recall test increases their familiarity and thus renders them more likely to be produced again and misremembered on a later test. Subjects in the cued recall condition produced fewer additional items, so their familiarity was not increased to the same degree. In Experiment 2, prior forced recall did not lead to greater false recall on the second test than did prior cued recall, probably because the initial confidence ratings given to each item produced on the first test called attention of older and younger adults to the source of their recollections. These confidence ratings served as an additional cue about the items and thereby increased recollection of their source on the second test. Interestingly, forced recall still led to more errors on the source monitoring recognition test for older adults. Apparently, when a copy cue of a generated item was presented to older adults, familiarity outweighed recollection, and they misattributed the item to the list. Younger adults were better able to oppose familiarity with recollection of specific features and so made this error less frequently. Older adults were generally more confident of their errors, a process that probably was also driven by high familiarity unopposed by recollection.

The current results can also be interpreted with Johnson's source monitoring framework (Johnson & Raye, 1977; Johnson et al., 1993). Source monitoring is subjects' ability to remember where or when they learned information. In making a memory decision, subjects must remember not only the information itself (i.e., the item in a word list) but also the source of the information (i.e., where or when they learned it). According to the source monitoring framework, subjects make source attributions by assessing the characteristics associated with a particular event and a particular source. Events that have been externally presented contain perceptual characteristics, whereas events that one produced or generated contain characteristics associated with the cognitive operations involved in the process of production and external characteristics. Imagined events are associated with cognitive operations but no features of external presentation (Johnson et al., 1993). In making a source attribution, then, subjects can use the retrieved features of an event to help determine the source. The process of attributing qualitative characteristics to their original source is not always successful, however, and memory errors arise when subjects misattribute events to the wrong source.

Older adults have been shown to have impaired source monitoring abilities (McIntyre & Craik, 1987). Within the source monitoring framework, older adults make more source errors because they are less able to retrieve accurately the qualitative characteristics associated with an event to correctly determine its source. Even though older adults may be able to recall or recognize target information, they have a difficult time attributing the information to the correct source.

The source monitoring framework can also account well for data from the current experiments by assuming that older adults were confused about the source of familiarity of items on the second test. More specifically, in the forced recall condition of Experiment 1 (without confidence ratings), older adults made dramatic errors on the final source monitoring test. They judged that 83% of the critical items had been presented in the original study phase, whereas in actuality none of these items was presented. During forced recall, subjects write down items they remember from the list, and they also write down items they produce as guesses to fill in the spaces required. In both cases, the subject is retrieving items from the list as well as related items and writing them down. Because source monitoring is more difficult when the possible sources are more similar (Johnson, Foley, & Leach, 1988; Johnson & Raye, 1981), and also when items from different sources are more similar (Henkel et al., 1998), subjects may have had a harder time differentiating list items from guesses when they had produced both earlier. Also, remembering that an item was written down is not discriminative; both studied items and critical items were written down. This fact leads to source monitoring errors, especially for older adults. In contrast, on the second cued recall test after an earlier cued recall test, remembering that an item was written down is more discriminative because subjects were asked to recall only items remembered from the list on the first test.

For subjects in the cued recall condition in the initial test, older adults still made more errors than younger adults. We account for this outcome in the same way that we have proposed to account for older adults' greater tendency to produce errors occurs in the DRM paradigm (Balota et al., 1999; Roediger, Balota, & Watson, 2001); because the critical items are strongly associated to the category and to the presented items, they may be aroused (consciously or unconsciously) during presentation of the list. Therefore, during the test subjects face a reality monitoring problem: Was the item presented, or did I only think about it? In both forced and cued recall, younger adults are better able to retrieve features of presented items and accurately determine source, so they make fewer errors. Older adults' greater memory errors may also result from a greater tendency to base their memory judgments on general, categorical information rather than source-specific details (see Mather et al., 1999).

The source monitoring framework can also explain the discrepant findings between Experiments 1 and 2 in the effects of prior forced recall on performance on the second test. In Experiment 1, prior forced recall resulted in higher levels of false recall on the second test than did prior cued recall, but there was no difference between prior forced and cued recall on later false recall in Experiment 2. As noted earlier, this outcome probably arose because confidence ratings on the initial recall test were required in Experiment 2 but not Experiment 1. The immediate confidence ratings called attention to the source of retrieved items (was the produced item presented in the original list or generated as a guess?), and consequently this cue helped subjects reduce errors on the later test. As has been shown in other paradigms, immediate attention to source information helps to reduce false memories (Jacoby et al., 1989; Lindsay & Johnson, 1989; Multhaup, 1995).

Finally, the most compelling evidence in support of a source monitoring deficit in older adults comes from the final recognition source monitoring test. In this test, subjects were tested explicitly on their memory for the source of information. The source task probably was difficult because subjects may already have recalled some items twice. All subjects had the same source options, and younger and older adults showed clear differences in their abilities to accurately determine sources. Older adults, in particular, had difficulty distinguishing items they themselves produced as guesses (items not on the list) and items they produced that had been presented on the list. Younger adults were much more successful on the final source test in distinguishing items produced on the forced recall test as guesses and items produced on the forced recall test as memories of the list. The finding of explicit source differences between younger and older adults adds further evidence for a source monitoring deficit in older adults and is especially impressive in that the age difference did not appear on the prior cued recall test. Of course, these same results could be explained in terms of dual process theory: The item was familiar, and older adults could not recollect its occurrence and so misattributed it to having occurred in the list. Again, the current study does not provide evidence to differentiate between the possible theoretical accounts, which are themselves rather similar (although see Jacoby et al., 2001). Further research is needed to try to determine which theoretical account better explains age difference in illusory recall. The primary import of the current experiments is to provide a robust means of producing false recollections and age difference in false recollection that are amenable to further study.

Notes

These experiments were based on a dissertation conducted by the first author under the supervision of the second. We thank other members of the committee: Janet Duchek, Leonard Green, Larry Jacoby, Alan Lambert, and James Wertsch. We also thank Jane McConnell for help in organizing and testing subjects and Fergus Craik and Marcia Johnson for helpful comments. This research was supported by National Institute of Aging Grant AG17481.

Correspondence about this article should be addressed to Michelle Meade, Department of Psychology, Montana State University, P.O. Box 173440, Bozeman, MT 59717 (e-mail: mlmeade@montana.edu). Received for publication March 28, 2005; revision received July 4, 2005.

References

Ackil, J. K., & Zaragoza, M. S. (1998). Memorial consequences of forced confabulation: Age differences in susceptibility of false memories. *Developmental Psychology, 34,* 1358–1372.

American Medical Association. (1986). Council report: Scientific status of refreshing recollection by the use of hypnosis. *International Journal of Clinical and Experimental Hypnosis, 34,* 1–12.

Anderson, J. R., & Bower, G. H. (1972). Recognition and retrieval processes in free recall. *Psychological Review, 79,* 97–123.

Balota, D. A., Cortese, M. J., Duchek, J. M., Adams, D., Roediger, H. L. III, McDermott, K. B., & Yerys, B. E. (1999). Veridical and false memories in healthy older adults and in dementia of the Alzheimer's type. *Cognitive Neuropsychology, 16,* 361–384.

Balota, D. A., Dolan, P. O., & Duchek, J. M. (2000). Memory changes in healthy older adults. In E. Tulving & F. I. M. Craik (Eds.), *The Oxford handbook of memory.* New York: Oxford University Press.

Battig, W. F., & Montague, W. E. (1969). Category norms for verbal items in 56 categories: A replication and extension of the Connecticut category norms. *Journal of Experimental Psychology Monograph, 80,* 1–45.

Butler, K. M., McDaniel, M. A., Dornburg, C. C., Price, A. L., & Roediger, H. L. III. (2004). Age differences in veridical and false recall are not inevitable: The role of frontal lobe function. *Psychonomic Bulletin & Review, 11,* 921–925.

Craik, F. I. M. (1983). On the transfer of information from primary to secondary memory. *Philosophical Transactions of the Royal Society of London, B302,* 341–359.

Craik, F. I. M., & McDowd, J. M. (1987). Age differences in recall and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 474–479.

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

Dywan, J., & Jacoby, L. L. (1990). Effects of aging on source monitoring: Differences in susceptibility to false fame. *Psychology and Aging, 5,* 379–387.

Erdelyi, M. H., & Becker, J. (1974). Hypermnesia for pictures: Incremental memory for pictures but not words in multiple recall trials. *Cognitive Psychology, 6,* 159–171.

Erdelyi, M. H., Finks, J., & Feigin-Pfau, M. B. (1989). The effect of response bias on recall performance, with some observations on processing bias. *Journal of Experimental Psychology: General, 118,* 245–254.

Gallo, D. A., Roberts, M. J., & Seamon, J. G. (1997). Remembering words not presented in lists: Can we avoid creating false memories? *Psychonomic Bulletin & Review, 4,* 271–276.

Gallo, D. A., & Roediger, H. L. III. (2002). Variability among word lists in evoking associative memory illusions. *Journal of Memory and Language, 47,* 469–497.

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

Gardiner, J. M. (1988). Functional aspects of recollective experience. *Memory & Cognition, 16,* 309–313.

Gardiner, J. M., Passmore, C., Herriot, P., & Klee, H. (1977). Memory for remembered events: Effects of response mode and response-produced feedback. *Journal of Verbal Learning and Verbal Behavior, 16,* 45–54.

Glisky, E. L., Polster, M. R., & Routhieaux, B. C. (1995). Double dissociation between item and source memory. *Neuropsychology, 9,* 229–235.

Glisky, E. L., Rubin, S. R., & Davidson, P. S. R. (2001). Source memory in older adults: An encoding or retrieval problem? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 1131–1146.

Hashtroudi, S., Johnson, M. K., & Chrosniak, L. D. (1989). Aging and source monitoring. *Psychology and Aging, 4,* 106–112.

Henkel, L. A. (2004). Erroneous memories arising from repeated attempts to remember. *Journal of Memory and Language, 50,* 26–46.

Henkel, L. A., Johnson, M. K., & De Leonardis, D. M. (1998). Aging and source monitoring: Cognitive processes and neuropsychological correlates. *Journal of Experimental Psychology: General, 127,* 251–268.

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, and Cognition, 25,* 3–22.

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., Dolan, P. O., & Marsh, E. J. (2001). Forms of bias: Age-related differences in memory and cognition. In M. Naveh-Benjamin, M. Moscovitch, & H. L. Roediger III (Eds.), *Perspectives on human memory and cognitive aging: Essays in honour of Fergus Craik.* New York: Psychology Press.

Jacoby, L. L., & Hollingshead, A. (1990). Reading student essays may be hazardous to your spelling: Effects of reading incorrectly and correctly spelled words. *Canadian Journal of Psychology, 44,* 345–358.

Jacoby, L. L., Kelley, C. M., Brown, J., & Jasechko, J. (1989). Becoming famous overnight: Limits on the ability to avoid unconscious influences of the past. *Journal of Personality and Social Psychology, 56,* 326–338.

Jennings, J. M., & Jacoby, L. L. (1993). Automatic versus intentional uses of memory: Aging, attention, and control. *Psychology and Aging, 8,* 283–293.

Johnson, M. K., Foley, M. A., & Leach, K. (1988). The consequences for memory of imagining in another person's voice. *Memory & Cognition, 16,* 337–342.

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

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

Kelley, C. M., & Rhodes, M. G. (2002). Making sense and nonsense of experience: Attributions in memory and judgment. *Psychology of Learning and Motivation, 41,* 293–320.

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

Koustaal, W., & Schacter, D. L. (1997). Gist-based false recognition of pictures in older and younger adults. *Journal of Memory and Language, 37,* 555–583.

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

Loftus, E. F., Miller, D. G., & Burns, H. J. (1978). Semantic integration of verbal information into a visual memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 4,* 19–31.

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

Mather, M., Johnson, M. K., & De Leonardis, D. M. (1999). Stereotype reliance in source monitoring: Age differences and neuropsychological test correlates. *Cognitive Neuropsychology, 16,* 437–458.

McDermott, K. B., & Roediger, H. L. III. (1998). Attempting to avoid illusory memories: Robust false recognition of associates persists under conditions of explicit warnings and immediate testing. *Journal of Memory and Language, 39,* 508–520.

McIntyre, J. S., & Craik, F. I. M. (1987). Age differences in memory for item and source information. *Canadian Journal of Psychology Special Issue: Aging and Cognition, 41,* 175–192.

McKelvie, S. J. (1999). Effect of retrieval instructions on false recall. *Perceptual and Motor Skills, 88,* 876–878.

McKelvie, S. J. (2001). Effects of free and forced retrieval instructions on false recall and recognition. *Journal of General Psychology, 128,* 261–278.

Meade, M. L., & Roediger, H. L. III. (2002). Explorations in the social contagion of memory. *Memory & Cognition, 30,* 995–1009.

Mitchell, K. J., Johnson, M. K., & Mather, M. (2003). Source monitoring and suggestibility to misinformation: Adult age-related differences. *Applied Cognitive Psychology, 17,* 107–119.

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

Norman, K. A., & Schacter, D. L. (1997). False recognition in younger and older adults: Exploring the characteristics of illusory memories. *Memory & Cognition, 25,* 838–848.

Rajaram, S. (1993). Remembering and knowing: Two means of access to the personal past. *Memory & Cognition, 21,* 89–102.

Reder, L. M., Wible, C., & Martin, J. (1986). Differential memory changes with age: Exact retrieval vs. plausible inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12,* 72–81.

Ritter, W., & Buschke, H. (1974). Free, forced, and restricted recall in verbal learning. *Journal of Experimental Psychology, 6,* 1204–1207.

Roediger, H. L. III, Balota, D. A., & Watson, J. M. (2001). Spreading activation and arousal of false memories. In H. L. Roediger III, J. S. Nairne, I. Neath, & A. M. Suprenant (Eds.), *The nature of remembering: Essays in honor of Robert G. Crowder.* Washington, DC: American Psychological Association.

Roediger, H. L. III, Jacoby, J. D., & McDermott, K. B. (1996). Misinformation effects in recall: Creating false memories through repeated retrieval. *Journal of Memory and Language, 35,* 300–318.

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

Roediger, H. L. III, & Payne, D. G. (1985). Recall criterion does not affect recall level or hypermnesia: A puzzle for generate/recognize theories. *Memory & Cognition, 13,* 1–7.

Roediger, H. L. III, Srinivas, K., & Waddill, P. (1989). How much does guessing influence recall? Comment on Erdelyi, Finks, and Feigin-Pfau. *Journal of Experimental Psychology: General, 118,* 255–257.

Roediger, H. L. III, Wheeler, M. A., & Rajaram, S. (1993). Remembering, knowing, and reconstructing the past. In D. L. Medin (Ed.), *The psychology of learning and motivation: Advances in research and theory.* San Diego, CA: Academic Press.

Schooler, J. W., Foster, R. A., & Loftus, E. F. (1988). Some deleterious consequences of the act of recollection. *Memory & Cognition, 16,* 243–251.

Smith, S. M., Ward, T. B., Tindell, D. R., Sifonis, C. M., & Wilkenfeld, M. J. (2000). Category structure and created memories. *Memory & Cognition, 28,* 386–395.

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

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review, 80,* 359–380.

Wright, D. B. (1993). Misinformation and warnings in eyewitness testimony: A new testing procedure to differentiate explanations. *Memory, 1,* 153–166.

Yonelinas, A. P., & Jacoby, L. L. (1994). Dissociations of processes in recognition memory: Effects of interference and of response speed. *Canadian Journal of Experimental Psychology, 48,* 516–534.

Zaragoza, M. S., Payment, K. E., Ackil, J. K., & Beck, M. (2001). Interviewing eyewitnesses: Forced confabulation and confirmatory feedback increase false memories. *Psychological Science, 12,* 473–477.

Appendix. Categorized study lists followed by the top five exemplars from each category that were used as critical items.

Birds: crow, bluebird, canary, parakeet, hawk, blackbird, wren, oriole, parrot, pigeon, hummingbird, starling, woodpecker, vulture, swallow, chicken, dove

Critical items: robin, sparrow, cardinal, blue jay, eagle

Human body parts: nose, finger, ear, hand, toe, mouth, stomach, hair, neck, heart, knee, chest, liver, brain, lungs, tooth, elbow

Critical items: legs, arms, head, eye, foot

Vegetables: tomato, lettuce, spinach, asparagus, broccoli, celery, cabbage, string beans, cauliflower, beets, lima beans, squash, onions, radishes, Brussels sprouts, cucumber, turnip

Critical items: carrot, peas, corn, bean, potato

Four-footed animals: tiger, elephant, pig, bear, mouse, rat, deer, sheep, giraffe, goat, zebra, squirrel, wolf, donkey, rabbit, leopard, mule

Critical items: dog, cat, horse, cow, lion

Articles of clothing: skirt, coat, dress, hat, sweater, tie, slip, jacket, slacks, gloves, belt, underwear, shorts, scarf, suit, T-shirt, vest

Critical items: shirt, socks, pants, shoes, blouse

Flowers: orchid, chrysanthemum, lily, pansy, petunia, gardenia, daffodil, dandelion, iris, lilac, geranium, peony, sunflower, azalea, gladiola, lily of the valley, snapdragon

Critical items: rose, tulip, carnation, daisy, violet