ACADEMIC PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Memory and Language 48 (2003) 704–721

Journal of Memory and Language

www.elsevier.com/locate/jml

Memory, monitoring, and control in the attainment of memory accuracy

Colleen M. Kelley* and Lili Sahakyan

Department of Psychology, Florida State University, Tallahassee, FL 32306-1270, USA

Received 8 November 2001; revision received 19 June 2002

Abstract

Three experiments assessed people's ability to strategically regulate memory accuracy in free report. Older adults were substantially less accurate than young adults in free report cued recall. Both older and younger adults made gains in memory accuracy from forced report to free report, but older adults did so at the expense of greater losses in quantity correct. This pattern of gains in accuracy at the cost of losses in quantity was mediated by the level of memory monitoring, and older adults showed less correspondence between their confidence judgments and the accuracy of their responses. When young adults encoded items with full vs. divided attention, the resulting differences in retention set off a cascade of effects including poorer memory monitoring and, ultimately, lower accuracy in free report. We suggest that older adults' problems with memory monitoring and memory accuracy stem from impairments in their ability to recollect details of events.

Keywords: Memory monitoring; Memory accuracy; Aging; Incentives

How do people regulate memory accuracy in the face of wide variations in their ability to remember? In answering that question it is critical to distinguish between conditions where responding to each item is forced, as in recognition tasks, and conditions where responding is under the control of the rememberer, as in recall tasks. Under free report conditions such as recall, people may choose not to answer memory queries unless they can recollect specific details of an event. That strategy would manifest itself as losses in the quantity of memory, which is the typical measure in memory experiments. Koriat and Goldsmith (1994) refer to quantity as an input-bound measure of memory performance, as it represents the number of correct memory responses as a function of the number of items input at study. In contrast, one can measure the accuracy of responses that people give under conditions of free report. Accuracy is an output-bound measure of performance as it represents the number of correct memory responses as a function of the number of items a person chooses to output. Even when candidate responses lack recollected details and are merely familiar, people could still maintain a high level of memory accuracy in that the responses that they do volunteer could be correct as often as those volunteered by people under conditions where recollection is high. Although losses in the quantity of memory performance are problematic, losses in accuracy would create additional difficulties, particularly for social interactions where others would assume that reported memories are indeed accurate.

In this paper, we propose that memory accuracy is modulated by the quality of evidence for a candidate response in conjunction with people's assessment of the quality of that evidence as they monitor those candidate responses. The quality of evidence for a candidate response refers to the validity of the evidence that the

* Corresponding author. Fax: 1-850-644-7739.
E-mail address: Kelley@psy.fsu.edu (C.M. Kelley).

0749-596X/02/$ - see front matter © 2002 Elsevier Science (USA). All rights reserved.
doi:10.1016/S0749-596X(02)00504-1

candidate response is actually a memory. For example, recollection of specific details of an episode is typically highly diagnostic of having experienced an episode. But people have other bases for responding when recollection fails such as responding with any familiar, plausible, or even easily generated alternative (cf. Jacoby & Hollingshead, 1990; Reder, 1987; Reder, Wible, & Martin, 1986). If such alternatives to recollection are less valid, then memory performance could suffer, both in terms of omissions, which would reduce memory quantity, and in terms of commissions or false memories, which would reduce memory accuracy.

Memory accuracy depends not only on the validity of the basis for responding, but on the metamemorial monitoring and control processes that lead people to either offer a response or withhold it (Koriat & Goldsmith, 1996; Nelson, 1996). If people use alternatives to recollection to respond to memory queries, but can sensitively monitor the validity of those alternatives, then they could adjust their responses accordingly. Less valid responses could be withheld when accuracy was most important and volunteered when quantity was most important. Thus, differences in free report memory performance across conditions and across populations need to be analyzed both in terms of potential differences in the bases for responding and in terms of potential differences in monitoring and/or control. We examine these issues in younger and older adults, two groups known to differ in memory retrieval, and possibly in terms of monitoring and control processes as well.

Memory monitoring and control

To assess the role of memory monitoring and control in cued recall, we use the framework developed by Koriat and Goldsmith (1996). Their model distinguishes between retrieval, monitoring, and control. Retrieval is captured by the quantity of correct answers generated in a forced report phase where people generate candidate responses, guessing if necessary. Then, people attempt to monitor the validity of a candidate response and assign it a probability, P_A, of being correct. They exercise control over their responding by setting a criterion, P_{RC}, and compare the assessed probability, P_A, to the response criterion, P_{RC}. Items for which P_A equals or exceeds the response criterion are output, and items that fall below the response criterion are withheld. People exercise control over responding by adjusting their response criteria in accord with situational factors such as explicit payoffs.

Memory performance in free report is thus dependent on three factors in addition to retrieval. The first is monitoring effectiveness, which is the degree to which assessed probabilities of correctness successfully differentiate between correct and incorrect candidate answers. The second factor is control sensitivity, which refers to the degree to which people base their decision to report or withhold an item on that item's assessed probability of being correct. The third factor is response criterion setting, which can be adjusted upward if there are large losses associated with a commission error, or downward if there is no penalty for a commission error and a premium is placed on the quantity of correct answers. Under high incentive conditions for accuracy compared to low or moderate incentives, people will be able to increase accuracy if they have effective monitoring of the probability of the correctness of candidate answers, good control sensitivity, and effective response criterion setting.

Koriat and Goldsmith (1996) have applied their model to experiments which probed general knowledge under various accuracy incentives. They demonstrated that monitoring effectiveness, control sensitivity, and response criterion setting all need to be taken into account to predict trade-offs in quantity and accuracy measures between forced and free report. In addition, they decomposed monitoring effectiveness into two factors: polarization and correspondence. Polarization refers to the distribution of probability assessments. If there is no variability in assessed probabilities of correctness, then the outcome of the monitoring process is not useful for control. The other extreme is high polarization, where a person might assign probabilities of only 0 or 100 to candidate responses. High polarization will be an effective basis for control, but only if there is also good correspondence between those assessed probabilities and actual probabilities of correctness. People could differ in their monitoring effectiveness because of differences in polarization and/or differences in correspondence.

Input to the monitoring process: The diagnosticity of evidence

The input to the monitoring process, that is, the quality of evidence that a response is a memory, may affect the monitoring process itself. When a possible memory includes the recollection of details, and even memories of events that preceded or followed from the event, the probability that it is indeed a memory rather than a product of imagination is high. But people also use other perhaps less diagnostic bases for responding to memory queries. The important question is whether they can effectively evaluate whether a candidate response that is merely familiar or plausible is indeed a memory.

Older adults in particular are apt to make recognition judgments based on memory for the gist of an episode or plausibility of a probe, rather than recollection of an event (Koutstaal & Schacter, 1997; Koutstaal, Schacter, Galluccio, & Stofer, 1999; Norman & Schacter, 1997;

Reder, 1987; Reder et al., 1986; Schacter, Koutstaal, Johnson, Gross, & Angell, 1997; Tun, Wingfield, Rosen, & Blanchard, 1998). Gist or plausibility is less diagnostic of prior presentation than are recollected details, and if people are not aware that gist has lower validity, they would presumably be more vulnerable to false memories in those situations.

People also use familiarity to make recognition judgments (Bartlett, Strater, & Fulton, 1991; Dywan & Jacoby, 1990; Jacoby, 1999b; Jennings & Jacoby, 1993, 1997), particularly when recollection is impaired by age or encoding manipulations such as divided attention (Jennings & Jacoby, 1993, 1997; Jacoby, 1999a). Like gist or plausibility, familiarity is also not as diagnostic of prior presentation as is the recollection of details. Familiarity can be the product of prior knowledge or result from subtle sources of familiarity that produce memory illusions (Jacoby & Whitehouse, 1989; Rajaram, 1993; Whittlesea & Williams, 2000, 2001).

The current experiments use a cued recall paradigm developed by Kato (1985) that produces many false recalls in younger adults. In the paradigm, people study pairs of words, half related and half unrelated. Recall of the second word in a pair is cued with the context word and three letters of the target word; thus if the pair CLOCK–DOLLAR were studied, the test cue would be CLOCK–DO_ _ _R. Some of the unrelated word pairs are deceptive in that an unstudied associate of the context word also fits the target cue. For a deceptive item such as NURSE–DOLLAR, the test cue would be NURSE–DO_ _ _R, and the associatively related competitor would be "doctor." In Kato's study, young adults mistakenly produced the strongly cued competitors for deceptive items (.42) almost as often as they recalled the studied item (.47). Kato suggested that the strongly cued competitor was accessed so easily that people assumed it had been studied, a form of the fluency heuristic (Jacoby & Hollingshead, 1990; Lindsay & Kelley, 1996). Alternatively, the strong associative relationship between the cue word and the deceptive response might be interpreted as episodic familiarity.

The deceptive items provide a particular challenge to memory monitoring. To the extent that people respond to the test cue with whatever word is easily accessible and familiar, they will make errors on the deceptive items. To avoid such errors, they would have to notice that deceptive cues such as NURSE–DO_ _ _R would lead them to think of "DOCTOR" even if they hadn't studied the word.

Aging, memory monitoring, and control

There is reason to suspect the existence of age-related changes in monitoring and control in light of evidence that the frontal lobe and some of the functions it subserves are impaired with age (Moscovitch & Winocur, 1995; West, 1996). The frontal lobe has been implicated in the monitoring and control of behavior (Stuss & Benson, 1986) and in memory functions such as determining the source of memories (Craik, Morris, Morris, & Loewen, 1990; Glisky, Polster, & Routhieaux, 1995; Schacter, Harbluk, & McLachlan, 1984; but see Henkel, Johnson, & De Leonardis, 1998), maintaining a retrieval set (Schacter, Alpert, Savage, Rauch, & Albert, 1996), and engaging in strategic retrieval activities (Moscovitch, 1992; Moscovitch & Winocur, 1995).

The evidence regarding older adults' memory monitoring abilities is mixed, perhaps reflecting the diversity of monitoring tasks. Older adults are unimpaired on item-by-item judgments of learning, or JOL's (Connor, Dunlosky, & Hertzog, 1997), but impaired on feeling-of-knowing judgments (Souchay, Isingrini, & Espagnet, 2000). Our interest is in the monitoring of the validity of candidate memory responses as expressed in the relation between confidence judgments and correctness (Lovelace & Marsh, 1985; Perfect & Stollery, 1993). If older adults use a different basis for memory responses than do younger adults (e.g., familiarity of candidate answers rather than recollection) that could be a valid basis for responding in some situations and invalid in others. Successful monitoring of the validity of candidate responses would hinge on being able to determine when familiarity or plausibility is invalid and to adjust confidence accordingly.

Older adults might experience impaired control sensitivity relative to younger adults. At an extreme, if older adults are less effective at maintaining a retrieval set they may simply complete cues rather than attempt retrieval. This would produce a weaker relationship between confidence assessments and the decision to respond or withhold a candidate response for older adults compared to younger adults. A less extreme impairment in the control of memory might appear as impaired response criterion setting, perhaps by showing less sensitivity to payoffs.

Overview of current experiments

In Experiment 1, we assessed how retrieval, memory monitoring, and memory control determine memory accuracy in younger and older adults. To preview, we found that older adults were much less accurate in cued recall than were young adults, even under high incentives for accuracy. The proximal cause of lower accuracy for older adults was their lower monitoring resolution. However, their lower monitoring resolution may in turn be a function of the quality of evidence that the candidate responses they generate are indeed memories. Variations in the quality of evidence in candidate responses may set off a cascade of processes, leading to lower memory monitoring resolution and, ultimately, lower memory accuracy.

In Experiment 2, we tested how the quality of evidence that a candidate response is a memory affects people's ability to monitor the validity of a response. We did so by comparing the memory monitoring of young adults who encoded items with full vs. divided attention. Our prediction was that encoding items with full rather than divided attention would increase the probability that items would be encoded distinctively and support later recollection of details. Recollection of details would in turn be good evidence that a candidate response was indeed studied, and so lead to better monitoring effectiveness. In Experiment 3, we compared the accuracy of cued recall for older adults and for young adults who encoded items with full attention, and for young adults who encoded items under conditions of divided attention. We predicted similar levels of memory accuracy for older adults and young adults who encoded items under conditions of divided attention.

Experiment 1

Experiment 1 was designed to compare the monitoring effectiveness, control sensitivity, and response criterion setting of older and younger adults as they responded to control and deceptive items. Participants were asked to generate a candidate response, guessing if necessary, and then assign it a probability of being correct. The relation between assessed probability and correctness provides a measure of monitoring resolution. Next, participants decided whether to volunteer that response or withhold it during the free report phase. The relation between assessed probability and the volunteering decision provides a measure of control sensitivity. Finally, participants responded under either moderate or high incentives for accuracy, and we assessed changes in response criterion setting, as well as the level of accuracy attained in free report.

Method

Participants. Participants were 60 college undergraduates (mean age = 18.8, $SD = .93$) randomly assigned to either the moderate or high incentive condition. Sixty older adults (mean age = 72.8, $SD = 5.6$, range 65–88) were recruited from local university alumni and were also randomly assigned to either of the two incentive conditions. Performance on the Mill–Hill Vocabulary test was higher for older ($M = 23.4$, $SD = 3.9$) than for younger adults ($M = 17.7$, $SD = 2.9$), $t(118) = 9.25$, and older adults had completed more years of schooling ($M = 17.5$, $SD = 1.8$) than younger adults.

Materials and procedure. The test materials consisted of 75 word pairs, one third of which were related filler items (e.g., morning–evening), and two-thirds of which were unrelated pairs. Half of the unrelated word pairs were deceptive items that had potentially interfering competitors that were not only associatively related but also shared the same first two letters and last letter with the target (e.g., "nurse–dollar" had an interfering competitor "doctor"). The remaining half of the unrelated word pairs appeared as control items, which paired the second target word with an unrelated cue word (e.g., "clock–dollar").

Deceptive items were selected from Kato (1985) and additional items were developed by the same procedure. Strong associates were first selected from word-association norms (the Edinborough Associative Norms). For each associate (e.g., doctor in the nurse–doctor pair), we used a dictionary to find a word of the same length and approximate word frequency that contained the same first two and final letters as the target word, but was not related to the cue or the associate (e.g. "dollar"). This word became the target of an unrelated deceptive word pair. The 50 unrelated word pairs were then randomly divided into two equivalent sets of 25 each to counterbalance as deceptive vs. control items.

The control items were constructed by pairing common words selected from the Kucera and Francis (1967) word norms with the second (target) words of the deceptive items. The same 25 cue words for the control items were used across the counterbalancing of the two sets of unrelated word pairs. Each participant studied 60 items, with 20 items of each type (control, deceptive, and related filler). Test lists consisted of the 60 study items plus 5 new items of each type for a total of 75 items. Items were counterbalanced for old vs. new status at test, and for presentation as control vs. deceptive items across lists.

Upon the completion of the demographic/health questionnaire, participants proceeded to the memory task. Sixty word pairs were presented for study on a computer screen at a rate of one every 8 s. The participants were told that pairs of words would appear one at a time in the center of the screen and they should study them for a later memory test. After the study phase, participants performed an unrelated filler task for five minutes, writing down the names of boys, girls, and US presidents.

The cued recall test consisted of two stages of report, done in succession item by item. The first stage involved cued forced-report: Participants were presented with the cue word and three letters of the target word and asked to recall the target word using the presented cues. They were instructed to guess the target word if unable to recall. One fifth of the test items were new, and subjects were instructed to say "new" if they found the cue word to be unfamiliar. Immediately after recalling or guessing the word, participants assessed the likelihood that the response they produced was correct using a 0–100% scale. After making the confidence judgment, participants were tested on the same item in cued free-report: They were free to report the item or withhold it by saying "pass."

Free report responses were made under conditions of moderate or high incentives. Participants were told that they could choose to give a response or pass on to the next item without being penalized or rewarded for omitted responses. In the moderate incentive condition, they received 25 cents for each correct item and lost the same amount for each incorrect answer. In the high incentive condition, they received 25 cents for each correct response but were penalized $2.50 for each incorrect answer. They were assured that they did not have to pay any losses if they did not break even.

Participants completed the Mill–Hill Vocabulary test at the end of the session.

Results

Overview: Our initial analyses are aimed at capturing memory performance, both in terms of memory accuracy and memory quantity, first at the forced report stage and then at the free report stage. Participants show gains in accuracy from forced to free report, which come at the cost of loss of some quantity correct. This suggests that people are able to monitor the correctness of candidate responses, although not perfectly, and withhold responses at the free report stage that they deem incorrect. We then analyze the monitoring processes, characterizing both how well calibrated people's confidence judgments are, and how well they can discriminate correct from incorrect candidate responses as indicated by measures of monitoring resolution. Finally, we analyze the memory control processes, in terms of the relation between confidence and the decision to report or withhold a candidate response, and in terms of where people set their response criteria.

The significance level for all statistical tests was set at $\alpha = .05$.

Memory accuracy in free report. Our first analysis focuses on the outcome of the retrieval, monitoring, and control processes by looking at accuracy in free report. Accuracy is scored as the number of correct responses divided by the number of responses offered at free report. An age × incentive × item type ANOVA on the accuracy scores in free report revealed that older adults were much less accurate (.53) than young adults (.75), accuracy was lower for deceptive items (.49) than for control items (.79), and the age difference was particularly pronounced on deceptive items, $F(1,116) = 4.28$, $MSE = .032$, for the interaction of age and item type (see Table 1). Across moderate to high incentive conditions, younger adults showed slightly higher accuracy (.73 vs. .78), whereas surprisingly, older adults were less accurate in the high incentive condition (.47) compared to the moderate incentive condition (.59), $F(1,116) = 5.64$, $MSE = .08$, for the interaction of incentive and age. The poorer performance of older adults in the high incentive condition may simply be spurious. Alternatively, if older adults are less able to retrieve studied items and more likely to respond with familiar items, "trying harder" in the high incentive condition may lead them to apply that strategy more vigorously, and so lead to even more errors.

Retrieval at forced report. To understand the lower accuracy of older adults at free report, and the lower accuracy of all participants on deceptive compared to control items, we turn next to how well retrieval operated as shown by performance under forced report. Quantity and accuracy are equivalent in the forced report stage as participants must produce a response to

Table 1
Mean and standard deviation of quantity and accuracy scores for the free and forced report condition by age, incentive, and item type in Experiment 1

Age group	Incentive	Item type	Report option					
			Forced		Free			
			Quantity and accuracy		Quantity		Accuracy	
			M	*SD*	*M*	*SD*	*M*	*SD*
Younger	Moderate	Control	.66	.16	.61	.18	.85	.15
		Deceptive	.47	.21	.48	.22	.60	.25
	High	Control	.63	.18	.54	.19	.91	.17
		Deceptive	.45	.21	.42	.22	.65	.28
Older	Moderate	Control	.54	.20	.45	.24	.78	.24
		Deceptive	.30	.22	.29	.22	.40	.28
	High	Control	.45	.16	.32	.17	.63	.23
		Deceptive	.22	.15	.19	.15	.31	.23

every item. Both older and younger adults retrieved a lower quantity of correct items for deceptive items (.36) compared to control items (.57), $F(1,116) = 161.43$, $MSE = .016$. Older adults retrieved a lower quantity of correct items at the forced report stage (.38) than younger adults (.55), $F(1,116) = 33.20$, $MSE = .054$. with no interaction of age and item type, $F(1,116) = 2.76$, $MSE = .016$, $p = .10$. Incentive had a marginal effect at the forced retrieval stage, with slightly higher retrieval in the moderate (.49) compared to the high incentive (.44) condition, $F(1,116) = 3.76$, $MSE = .054$, $p = .055$.

Gains in memory accuracy from forced to free report. Were older and younger adults able to use monitoring and control processes to improve their accuracy between forced and free report? The answer is yes, with accuracy increasing from .46 to .64 (see Table 1). An age × incentive × item type × report option (forced vs. free report) mixed model ANOVA on the memory accuracy scores revealed a triple interaction of report option, age, and incentive, $F(1,116) = 8.53$, $MSE = .011$. The highest gains in accuracy were made by young adults in the high incentive condition. Older participants in the high incentive condition made about the same gains in accuracy as older participants in the moderate incentive condition,

Accuracy overall was much higher for control compared to deceptive items, $F(1,116) = 193.00$, $MSE = .040$. Participants were able to make greater gains in accuracy on control items compared to deceptive items, $F(1,116) = 36.19$, $MSE = .007$, for the interaction of report option and item type.

Losses in memory quantity from forced to free report. If participants cannot perfectly distinguish which of their candidate responses in forced report are correct and which are incorrect, then increases in accuracy in free report may come at the expense of quantity correct. We performed an age × incentive × item type × report option mixed ANOVA on the memory quantity performance. Exercising the option of free report led to gains in accuracy with a tradeoff of some loss of quantity for the control items, but little loss of quantity for deceptive items, $F(1,116) = 74.15$, $MSE = .002$, for the interaction of report option and item type. Importantly, there was a triple interaction of report option, item type, and age, $F(1,116) = 5.16$, $MSE = .002$. Older adults lost twice as much in quantity correct on control items as did younger adults as they moved from forced to free report. Both older and younger adults showed negligible losses of quantity for deceptive items.

As noted above, the incentive manipulation affected gains in accuracy only for young adults, with greater gains in the high incentive condition. However, both older and younger adults lost more quantity between forced and free report in the high incentive condition than in the moderate incentive condition, $F(1,116) = 7.36$, $MSE = .003$ for the interaction of incentive and report option.

In summary, the quantity and accuracy analyses showed that both younger adults and older adults increased their accuracy substantially from forced report to free report. For control items, the younger adults showed a gain of 24 percentage points in accuracy from forced to free report, with only a 6 percentage point loss of quantity. For deceptive items, they showed a gain of 16 percentage points, with a loss of only one percentage point in quantity. The older adults also achieved a substantial 20 percentage point gain in accuracy on control items, but at a much higher cost in quantity than younger adults, a drop of 12 percentage points. Older adults' gains were smaller for deceptive items, only 9 percentage points, with a loss of 2 percentage points in quantity.

Overall, older adults show lower accuracy in free report than younger adults, even though all testing occurred under conditions that clearly specified the costs of responding incorrectly. They began with fewer correct items retrieved in forced report compared to young adults. Although older adults made gains in accuracy from forced to free report that were nearly as great as those made by young adults, they did so at the expense of greater losses in quantity correct. To understand the pattern of gains in accuracy and losses of quantity from forced to free report, we turn next to an analysis of monitoring.

Monitoring effectiveness: Subjective confidence and the likelihood of being correct. The confidence rating on each item solicited immediately after forced report is equivalent to the assessed probability that the response is a memory, which is then assumed to be the basis for the decision to volunteer or withhold it. On 1.5% of the items, participants changed their response between forced and free report, and those items are omitted from the following analyses.

Two indices of monitoring effectiveness are evaluated. Calibration taps the absolute correspondence between the assessed probability and the actual proportion correct. Monitoring resolution is a form of relative correspondence, as it represents the ability to discriminate between correct and incorrect answers, and is measured by the Kruskal–Goodman γ correlation (see Koriat & Goldsmith, 1996; Nelson, 1984).

The probability judgments were grouped into 12 levels (.0, .01–.10, .11–20,..., .91–.99, 1.0). Because the incentive was applied only to the free recall performance, the data for the probability judgments was collapsed across the two incentive conditions. The calibration curves based on the forced report performance are presented in Figs. 1 and 2 for control and deceptive items for younger and older adults. The proportion correct is plotted against the mean assessed probability across participants; the diagonal line with an

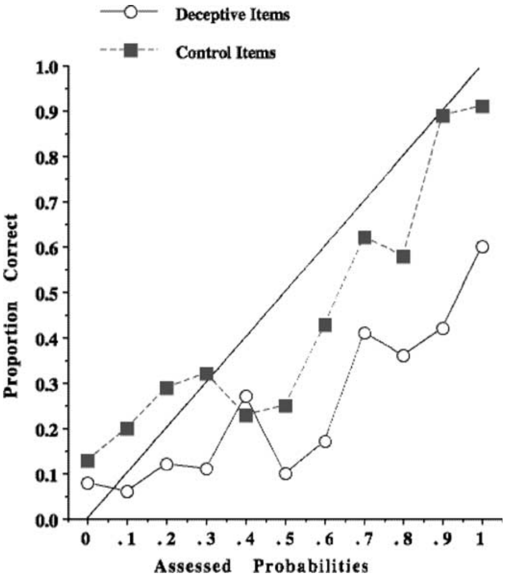

Fig. 1. Calibration curves for young adults on deceptive and control items in Experiment 1. Diagonal line represents perfect calibration.

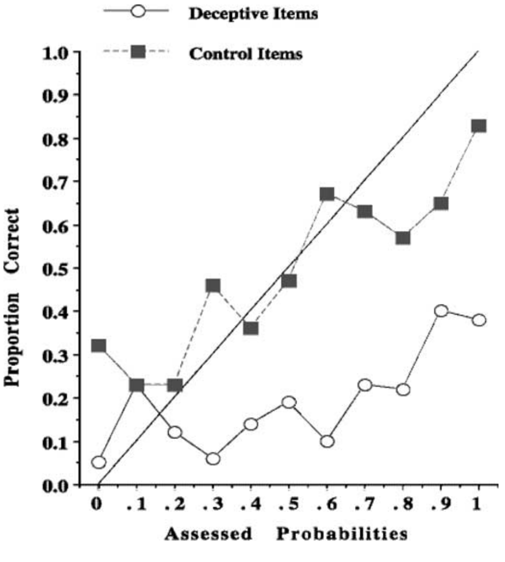

Fig. 2. Calibration curves for older adults on deceptive and control items in Experiment 1. Diagonal line represents perfect calibration.

intercept of .0 and a slope of 1.0 indicates perfect calibration.

On the control items, both age groups were very well calibrated. The younger adults' confidence averaged .67 whereas their actual proportion correct was .64, the respective values for the older adults were .49 and .49. On the deceptive items, however, participants were less well calibrated. This was reflected in the form of overconfidence, which was especially exaggerated for older adults. For the deceptive items, younger adults' confidence averaged .74 when the actual proportion correct was only .46; the respective values for the older adults were .69 and .26.

Participants' individual calibration error scores were computed as the weighted mean of the absolute difference between the actual proportion correct and the mean assessed probability for each category. On the control items, the calibration error scores averaged .18 for the younger adults and .21 for the older adults. On deceptive items, the corresponding values were .33 for younger adults and .46 for the older adults. An age by item type ANOVA on the calibration error scores revealed an interaction between age and item type, $F(1, 115) = 5.46$, $MSE = .027$.

We now turn to the second index of monitoring effectiveness, resolution. The γ correlations between the assessed confidence and the correctness of each answer were computed for each participant. Occasionally, γs were incalculable (e.g., when no correct responses were made, or when only one probability category was used) and the data from the participant was not included in the analysis. This occurred for deceptive items for three older adults in the high incentive condition and one older adult in the moderate incentive condition. An age × item type ANOVA for the γ scores yielded a significant effect of age, $F(1, 112) = 12.98$, $MSE = .16$, and of item type, $F(1, 112) = 16.41$, $MSE = .12$, with no interaction $F(1, 112) = 3.32$, $MSE = .12$, $p = .07$. Older adults had much lower monitoring resolution than younger adults both on deceptive items (mean $\gamma = .46$ vs. .74) and control items (.74 vs. .85).[1] People achieved higher monitoring resolution on control compared to deceptive items.

Memory resolution depends upon having a polarized distribution of assessed probabilities, as well as good correspondence between assessed probability and correctness. There is no apparent age difference in polarization. For control items, the extreme categories of 0

[1] Following Koriat and Goldsmith (1996) we computed a second measure of monitoring resolution (Yaniv, Yates, & Smith, 1991), the adjusted normalized discrimination index (ANDI). This index is interpreted as the proportion of variance in the correctness of the answers that is accounted for by participants' confidence judgments. ANDI scores were calculated for each participant individually and subjected to an age by item type ANOVA. As in the analysis of γs, there was a significant age effect, $F(1, 110) = 20.9$, $MSE = .098$, and a significant effect of Item Type, $F(1, 110) = 23.81$, $MSE = .068$, with no interaction, $F < 1$. The ANDIs revealed lower monitoring resolution for older adults (.37 on control items and .20 on deceptive items) compared to younger adults (.56 on control items and .39 on deceptive items).

and 100% were used by older adults 57% of the time and by younger adults 64% of the time. The corresponding figures for deceptive items are 66% vs. 62%. The lower memory resolution for older adults appears to be a problem of correspondence between assessed probability and correctness. The lower correspondence reflected in the resolution scores for older adults converges with the evidence from the calibration curves and calibration error scores, indicating particular problems with overconfidence for memory responses with high assessed probabilities.[2]

Control sensitivity: Subjective confidence and the decision to respond. Next we analyze the memory control processes that led particular responses to be withheld or volunteered. To assess control sensitivity, the link between assessed confidence and the decision to volunteer an answer was computed in the form of individual γs. The γs were very high for both age groups; γ averaged .95 for older adults and .93 for younger adults, indicating a tight link between confidence and the decision to report a candidate response. An age × incentive × item type mixed model ANOVA on the volunteering γs revealed only an effect of item type, $F(1, 111) = 6.37$, $MSE = .03$, with γs slightly higher on control items (.97) compared to deceptive items (.91).

Response criterion setting. How did changing the incentives for accuracy affect response decisions? In the moderate incentive condition, the monetary loss for a commission error is the same as the gain for correct recall, but in the high incentive conditions the monetary loss for a commission error is 10 times the potential gain for correct recall. Such a payoff structure should induce participants to use a more conservative response criterion. An age × incentive mixed model ANOVA on the proportion of volunteered items at the free report stage yielded a significant main effect of age, $F(1, 114) = 5.59$, $MSE = .026$, and incentive, $F(1, 114) = 9.37$, $MSE = .026$, with no interaction, $F(1, 114) = 1.23$, $MSE = .026$.

Table 2
Mean response criteria, mean fit ratios, and number of responses volunteered and withheld at confidence levels above and below the response criteria, by age, incentive, and item type in Experiment 1

Variable		Control items				Deceptive items			
		Moderate incentive		High incentive		Moderate incentive		High incentive	
		$P_A \geqslant P_{RC}$	$P_A < P_{RC}$	$P_A \geqslant P_{RC}$	$P_A < P_{RC}$	$P_A \geqslant P_{RC}$	$P_A < P_{RC}$	$P_A \geqslant P_{RC}$	$P_A < P_{RC}$
Mean number of responses volunteered	Young	13.90	.23	11.37	.27	14.87	.60	12.47	.40
	Older	9.79	.46	8.67	.67	13.64	.64	11.70	.59
Mean number of responses withheld	Young	.67	5.13	.43	7.83	.80	3.57	.67	6.40
	Older	.64	9.11	.59	10.15	1.18	4.50	1.11	6.44
P_{RC}	Young	.67		.79		.66		.79	
	Older	.70		.60		.69		.72	
Fit ratio	Young	.95		.96		.93		.95	
	Older	.95		.94		.91		.91	

[2] Because of the possibility of ceiling effects on the γs, the confidence-volunteering decision was also evaluated using the ANDI measure (see 1). An age by incentive by item type mixed model ANOVA performed on individual ANDI scores revealed a significant effect of item type, $F(1, 110) = 25.64$, $MSE = .064$. The assessed probability of correctness accounted for a considerable amount of the variance in the decision to volunteer or withhold an item; however, it played somewhat less of a role in the decision to volunteer answers for deceptive items: On the control items, ANDI averaged .80, while on deceptive items it averaged .64. There was also a significant age by incentive interaction, $F(1, 110) = 4.31$, $MSE = .088$, such that for younger adults, the ANDIs increased as the incentive for accuracy increased (.80 in the moderate incentive condition, .88 in the high incentive condition), but did not for the older adults (.80 in the moderate condition, .74 in the high incentive condition).

Younger adults volunteered more items in free report than did older adults (Table 2), which is reasonable given that younger participants retrieved more correct responses in the forced report stage and had higher confidence in candidate responses.[3] As predicted, participants in the high incentive condition volunteered fewer items (.60) than participants in the moderate incentive condition (.69).

Response criterion estimates were derived following the procedure developed by Koriat and Goldsmith (1996) for both item types. They defined response criterion as the value on the confidence scale that determines the volunteering decision rule: Any item receiving a confidence rating equal to or greater than the response criterion is volunteered, while those with confidence ratings below the criterion are withheld. To determine the response criterion for each individual, we treat each probability rating used by a participant as a possible candidate for a response criterion, and then compute the proportion of items conforming to the decision rule. That is, hits are defined as volunteered items that had an assessed probability greater than or equal to the response criterion, and correct rejections are defined as withheld items that had an assessed probability less than the response criterion. The proportion of hits and correct rejections for a possible response criterion is called the fit ratio, and the judgment category with the largest fit ratio is considered to be that participant's response criterion, P_{RC}.

An age × incentive × item type mixed model ANOVA on response criterion estimates yielded only a trend toward an interaction of age and incentive, $F(1,111) = 5.49$, $MSE = .10$. $p = .06$ (see Table 2). Separate analyses of each age group by incentive and item type revealed that younger adults in the high incentive condition used a higher response criterion (.79) than the younger adults in the moderate incentive condition (.66), $F(1,58) = 6.82$, $MSE = .07$. In contrast, older adults in the high incentive condition (.66) did not set their criterion higher than the older adults in the moderate incentive condition (.69), $F < 1$.

Discussion

Older adults had much lower memory accuracy in free recall compared to younger adults. Their lower accuracy was not confined to deceptive items, suggesting a general problem achieving high memory accuracy, rather than a problem only in specialized paradigms that set the stage for high levels of false memories. Older adults' lower accuracy in free report can be traced back to the retrieval stage, where they retrieved fewer correct responses than did young adults. If older adults' monitoring processes were perfect or near perfect, they could selectively withhold only incorrect responses from free report and so narrow the accuracy gap. But older adults' retrieval problems were compounded by difficulties in monitoring the correctness of candidate responses, as indicated by lower γs compared to younger adults. Although the older adults were as well calibrated as young adults on control items, they were poorly calibrated on deceptive items. The poorer monitoring for older adults nonetheless supported large gains in accuracy between forced report and free report, but at a higher cost in quantity correct on control items compared to younger adults.

Memory accuracy in free report was lower for deceptive items compared to control items. As in the case of the population differences, that lower accuracy is evident at the retrieval stage, where fewer correct answers were retrieved for deceptive items. However, monitoring was particularly poor for candidate responses to deceptive items, both in terms of calibration and monitoring resolution. People were confident that incorrect responses to deceptive items were correct. The poorer monitoring for deceptive items produced lower gains in accuracy from forced to free report compared to control items.

While a proximal cause of low accuracy in free report is poor memory monitoring, for both the population differences and item differences, the more distal cause could be the quality of the evidence available at retrieval. The quantity of items retrieved at the forced report stage was lower for older adults compared to younger adults, and lower for deceptive compared to control items. If in addition there are differences in the quality of the evidence retrieved, effective memory monitoring could be more difficult for older adults compared to younger adults, and for deceptive items compared to control items. We address those issues in Experiment 2.

In contrast to the concern that older adults would show impaired control sensitivity and simply complete cues rather than stay on task and attempt retrieval, both older and younger adults showed excellent control sensitivity. They based their decision to volunteer items on their assessed probability of correctness for that item. Age differences in memory control did appear in re-

[3] For younger adults, the items that were volunteered in the free report stage had a mean confidence of .90 compared with .27 on withheld items. The corresponding values for the older adults were .86 for the volunteered items and .17 for the withheld items. An age by incentive ANOVA performed on confidence scores for items that were withheld at the free report stage revealed a significant age effect, $F(1,116) = 11.84$, $MSE = 215.19$, in that younger adults were more confident on withheld items than were older adults. The same analysis on confidence scores of the volunteered items also revealed a main effect of age, $F(1,116) = 7.91$, $MSE = 79.33$, as younger adults were more confident in their volunteered answers than older adults.

sponse criterion setting, however. Only the younger adults responded to the high incentive for accuracy by setting their response criteria higher. Older adults have been shown to respond to various payoff matrixes regarding bias on a recognition test, although not to the same degree as younger adults (Baron & Surdy, 1990). The responsiveness to incentives for accuracy in older adults warrants further study.

Experiment 2

Experiment 1 found poorer memory monitoring for older adults compared to younger adults, which is the proximal cause of their lower levels of memory accuracy. However, older adults' poorer memory monitoring may itself depend on the input to the monitoring process, that is, on the quality of the evidence that a candidate response is a memory. The quality of evidence in a candidate response may in turn depend on elaboration at encoding and the binding of features to create a memory rich in detail.

Changes in the probability of responding on the basis of recollection vs. familiarity of candidate responses could lead to changes in monitoring effectiveness by changing the polarization or correspondence of probability assessments. Analyses of the receiver-operating characteristics or ROC curves in recognition memory find that familiarity-based responses receive a range of confidence ratings, but responses based on recollection typically receive very high confidence ratings (Yonelinas, 1994, 1997). Further, remember responses typically are more discriminative of old/new status in recognition then are Know responses (Gardiner & Java, 1990). When the probability of recollection is relatively high, monitoring effectiveness may be better because it is based on a more polarized set of probability assessments. In addition, when the probability of recollection of details is relatively high, the correspondence between assessed probability and correctness may be relatively high.

We attempted to lower young adults' ability to encode the distinctive details that support recollection by requiring them to study items under conditions of divided attention. Jacoby and colleagues have found that divided attention reduces recollection but leaves familiarity unaffected (Jacoby, Toth, & Yonelinas, 1993). We will assess whether such a change in the evidence that is input to the monitoring process changes monitoring effectiveness. If familiarity does not support good memory monitoring accuracy, then accuracy in free report will be lower for divided attention participants.

Method

Subjects. Participants were 60 undergraduates recruited from Florida State University who received experimental credit for participation. They were randomly assigned to either the full or divided attention study conditions.

Materials. The materials were the same as in Experiment 1.

Procedure. The study and test procedure was the same as for the moderate incentive conditions in Experiment 1 with one exception. Participants in the divided attention condition studied the list of 60 word pairs while simultaneously monitoring an auditory list of random digits presented at a rate of one every 2 s. They were instructed to tap the table whenever three odd digits in a row were detected. The criterion for the digit-monitoring task was that participants could not miss more than two successive sequences, and all participants met that criterion.

Results and discussion

As seen in Table 3, full attention participants generated a higher quantity of correct responses in forced report than did divided attention participants, and then attained a higher level of accuracy when they exercised the option of free report than did participants in the divided attention condition. This was confirmed in an encoding condition × report option × item type mixed model ANOVA on accuracy, with a main effect of condition, $F(1,58) = 17.91$, $MSE = .101$. Accuracy was higher for both conditions during free report compared to forced report, $F(1,58) = 111.90$, $MSE = .098$, and for control items compared to deceptive items, $F(1,58) = 191.43$, $MSE = .027$.

There was a trade-off of quantity and accuracy as participants moved from forced report to free (see Table 3). A mixed model analysis of variance on the quantity of correct answers revealed a drop in quantity from forced to free report, with the greatest loss as participants in the divided attention moved from forced report of control items to free report, $F(1,58) = 4.65$, $MSE = .002$, for the triple interaction of report option, item type, and condition.

The question of particular interest in Experiment 2 is whether divided attention also led to lower memory monitoring resolution as reflected in the γs relating confidence and accuracy for responses produced during forced report. As seen in Table 3, participants who encoded the list with full attention had higher γs than did participants who encoded the list with divided attention, $F(1,57) = 7.20$, $MSE = .178$. In contrast to Experiment 1, where a population difference in monitoring ability per se could have led older adults to have lower monitoring resolution than young adults, the young adults in the divided attention condition have monitoring problems that stem from the difference in conditions of encoding. As in Experiment 1, monitoring was worse for deceptive items compared to control items, for both full

Table 3
Means and standard deviations of quantity and accuracy scores, monitoring and control γ coefficients and response criterion by condition, item type, and report option in Experiment 2

Report option	Measures	Condition							
		Full attention				Divided attention			
		Control items		Deceptive items		Control items		Deceptive items	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Forced report	Quantity and accuracy	.63	.19	.39	.18	.48	.14	.23	.13
Free report	Quantity	.57	.23	.36	.18	.38	.19	.20	.13
	Accuracy	.82	.24	.51	.25	.67	.19	.28	.17
	Monitoring γ	.87	.18	.69	.53	.71	.25	.44	.63
	Control γ	.98	.04	.94	.09	.94	.15	.97	.05
	Response criterion (P_{RC})	.72	.25	.67	.27	.53	.23	.55	.27

and divided attention participants, $F(1,57) = 7.18$, $MSE = .205$.

Participants in the divided attention condition set their response criterion for free report at a lower level (.54) did participants in the full attention condition (.74), $F(1,53) = 7.50$, $MSE = .091$. The distribution of confidence ratings also differed somewhat between conditions, with participants in the full attention condition being more likely to assign a confidence rating of 100, which they used for 50% of their responses, compared to 39% of the responses for participants in the divided attention condition. The full attention participants were as likely to use the confidence rating of 100 for deceptive items (51% of responses) as control items (49%), but divided attention participants were more likely to use the confidence rating of 100 for deceptive items (43%) compared to control items (34%).

The relation between confidence ratings and the decision to report or withhold an item at the free report stage was very high, with γs ranging from .94 to .98, suggesting that both full and divided attention participants exercised a high level of control over responding.

In sum, divided attention at encoding led to lower quantity correct in forced report compared to full attention, and to poorer monitoring resolution. Nonetheless, divided attention participants increased the accuracy of their responses from forced to free report, but never attained the same level of accuracy as did full attention participants. They lost a higher quantity of correct items from forced to free report, particularly for control items, even though they used a much lower response criterion for free report.

The results illustrate how the input to the monitoring and control processes have important ramifications for how those processes function. Our interpretation is that divided attention at encoding reduced the probability that participants could recollect the details of prior study of items. As a consequence, more of the candidate responses in forced report were merely familiar. In the case of deceptive items in particular, that evidence that a candidate response was a memory was misleading, a fact that was not fully appreciated by participants in the monitoring process. The result was that accuracy in free report was extremely low.

Given the very large effect of the study manipulation on monitoring resolution, one might ask whether the monitoring process is completely a function of retention as measured by quantity in forced report, or whether the two can diverge. In the 12 conditions of Experiments 1 and 2, monitoring is generally a monotonic function of retention level in forced report, without any conditions showing a major reversal (e.g., greater retention, but lower average γ). To illuminate the contributions of retention and monitoring to free report accuracy, we performed a separate stepwise regression analysis of free report accuracy for control vs. deceptive items, using quantity in forced report and monitoring γs as predictors. For control items, the overall model with both predictors accounted for a significant proportion of the variance in free report accuracy, $R^2 = .73$, $F(2,58) = 77.92, p < .001$. The quantity in forced report accounted for significant unique variance, $\Delta R^2 = .23$, $F(1,60) = 49.21$, $p < .001$, and the monitoring γs also accounted for significant unique variance, $\Delta R^2 = .12$, $F(1,58) = 26.24$, $p < .001$.

The same analysis on deceptive items found that the overall model accounted for a significant proportion of the variance in free report accuracy, $R^2 = .86$, $F(2,56) = 171.32$, $p < .001$. The quantity in forced report accounted for significant unique variance, $\Delta R^2 = .57$, $F(1,56) = 228.71$, $p < .001$, and the monitoring γs also account for significant unique variance $\Delta R^2 = .04$, $F(1,56) = 17.46$, $p < .001$. Although retention as measured by forced report quantity and monitoring as measured by γs show moderate simple correlations, monitoring nonetheless contributes independently to the level of accuracy in free report.

Experiment 3

Experiment 3 compared the cued recall performance of younger and older adults, and younger adults whose attention was divided during encoding. Experiment 3 was actually the first study performed in the series, but we report it here because accuracy in cued recall is the final result of the monitoring and control processes. We have seen relatively low levels of accuracy in free recall for older adults (Experiment 1), and for young adults who encoded items under conditions of divided attention (Experiment 2). We trace that low level of accuracy to differences in retention as the input to the monitoring function and to the level of monitoring resolution obtained. However, such low accuracy may be partially due to the method used to assess monitoring and control. By forcing participants to guess in the first phase, we may have altered their approach to the task, possibly by inducing a strategy of generate-recognize rather than using the cue to retrieve the target word. The method used in Experiments 1 and 2 could also have inflated accuracy, because it forced participants to explicitly monitor candidate responses before the decision to produce or withhold the response in free report. Participants in Experiment 3 are simply given the context word and fragment as cues to recall, without being asked to provide confidence judgments first.

A second question in Experiment 3 is whether the performance of older adults can be mimicked by testing young adults whose encoding has occurred under conditions of divided attention. Jennings and Jacoby (1993) demonstrated that the high false fame errors of older adults also occurred in younger adults whose attention had been divided at study. They argued that older adults' problem was one of recollection of the study list, which was mimicked by the young adults who encoded items with divided attention. Thus, both groups accepted items as famous when they were familiar due to presentation on the study list. In contrast, Koutstaal, Schacter, and Brenner (2001) matched the performance of young divided attention participants to the level of performance of older adults on veridical recognition memory of one-of-a-kind pictures, but still found the older adults to have elevated false recognition of pictures when many instances of the same category had been studied. They suggested that older adults might engage in less stringent retrieval monitoring than younger adults, or have problems binding features of items, and so accept lures in recognition that share features with studied items. In our paradigm, the erroneous responses to deceptive items share orthographic and phonological features with the studied items, which might lead older adults to accept them as having been studied, as in Koutstaal et al. However, erroneous responses are also semantic associates of the context word, and so would be highly familiar. If both older adults and younger adults whose attention has been divided at study rely on familiarity because of problems recollecting study details, we may see matched levels of accuracy on deceptive items, as in Jennings and Jacoby.

Method

Subjects. Participants were 64 undergraduates recruited from Florida State University who received experimental credit for participation. They were randomly assigned to either the full or divided attention study conditions. Thirty-two older participants (mean age of 76.4, $SD = 6.0$; mean education of 16.6 years, $SD = 2.7$) were recruited from a panel of subjects and alumni, and tested with full attention during study.

Materials. The materials were largely the same as those in the previous experiments. Each study list consisted of 18 related pairs, 18 unrelated deceptive pairs, and 18 unrelated control pairs. The cued recall test included all studied items plus 9 new cues, three of each item type. Items were counterbalanced for old vs. new status at test and for presentation as control vs. deceptive items.

Procedure. After completing a brief demographic/health questionnaire, the participants were presented with a list of 54 word pairs, one pair at a time for 8 s, on a computer screen and told to study the items for the later memory test. Participants in the young adult divided attention condition simultaneously monitored an auditory list of random digits presented at a rate of one per two seconds, and were instructed to tap the table whenever three odd digits in a row were detected. They also repeated the word pairs aloud continuously. Upon completion of the study phase, participants performed an unrelated filler task for five minutes, writing down the names of boys, girls, and US presidents.

Immediately after the filler task, participants were given a cued recall test. They were warned that some cues had not been on the study list in which case they were instructed to say "new." They were also encouraged not to guess but to "pass" on a retrieval cue if they felt they could not recall the item. Participants had 12 seconds per item for recall.

Results and discussion

Quantity. A subject group × item type mixed-model analysis of variance (ANOVA) was used to analyze the quantity of correct cued recall. There was a main effect of subject group, $F(2,93) = 7.67$, $MSE = .061$, with the older adults and young adults in the divided attention

Table 4
Means and standard deviations of quantity and accuracy scores by condition in Experiment 3

Condition	Quantity				Accuracy			
	Control items		Deceptive items		Control items		Deceptive items	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Young full attention	.50	.20	.39	.20	.87	.14	.55	.23
Older full attention	.37	.21	.22	.19	.65	.24	.29	.24
Young divided attention	.36	.18	.26	.14	.79	.18	.34	.17

condition producing lower quantities of correct cued recall than the young adults in the full attention condition (see Table 4). The divided attention manipulation equated the young adults' mean level of correct recall on control items to that found in the older adults (.36 vs. .37). Quantity recalled for deceptive items was substantially lower than quantity recalled for control items $F(1,93) = 62.8$, $MSE = .011$.

Accuracy. Participants in the young full attention condition achieved higher accuracy than participants in the young divided attention condition, who in turn achieved higher accuracy than the older adults (see Table 4), $F(2,93) = 17.25$, $MSE = .054$. Neuman Keul's tests revealed that each group was significantly different from the others. Thus, even though the divided attention manipulation equated young adults and older adults on free-report quantity correct for control items, older adults were nonetheless more prone to errors of commission. As in Experiments 1 and 2, accuracy was higher on control than on deceptive items, $F(1,93) = 236.02$, $MSE = .029$.

The interaction of item type and groups was not significant, $F(2,93) = 2.40$, $p = .10$. However, because of the trend toward an interaction, we looked more closely at the group differences for each item type. A one-way ANOVA on accuracy on the control items found significant group differences, and Neuman Keul's tests revealed that the older adults were less accurate on control items than either the full attention young participants or the divided attention participants, who did not differ from each other. In contrast, for the deceptive items, the one way ANOVA on accuracy found a different pattern, confirmed with Neuman Keul's tests: Older adults and young divided attention participants were both less accurate than young full attention participants, but did not differ from each other.

This variation in the patterns of commission errors across groups could come about if young divided attention participants tended to output responses that came to mind readily in response to the memory cues *and* which were familiar. That would lead to high levels of errors on deceptive items, but fewer on control items. However, older adults committed commission errors even on control items, albeit to a much smaller degree than on deceptive items. Older adults may have been willing to endorse candidate responses that merely came to mind readily, even when the response lacked an associative relationship to the cue.

In sum, older adults and young adults whose attention was divided at encoding exhibited much lower memory accuracy than young adults who had studied the items with full attention. Our attempt at matching the older adults and young divided attention adults on free-report quantity correct recall for control items succeeded, but older adults nonetheless did not attain the same level of accuracy as the young divided attention participants. Similar to young full attention participants, young divided attention participants avoided commission errors on control items while older adults were less able to do so. In contrast, both young divided attention participants and older adults made more than twice as many commission errors as correct recalls on deceptive items. The parallels between the performance of older adults and young participants whose attention was divided during encoding suggests that their performance on deceptive items reflects a greater reliance on familiarity of a candidate response, rather than recollection. This was a fairly effective strategy for control items, but quite ineffective for deceptive items.

General discussion

The current experiments studied variations in memory accuracy in free report that are due to: (1) population differences between older and younger adults; (2) variations in the qualities of the items studied (control vs. deceptive items); (3) experimentally manipulated encoding conditions, and (4) different incentives for accuracy. We then studied how those factors affected retrieval, memory monitoring, and memory control processes to produce the final variations in free report memory accuracy.

Accuracy, monitoring, and control in episodic memory

Experiments 1 and 2 provide an extension of Koriat and Goldsmith's (1996) framework from answering general knowledge questions to answering episodic memory queries. The framework and method produced parallel results in the current study, which we will review first in terms of the results for young adults who encoded items with full attention. Accuracy increased dramatically from forced to free report, with only minimal losses in quantity correct. These increases in memory accuracy from forced to free report were made possible by the quality of participants' memory monitoring. High memory accuracy was also due to excellent control functions, in that the decision to respond in free report was based on the assessed probability that an item was correct.

Experiment 1 assessed how different incentives for accuracy would affect memory control functions and, consequently, affect the gains in accuracy between forced and free report. In the young adults, a high incentive for accuracy led to greater gains in accuracy from forced to free report compared to the moderate incentive, at the cost of slightly higher losses in quantity. The greater gains in accuracy in the high incentive condition came about as young adults set their response criteria higher than in the moderate incentive condition.

We also found that accuracy in free report was much lower for deceptive items compared to control items (cf. Koriat and Goldsmith's Experiment 2 on deceptive general knowledge questions). Lower memory accuracy on deceptive items can be traced back to the retrieval stage, where there was a lower quantity of correct answers retrieved for deceptive compared to control items. Conceptually, lower quantity in retrieval does not dictate the level of accuracy in free report, although it does limit the quantity correct in free report. However, because people were less able to monitor the correctness of candidate memory responses for deceptive items compared to control items, they were unable to make such large gains in accuracy from forced to free report for deceptive items as they were able to make for control items. Consequently, poorer memory monitoring for deceptive items compounded the problem created by poorer retrieval, and led to lower levels of memory accuracy on deceptive compared to control items in free report.

Typical studies of free recall treat performance as a direct measure of retention, without acknowledging the role of monitoring and control over responding that must be occurring to produce performance. Studies which compare performance across different retrieval monitoring conditions (e.g., Koutstaal et al., 1999; Multhaup, 1995) point to the dynamic quality of monitoring and control and its consequences for memory accuracy. The effect of incentives on memory accuracy (Hirt, 1990) or memory quantity (Weldon, Blair, & Huebsch, 2000) have rarely been explored. The strength of the Koriat and Goldsmith paradigm is that it pulls together all of the processes, retention, monitoring, and control, and reveals how they combine to produce memory accuracy in free report.

Input to the monitoring process: The diagnosticity of memory cues

The current experiments expand Koriat and Goldsmith's model by illustrating the importance of the quality of evidence that a candidate response is a memory. Experiment 2 found that divided attention at encoding lowered monitoring resolution, which in turn led to lower memory accuracy. The regression analyses revealed that retrieval as indexed by quantity correct at forced report makes a large contribution to memory accuracy at free report. In addition, monitoring makes a unique contribution to free report accuracy, particularly for control items.

Dividing attention at encoding lowered monitoring resolution several ways. First, it reduced participants' ability to encode details that could provide a recollective experience at test. The presence or absence of details of the study episode would provide a highly diagnostic basis for claiming the response as a memory (Robinson, Johnson, & Robertson, 2000), compared to familiarity. Recollection may also produce more polarization in the distribution of confidence ratings as recollection is typically accompanied by the highest level of confidence (Yonelinas, 1994, 1997), which in turn would afford higher monitoring resolution compared to responses based on familiarity.

Variations in encoding can also affect later monitoring resolution by changing the proportion of correct responses based on guesses. When correct answers can be guessed, probability assessments are less polarized, supporting lower monitoring resolution (Koriat & Goldsmith, 1996). In our Experiment 2, dividing attention at study reduced retention and so increased the probability that correct candidate responses to control items would be guessed, which lowered monitoring resolution.

The third way in which monitoring resolution varies with encoding applies mainly to the deceptive items. The structure of deceptive items is such that when memory fails, guesses will seldom be correct. Therefore, poorer retention will increase the generation of incorrect responses (e.g., "doctor" errors in the nurse dollar case). Because there is poor correspondence between assessed probability and the actual probability correct (the deceptive errors are confidently endorsed), monitoring resolution is lower. Any encoding manipulation that increases the proportion of such responses in the pool of candidate responses should lower monitoring resolution.

One consequence of lower correct guessing rates is that there is a smaller tradeoff between accuracy and quantity as one moves from forced to free report (Koriat & Goldsmith, 1996), which we found for deceptive items compared to control items. As seen in the calibration curves of Experiment 1 (Figs. 1 and 2), there is a much lower probability of generating correct answers to deceptive items for assessed probabilities between 0 and 50 compared to control items. Thus, moving from forced report to free report using the criterion levels adopted in Experiment 1 (mean criteria ranged from .60 to .79) led to little loss of quantity for deceptive items, but some loss of quantity for control items.

In sum, variations in the input to the monitoring process changed the effectiveness of monitoring. Monitoring resolution is higher when candidate responses are accompanied by details of the study episode, rather than merely familiar, as recollection supports a more polarized distribution of assessed probabilities and better correspondence than does familiarity.

Our thinking about the input to the monitoring process is similar to the cue utilization approach of Gigerenzer and colleagues (Gigerenzer, Hoffrage, & Kleinbolting, 1991), in that there are cues that indicate a response is a memory, and those cues vary in terms of utility. People's overconfidence on deceptive items might be taken to indicate that they are not well tuned to the validity of cues in memory. However, the items are deceptive in that what is typically a fairly valid cue (the easy generation of a familiar response) has been made invalid by a contrived selection of items. This is similar to Gigerenzer et al.'s analysis of overconfidence on responses to general knowledge questions: Experimenters select a nonrepresentative sample of questions where cues to correctness are not as valid as in the natural environment. In the case of memory illusions, the contrived situation allows us to show that the cue is a basis for the attribution that one is remembering. However, they do not necessarily imply that people are suboptimal in their assessment of the utility of that cue.

Age-related changes in retrieval, monitoring, and control

An important focus of Experiment 1 was the comparison of older and younger adults in terms of memory accuracy in free report, and how that related to differences in the retrieval, monitoring, and control of memory. Older adults were much less accurate than younger adults during free report cued recall. The problem for older adults began in forced report, where older adults showed lower quantity correct. Then, older adults had poorer monitoring resolution as reflected by lower average γs compared to younger adults, which prevented them from catching up to the younger adults in accuracy at free report. Lower monitoring resolution also led older adults to suffer larger losses in quantity correct from forced to free report, particularly for control items.

Retrieval and monitoring. One interpretation of the older adults' lower memory accuracy in Experiments 1 and 3 is that it derives from a general problem with memory monitoring. Older adults' assessed probability that a candidate response is correct was less related to the actual probability correct than for younger adults, as measured by γs. However, we think that older adults' poorer memory monitoring resolution derives primarily from their increased reliance on familiarity of candidate responses rather than recollection of details of the study experience (Jacoby, 1999b; Jacoby, Debner, & Hay, 2001; Parkin & Walter, 1992). Just as divided attention during encoding led young adults to have lower monitoring resolution compared to their full attention counterparts (Experiment 2), the older adults' monitoring problems may have followed from their encoding difficulties. Additional evidence for our interpretation comes from the finding that the young divided attention participants in Experiment 3 had levels of accuracy on deceptive items as low as the older adults.

Our match of older adults and young divided attention participants on free-report quantity correct on control items in Experiment 3 does not reveal whether it occurred despite differences in quantity retrieved, quality of monitoring, or control processes prior to free report. Because Experiments 1 and 2 used the same procedure and materials, we also did a cross-experiment comparison of older adults (Experiment 1, moderate incentive condition) and young divided attention adults (Experiment 2). In that case, we equated memory performance at the forced report stage for control items. To do so, we had to analyze a subset of the divided attention participants ($n = 21$), as overall they had higher forced report performance than the older adults. When older and young divided attention adults were equated on forced report retrieval of control items ($M = .55$ for older adults, $SD = .19$, $M = .55$ for young divided attention adults, $SD = .12$), all other monitoring and free report accuracy measures were also equated. In this comparison, older and young divided attention adults attained the same level of free report accuracy on control items (.78 for older adults, .72 for younger adults), and deceptive items (.40 for older adults, .31 for younger adults), and the same level of monitoring resolution (γ averaged .67 for older adults and .62 for younger adults). This evidence converges with the parallel patterns of results in Experiment 3 for young divided attention and older adults on deceptive items.

There may nonetheless be differences in monitoring resolution between older and younger adults in some situations that are independent of the quantity and quality of evidence in memory. Confidence judgments may have both nonanalytic and analytic components (Kelley & Jacoby, 1996), parallel to Koriat's (1997) distinction between subjective mnemonic cues vs. intrinsic and extrinsic cues in JOL's. The nonanalytic component might consist of a simple assessment of subjective familiarity, whereas analytic components to confidence would involve interpretation of cues such as whether word pairs are related or not. In the current paradigm, part of the analytic component to assessments of probability for candidate answers was likely the participant's appreciation of the structure of the deceptive items. Some participants exclaimed during testing "Oh, you're trying to trick me" when they encountered the deceptive items. Understanding the structure of the deceptive items could lead participants to adopt a more stringent criterion, or rely strictly on recollection when they realized an item was deceptive, in a process similar to use of the distinctiveness heuristic (Schacter, Israel, & Racine, 1999). On-line analytic reasoning about the structure of deceptive items might be more difficult for older adults than younger adults, contributing to lower memory monitoring resolution.

Our results have some similarities to those of Jacoby (1999a), who suggested that deficits in recollection lead older adults to rely on whatever response is highly accessible during recall. Younger and older adults studied lists of related word pairs, and later attempted to recall given cues such as "bed s_ee_" as a cue for "sheet." A prime word was presented before each test trial that was either the correct answer for congruent trials or another related word for incongruent trials (sleep). Participants were warned to try to recall the studied item to avoid being misled by the prime. Older adults were more likely to respond with the prime word than younger adults, even when given the option to pass. Younger adults whose attention had been divided during study tended to avoid the prime word, suggesting that they were not willing to risk being misled by the prime when they were not able to remember the target. In our experiments, we did not warn participants about deceptive items as did Jacoby, and the origin of the incorrect response for deceptive items was perhaps not as obvious. This could perhaps account for why younger participants whose attention had been divided at study produced as many false recalls as older participants.

Control processes. The performance of the older and younger adults differed in terms of responsiveness to incentives. Younger adults responded to the high incentive with a higher response criterion, whereas older adults did not. This shows a lack of control sensitivity in older adults, and future research should explore incentives for accuracy in a within-subjects design. Higher response criteria can lead to higher accuracy, but at the cost of memory quantity. Koriat and Goldsmith's (1996) simulations of accuracy and quantity tradeoff patterns as a function of response criterion setting found that accuracy increases as a linear function of response criterion, but that quantity performance decreases as a positively accelerated function of response criterion. They assumed perfect calibration and a uniform distribution of response probabilities. In the simulation, there was no cost of increased accuracy for relatively low response criteria, but as response criterion was raised, there were increasing costs. A parametric analysis of people's responses to incentives across levels of retrieval would reveal when further increases in response criteria are not functional due to high losses in quantity.

To summarize, the consequence of lower retention in combination with lower memory monitoring resolution was that older adults were not able to achieve high levels of memory accuracy. It is important to determine the extent of this problem with memory accuracy by testing in other memory paradigms. For example, some tasks such as associative recognition, are almost entirely accomplished via recollection rather than familiarity (Yonelinas, 1997). Older adults may have better monitoring resolution when recollection or guessing are the only ways to respond.

Summary and conclusion

We found that memory accuracy was lower when people were unable to recollect the details of prior presentation and so relied on other bases for remembering. The critical mediator between retention and accuracy in free report was the effectiveness of memory monitoring, which was lower for older adults than younger adults, for deceptive compared to control items, and for participants who encoded items with divided rather than full attention. People did not completely appreciate the lower validity of cues that a candidate response was a memory when using bases for remembering other than recollection.

Acknowledgments

Thanks to Lara Heflin, Ian Mathis, and Katherine Raynor for assistance in testing participants, and to Jeffrey Feddon for programming.

References

Bartlett, J. C., Strater, L., & Fulton, A. (1991). False recency and false fame in faces in young adulthood and old age. *Memory & Cognition, 19*, 177–188.

Baron, A., & Surdy, T. M. (1990). Recognition memory in older adults: Adjustment to changing contingencies. *Journal of the Experimental Analysis of Behavior, 54*, 201–212.

Connor, L. T., Dunlosky, J., & Hertzog, C. (1997). Age-related differences in absolute but not relative metamemory. *Psychology and Aging, 12*, 50–71.

Craik, F. I. M., Morris, L. W., Morris, R. G., & Loewen, E. R. (1990). Relations between source amnesia and frontal lobe functioning in older adults. *Psychology and Aging, 5*, 148–151.

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

Gardiner, J. M., & Java, R. I. (1990). Recollective experience in word and nonword recognition. *Memory & Cognition, 18*, 23–30.

Gigerenzer, G., Hoffrage, U., & Kleinbolting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98*, 506–528.

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

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*, 1–18.

Hirt, E. R. (1990). Do I see only what I expect? Evidence for an expectancy-guided retrieval model. *Journal of Personality and Social Psychology, 58*, 937–951.

Jacoby, L. L. (1999a). Deceiving the elderly: Effects of accessibility bias in cued recall performance. *Cognitive Neuropsychology, 16*, 417–436.

Jacoby, L. L. (1999b). Ironic effects of repetition: Measuring age-related differences in memory. *Journal of Experimental Psychology: Learning, Memory, and 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., & Hollingshead, A. (1990). Toward a generate/recognize model of performance on direct and indirect tests of memory. *Journal of Memory and Language, 29*, 433–454.

Jacoby, L. L., & Whitehouse, K. (1989). An illusion of memory: False recognition influenced by unconscious perception. *Journal of Experimental Psychology: General, 118*, 126–135.

Jacoby, L. L., Toth, J. P., & Yonelinas, A. P. (1993). Separating conscious and unconscious influences of memory: Measuring recollection. *Journal of Experimental Psychology: General, 122*, 1–16.

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

Jennings, J. M., & Jacoby, L. L. (1997). An opposition procedure for detecting age-related deficits in recollection: Telling effects of repetition. *Psychology and Aging, 12*, 352–361.

Kato, T. (1985). Semantic-memory sources of episodic retrieval failure. *Memory & Cognition, 13*, 442–452.

Kelley, C. M., & Jacoby, L. L. (1996). Adult egocentrism: Subjective experience versus analytic bases for judgment. *Journal of Memory and Language, 35*, 157–175.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126*, 349–370.

Koriat, A., & Goldsmith, M. (1994). Memory in naturalistic and laboratory contexts: Distinguishing the accuracy-oriented and quantity-oriented approaches to memory assessment. *Journal of Experimental Psychology: General, 123*, 297–316.

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

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

Koutstaal, W., Schacter, D. L., & Brenner, C. (2001). Dual task demands and gist-based false recognition of pictures in younger and older adults. *Journal of Memory and Language, 44*, 399–426.

Koutstaal, W., Schacter, D. L., Galluccio, L., & Stofer, K. A. (1999). Reducing gist-based false recognition in older adults: Encoding and retrieval manipulations. *Psychology and Aging, 14*, 220–237.

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

Lindsay, D. S., & Kelley, C. M. (1996). Creating illusions of familiarity in a cued recall remember/know paradigm. *Journal of Memory and Language, 35*, 197–211.

Lovelace, E. A., & Marsh, G. R. (1985). Prediction and evaluation of memory performance by young and old adults. *Journal of Gerontology, 40*, 192–197.

Moscovitch, M. (1992). Memory and working with memory: A component process model based on modules and central systems. *Journal of Cognitive Neuroscience, 4*, 257–267.

Moscovitch, M., & Winocur, G. (1995). Frontal lobes memory and aging. In J. Grafman, K. Holyoak, & F. Boller (Eds.), *Structure and functions of the human prefrontal cortex* (Vol. 769, pp. 119–150). New York: Annals of the New York Academy of Sciences.

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

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95*, 109–133.

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

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.

Parkin, A. J., & Walter, B. M. (1992). Recollective experience, normal aging, and frontal dysfunction. *Psychology and Aging, 7*, 290–298.

Perfect, T. J., & Stollery, B. (1993). Memory and metamemory performance in older adults: One deficit or two? *The Quarterly Journal of Experimental Psychology, 46A*, 119–135.

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

Reder, L. M. (1987). Strategy selection in question answering. *Cognitive Psychology, 19*, 90–137.

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

Robinson, M. D., Johnson, J. T., & Robertson, D. A. (2000). Process versus content in eyewitness metamemory monitoring. *Journal of Experimental Psychology: Applied, 6*, 207–221.

Schacter, D. L., Alpert, N. M., Savage, C. R., Rauch, S. L., & Albert, M. S. (1996). Conscious recollection and the human hippocampal formation: Evidence from positron emission tomography. *Proceedings of the National Academy of Science, USA, 93*, 321–325.

Schacter, D. L., Harbluk, J. L., & McLachlan, D. R. (1984). Retrieval without recollection: An experimental analysis of source amnesia. *Journal of Verbal Learning and Verbal Behavior, 23*, 593–611.

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

Schacter, D. L., Koutstaal, W., Johnson, M. K., Gross, M. S., & Angell, K. E. (1997). *Psychology and Aging, 12*, 203–215.

Souchay, C., Isingrini, M., & Espagnet, L. (2000). Aging, episodic memory feeling-of-knowing, and frontal functioning. *Neuropsychology, 14*, 299–309.

Stuss, D. T., & Benson, D. F. (1986). *The frontal lobes*. New York: Raven Press.

Tun, P. A., Wingfield, A., Rosen, M. J., & Blanchard, L. (1998). Older adults show greater susceptibility to false memory than young adults: Temporal characteristics of false recognition. *Psychology and Aging, 13*, 230–241.

Weldon, M. S., Blair, C., & Huebsch, P. D. (2000). Group remembering: Does social loafing underlie collaborative inhibition? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 1568–1577.

West, R. L. (1996). An application of prefrontal cortex function theory to cognitive aging. *Psychological Bulletin, 120*, 272–292.

Whittlesea, B. W. A., & Williams, L. D. (2000). The source of feelings of familiarity: The discrepancy-attribution hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 547–565.

Whittlesea, B. W. A., & Williams, L. D. (2001). The discrepancy-attribution hypothesis, part II: Expectation, uncertainty, surprise, and feelings of familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 14–33.

Yaniv, I., Yates, J. F., & Smith, J. E. K. (1991). Measures of discrimination skill in probabilistic judgment. *Psychological Bulletin, 110*, 611–617.

Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: Evidence for a dual-process model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 1341–1354.

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