# Adult Age Differences in Production and Monitoring in Dual-List Free Recall

By: Christopher N. Wahlheim, B. Hunter Ball, Lauren L. Richmond

****\*Note: Figures may be missing from this format of the document**

## Abstract:

The present experiment examined adult age differences in the production and monitoring of responses in dual-list free recall. Younger and older adults studied 2 lists of unrelated words and were instructed to recall from List 1, List 2, or List 1 and List 2. An externalized free recall procedure required participants to: (a) report all responses that came to mind while recalling from specific lists, (b) classify responses as correct or incorrect, and (c) provide confidence judgments for their accuracy classifications. Relative to younger adults, older adults showed a monitoring deficit by misclassifying proportionally more responses and discriminating more poorly between correct and incorrect responses in their confidence judgments. This deficit was especially pronounced under conditions of retroactive interference that occurred when participants recalled from List 1 only. A comparison of retrieval dynamics for all responses produced and for those that participants were reasonably confident were correct provided information about age differences in preretrieval context reinstatement and postretrieval monitoring of retrieved context. One noteworthy finding was that total production when recalling from List 1 showed that List 2 responses remained more accessible across the first several retrieval attempts for older than younger adults, which indicated a substantial age difference in the ability to reinstate List 1 context. Overall, the present findings provide a nuanced characterization of age differences in the operation of production and monitoring mechanisms under conditions of proactive and retroactive interference that can inform models of free recall.

**Keywords:** aging | free recall | interference | metacognition | monitoring

## Article:

Episodic memory deficits experienced by older adults are most pronounced when retrieval is self-initiated and competing information creates interference (for reviews, see Balota, Dolan, & Duchek, 2000; Zacks, Hasher, & Li, 2000). Dual-list free recall is an ideal task for examining these deficits because it provides little environmental support (Craik, 1986) and offers flexible analysis options that reveal underlying contextually based mechanisms (Kahana, 1996). A recent computational model proposed that differences in context reinstatement and monitoring of retrieved context can in part explain age-related deficits in free recall (Healey & Kahana, 2016).

Additionally, a recent behavioral approach using dual-list free recall has implicated roles for these mechanisms in older adults' greater susceptibility to proactive and retroactive interference (Wahlheim & Huff, 2015; Wahlheim, Richmond, Huff, & Dobbins, 2016). The primary aim of the present experiment is to further examine the mechanisms underlying these age differences using a behavioral approach. Specifically, we leverage conceptual notions and empirical methods from the metacognition literature (Goldsmith, 2016; Koriat & Goldsmith, 1996) to characterize the operation of pre- and postretrieval mechanisms in these age differences. We provide a brief overview of relevant literature before describing the present experiment.

Free Recall Dynamics

Age-related deficits in free-recall tasks are well-established in the memory and aging literature (Ceci & Tabor, 1981; Craik, 1968; Hultsch, 1969; Schonfield & Robertson, 1966). These deficits commonly result in older adults recalling fewer correct items and committing more intrusions relative to younger adults (Hartley & Walsh, 1980; Kahana, Dolan, Sauder, & Wingfield, 2005; Kahana, Howard, Zaromb, & Wingfield, 2002; Stine & Wingfield, 1987; Wahlheim & Huff, 2015; Wahlheim et al., 2016). One method for characterizing the role of retrieval processes in these deficits is to decompose the retrieval sequence to reveal differences in both the manner of retrieval initiation and the transitions that follow across subsequent retrievals. This decomposition method has typically revealed no age differences in retrieval initiation patterns when comparing probability of first recall (PFR) curves that plot first-recalled items conditionalized on input position. When recalling from a single list, PFR curves show recency effects on immediate tests and primacy effects on delayed tests. However, retrieval transitions throughout recall show that younger adults are more likely than older adults to subsequently recall items from adjacent input positions (e.g., Kahana et al., 2002; Wahlheim & Huff, 2015). The diminished temporal contiguity of responses exhibited by older adults is considered to partly reflect deficits in the ability reinstate and monitor context, which we define as internal states and external features associated with but not including the items themselves.

A recent context-based computational model has proposed a more comprehensive account of age differences in free recall dynamics (Healey & Kahana, 2016). Specifically, the model proposes that four candidate processes can account for such age differences. The model assumes that older adults have deficits in sustained attention, reinstatement of context, source monitoring to reject intrusions, and the resolution of internal evidence used for reporting decisions. Despite its elegance, one limitation is that the model has only been tested in proactive interference situations in which participants can use time-of-test context to retrieve from target lists. For example, the paradigm typically used to assess free recall dynamics involves many study-test cycles with participants always recalling from an immediately preceding list (e.g., Kahana et al., 2002). In such procedures, intrusions originate either from prior lists or outside of the experiment. However, these procedures do not capture the everyday phenomenon that individuals often must retrieve earlier episodes in the face of retroactive interference from subsequent competing episodes. Thus, additional investigation of the mechanisms underlying age differences in retroactive interference in free recall is warranted.

In this vein, Wahlheim and colleagues (Wahlheim & Huff, 2015; Wahlheim et al., 2016) recently investigated the mechanisms underlying age differences in both proactive and retroactive interference using a dual-list free recall paradigm, inspired by earlier studies (Epstein, 1969, 1970; Jang & Huber, 2008; Sahakyan & Hendricks, 2012; Shiffrin, 1970; Unsworth, Brewer, & Spillers, 2013; Unsworth, Spillers, & Brewer, 2012; Ward & Tan, 2004). Their experiments included several trials comprised of two study lists separated by a context break (e.g., a space bar press), each followed by a recall test of List 1, List 2, or List 1 and List 2. When recalling from individual lists, older adults recalled fewer correct responses and committed more intrusions from non-target lists. Decomposition of these retrieval sequences revealed dynamics consistent with prior findings: PFR curves produced recency effects on immediate tests and primacy effects on delayed tests, and response transitions originated from adjacent input positions more often for younger than older adults when semantic associations among items were minimized (Wahlheim & Huff, 2015).

Wahlheim and colleagues also found unique age differences in response monitoring when items were semantically associated within and between lists and in retrieval initiation when participants recalled from both lists. For monitoring, when semantic associations were present within and between lists (Wahlheim et al., 2016), retrieval sequences from individual lists showed more within-category transitions within than between lists, with the difference being greater for younger than older adults. This suggested that younger adults could more effectively monitor episodic context, which refers to the ability to accurately remember specific details about the source of retrieved items, when semantic associations made the sources difficult to discriminate. These findings are consistent with other studies showing that older adults are more prone to mistakenly remember information from nontarget sources being from target sources due to impaired memory for source details (e.g., Dodson, Koutstaal, & Schacter, 2000). For retrieval initiation on trials where participants were instructed to recall from both lists, younger adults showed List 2 recency effects akin to those on immediate tests, whereas older adults showed List 2 recency effects and List 1 primacy effects. These differences reflected older adults initiating retrieval more variably across trials, which was not because of their lower memory ability nor did it reflect their tendency to commit more intrusions.

Taken together, these studies of age differences in free recall provide converging evidence that older adults are impaired in their abilities to reinstate context from specific lists and monitor the source of retrieved items to decide whether to report them. Importantly, no studies to our knowledge have directly assessed the role of metacognitive processes in age differences in free recall under conditions of proactive and retroactive interference. Also, the mechanism underlying age differences in retrieval initiation pattern when recalling from two lists remains to be clarified. We addressed these issues here by examining younger and older adults' response production and monitoring in dual-list free recall.

Metacognitive Monitoring

The notion that optimal recall performance depends on monitoring ability is a central assumption of classic process models of episodic recall (Raaijmakers, 2003; Raaijmakers & Shiffrin, 1981). This assumption is consistent with generate-recognize models positing that participants retrieve

items from both correct and incorrect sources, but can withhold reporting of items from inappropriate sources (Anderson & Bower, 1972; Keppel, 1968; Raaijmakers & Shiffrin, 1980; Wixted & Rohrer, 1994). In this vein, Kahana et al. (2005)assessed whether age differences in free recall in part reflect differences in monitoring ability using a variant of the externalized free recall (EFR) procedure (Bousfield & Rosner, 1970; Roediger & Payne, 1985). Their EFR procedure required participants to report all responses that came to mind when recalling from a list and to press a key following responses that were not from the list. Younger adults produced more intrusions than older adults, but older adults rejected fewer intrusions, suggesting that impaired monitoring contributed to age-related deficits in the precision of recall.

The EFR method of assessing response production and monitoring in free recall is similar to procedures used to test proposals of a contemporary model of metacognition that emphasizes the strategic regulation of memory accuracy (Koriat & Goldsmith, 1996). This model proposes that memory accuracy, which is a measure of the ability to report only correct information, requires accurate evaluation of the contents of memory and subsequent control over reporting decisions. In the procedure used to test this, participants respond to every item on a memory test (forced-report), evaluate the accuracy of each response using confidence judgments (monitoring), and decide whether each response should count toward their overall performance on the task (free-report). The relationship between confidence and accuracy determines how well participants can monitor for correct and incorrect responses, the relationship between confidence and report decisions determines one's confidence criterion for outputting a response, and the relationship between memory performance on forced- and free-report measures determines the extent to which participants can regulate their memory accuracy by volunteering correct responses and withholding incorrect responses. Recent studies have used this approach to examine the role of metacognitive processes in age differences in memory performance.

For example, Kelley and Sahakyan (2003) used the strategic regulation approach to examine age differences in an associative interference task. In this task, memory for word pairs in a deceptive condition (e.g., nurse-dollar) was later tested using cue-fragment pairs (e.g., nurse-do_ _ _r) for which the fragment could be completed by an extralist response that was a strong associate of the cue (e.g., doctor). In contrast, memory for word pairs in a control condition (e.g., clock-dollar) was tested using cue-fragment pairs (e.g., clock-d_ _ _r) for which the extralist responses that could complete the fragment (e.g., doctor) were unrelated to the cue. Results showed that memory performance was better for control than deceptive items and that both age groups could use the free report option to improve their accuracy on those items. However, older adults showed poorer metacognitive monitoring for both control and deceptive items that presumably resulted from impaired retrieval quality. Following this, Rhodes and Kelley (2005) used the same task to show that deficits in memory accuracy resulting from impaired monitoring were associated with impaired executive functioning in younger and older adults. More recently, Pansky, Goldsmith, Koriat, and Pearlman-Avnion (2009) examined age differences in memory accuracy using a more naturalistic task of remembering a slide show depicting an event in the life of a family. They found that older adults had poorer monitoring and free-report memory accuracy resulting from lower retrieval quality and more volunteering of incorrect responses.

These studies demonstrate the utility of the strategic regulation approach for assessing the roles of postretrieval monitoring and control processes in age-related memory deficits. However, this approach does not adequately highlight the role of preretrieval processes that determine the quality of retrieved information that serves as a basis for monitoring decisions. This is despite the fact that neuropsychological evidence (e.g., Burgess & Shallice, 1996; Moscovitch & Melo, 1997), verbal theories (e.g., Jacoby, 1999; Jacoby, Kelley, & McElree, 1999; Jacoby, Shimizu, Daniels, & Rhodes, 2005), and context-based computational models of free recall (e.g., Healey & Kahana, 2016; Polyn, Norman, & Kahana, 2009) all implicate a preretrieval mechanism that reinstates context in the service of facilitating production from a target source. To address this, Goldsmith and colleagues (Goldsmith, 2016; Halamish, Goldsmith, & Jacoby, 2012) recently updated the strategic regulation approach to include a preretrieval mechanism that serves to elaborate cues and improve later monitoring and memory accuracy in cued recall. With this addition, their model is more consistent with recent computational models of free recall (e.g., Healey & Kahana, 2016; Polyn et al., 2009). The similarity between approaches suggests that they could be integrated, perhaps by utilizing methods from the strategic regulation approach when examining age differences in free recall. Doing so would allow for the decomposition of not only retrieval sequences but also the component metacognitive processes that contribute to the age-related memory differences. Indeed, employing variants of the measurement techniques from the strategic regulation approach in free recall can elaborate on the mechanisms proposed by computational models and implicated by earlier behavioral results. We took this approach in the present experiment by modifying the EFR procedure to incorporate both accuracy classifications and confidence judgments as measures of monitoring ability in a dual-list free recall paradigm.

The Present Experiment

The present experiment employed a variant of the EFR procedure designed to assess age differences in response production and monitoring in a dual-list free recall paradigm that requires participants to recall from either one or two lists. Based on earlier findings, we expected younger adults to produce more intrusions and more effectively monitor those intrusions relative to older adults (cf. Kahana et al., 2005). One novel contribution of the present experiment was that we employed an EFR with Confidence (EFR-C) procedure that combines free- and forced-report recall with subjective evaluations of the accuracy of response classifications. Although the EFR-C is similar to the methodology used in the strategic regulation approach described above, it differs in that participants are asked to: (a) report any response that comes to mind while attempting to retrieve from specific lists; then (b) indicate whether each response is correct or incorrect (accuracy classification); and finally, (c) provide a confidence judgment evaluating the accuracy classification. The critical difference between the strategic regulation and EFR-C procedures is that confidence judgments in the strategic regulation approach are made prior to report decisions and evaluate the likelihood that responses are correct. In contrast, the EFR-C requires participants to first make a response classification that provides initial information about the ability to monitor response accuracy and then make a confidence judgment to precisely evaluate the classification. We adopted this method to more closely approximate traditional EFR procedures that elicit accuracy classifications immediately following response production. We

added confidence judgments to provide more precision regarding the extent to which both age groups can discriminate between correct and incorrect recalls.

The first way that we examined the role of monitoring in age differences in recall was to compute the relative proportion of accurately classified responses. Based on earlier findings (e.g., Kahana et al., 2005), we expected that older adults would produce fewer intrusions than younger adults and also misclassify proportionally more of those intrusions as being correct responses. In contrast to earlier studies, we also examined participants' ability to endorse correct recalls as such. If older adults are generally impaired in their ability to evaluate the original source of productions, they should also misclassify proportionally more correct recalls as being incorrect than younger adults. The second way we examined the role of monitoring was by comparing confidence judgment magnitudes for correct recalls and intrusions. We assumed that the extent to which confidence magnitudes are greater for accurate than inaccurate classifications provides another index of participants' monitoring ability. We expected that age differences in monitoring would be shown by greater differences in confidence magnitudes for younger than older adults. We also examined whether the predicted age-related monitoring impairments would differ between proactive and retroactive interference situations. Given that older adults are impaired in their ability to reinstate context and that the demands on such reinstatement are greater in retroactive than proactive interference situations, we expected that both accuracy classifications and confidence judgments would show the largest age differences when participants attempted to recall from List 1 while avoiding intrusions from List 2.

We also expected the EFR-C aspect of the current procedure to provide a more accurate characterization of age differences in response output than has been shown in standard free recall. Younger and older adults typically show similar patterns of retrieval initiation in their PFR curves when recalling from individual lists in standard free recall (e.g., Healey & Kahana, 2016). However, older adults sometimes retrieve fewer first-recalled items from target lists than younger adults (e.g., Wahlheim et al., 2016). Taken with older adults' well-established deficit in context reinstatement, this suggests that response output in standard free recall may underestimate the extent to which older adults produce first-recalled items from nontarget lists due to selective reporting. Beyond first-recalled items, there are also substantial age differences in the patterns of response output across the entire recall period (e.g., Wahlheim & Huff, 2015). Age differences in selective reporting might also cause output profiles in standard recall to misrepresent the characteristics of response production throughout recall. To examine the extent to which standard recall results reflect actual age differences in response production, we compared PFR curves and output profiles produced under EFR instructions with the same measures conditionalized on responses judged to be correct with medium to high confidence. We describe the specific comparisons prior to the relevant analyses below.

**Method**

The research reported here was approved by the Institutional Review Board at Washington University in St. Louis.

**Participants**

The participants included in the analyses were 30 younger adults ($M$age = 18.97 years, $SD$ = 0.81, Range = 18–21) and 30 older adults ($M$age = 76.97 years, $SD$ = 6.42, Range = 66–90). Data from one additional older adult were not included because the participant failed to comprehend the task instructions. We selected these sample sizes because they were larger than the samples of 24 participants in each age group used by Wahlheim and Huff (2015) that were sufficient for detecting the effects of age on a variety of free recall measures that conceptually replicated findings from earlier studies. We increased the sample size a bit here because we have never examined age effects on metacognitive measures in free recall, and we wanted to give ourselves a reasonable chance to detect age differences on these measures. Younger adults were recruited from the participant pool at Washington University in St. Louis and were given partial course credit or $10. Older adults were recruited from participant pools maintained by the School of Medicine and the Department of Psychological and Brain Sciences at Washington University in St. Louis and were given $15. Older adults reported significantly more years of education ($M$ = 15.54, $SD$ = 2.65) than younger adults ($M$ = 13.07, $SD$ = .83), $t(56)$ = 4.87, $p <$ 001. Two older adults did not report their years of education. Vocabulary scores on the Shipley Institute of Living Scale (Shipley, 1986) were significantly higher for older ($M$ = 36.13, $SD$ = 2.42) than younger ($M$ = 33.30, $SD$ = 2.58) adults, $t(58)$ = 4.40, $p < .001$, $d$ = 1.13.

### Design and Materials

A 2 (Age: Younger vs. Older) × 3 (Trial: List 1 vs. List 2 vs. List Both) mixed design was used. Age was a between-subjects variable, and trial was manipulated within-subjects. The experiment consisted of 15 study-test trials that each included two 10-word study lists followed by a test. The 15 trials comprised five blocks of three trials, with each block containing one from each of the trial conditions. The presentation order of conditions was randomized within blocks.

Materials were 300 concrete nouns from the MRC Psycholinguistic Database (Coltheart, 1981). Words were four to nine letters in length ($M$ = 5.42, $SD$ = 1.39), had concreteness ratings ranging from 502–670 ($M$ = 578.6, $SD$ = 30.8, Scale = 100–700), and Hyperspace Analog to Language log frequency counts that ranged from 6.94−12.60 ($M$ = 9.63, $SD$ = 1.12).

To counterbalance items across conditions, the 300 words were divided into 30 groups of 10-word lists that were matched on length, concreteness, and frequency. The groups were then clustered into five larger ensembles each consisting of six groups of 10-word lists. Each ensemble was assigned to one of the five trial blocks that were each comprised of two lists from each of the three trial conditions. The assignment of ensembles to blocks was fixed. The 10-word lists within the ensembles were rotated through the two list positions and three trial conditions within each block, resulting in six experimental formats.

### Procedure

Participants were tested individually. Participants first read an overview of the experiment describing the three different trial conditions and the EFR procedure. Before each trial, participants were told that they would study two lists and that their tasks were to read words aloud and remember them for an upcoming test. Each trial began when participants pressed the space bar. Each list within a trial began following the presentation of the list name (i.e., List 1 or

List 2), which appeared for 3 s. Each word within the lists appeared for 1 s in the center of the screen followed by a blank screen for 1 s. After studying both lists, participants were instructed to recall words in any order from either List 1, List 2, or List 1 and List 2. A prompt that read "List 1," "List 2," or "Lists 1 and 2" appeared on the screen for 3 s to indicate the list(s) from which to recall. The recall phase began after the prompt disappeared. No other intervening task occurred between List 2 and the recall phase.

During recall, participants were instructed to report all the words that came to mind while they attempted to recall from target lists. Participants (or the experimenter) typed responses onto the screen and pressed enter after each response. The experimenter typed responses for a few older adults who were not comfortable typing for themselves. Following each response, participants indicated whether the response was from the target source (i.e., correct) by pressing the 1 key or from a nontarget source (i.e., incorrect) by pressing the 2 key. After making these accuracy classifications, participants rated how confident they were in those classifications by pressing the 1 key to indicate low confidence, the 2 key to indicate medium confidence, and the 3 key to indicate high confidence. Prior to completing the 15 critical trials, participants were given a brief practice phase in which the lists and duration of the recall phase were shortened. This allowed the experimenter to discuss the procedure with participants, to assess their understanding of the instructions, to resolve any confusion, and to determine which participants were not comfortable typing for themselves.

### Results

The level for significance was set at $\alpha = .05$. Note that variations in degrees of freedom for conditional analyses below occur when some participants could not be included because they did not provide at least one observation in each cell.

### Overall Recall and Accuracy Classifications

In the following analyses, we computed response frequencies for correct recalls, intratrial intrusions, and extratrial intrusions (collapsed across prior-trial and extraexperimental intrusions) and segmented them based on whether they were classified as correct or incorrect (Figures 1–3). Our analysis plan for each response type was to first compare the total number of responses produced, and then compare the relative proportion of inaccurate classifications (i.e., correct recalls classified as incorrect and intrusions classified as correct) by dividing the number of inaccurately classified responses by the total number of responses produced. We chose to analyze inaccurate classifications to focus on differences in monitoring errors between age groups. We submitted comparisons for each response type to separate Age × Trial ANOVAs.

*Figure 1. Mean number of correct recalls per trial as a function of age, trial, and accuracy classification. Note that the total number of possible correct recalls for the List 1 and List 2 conditions (10) was lower than for the List Both condition (20). Error bars are 95% confidence intervals.*
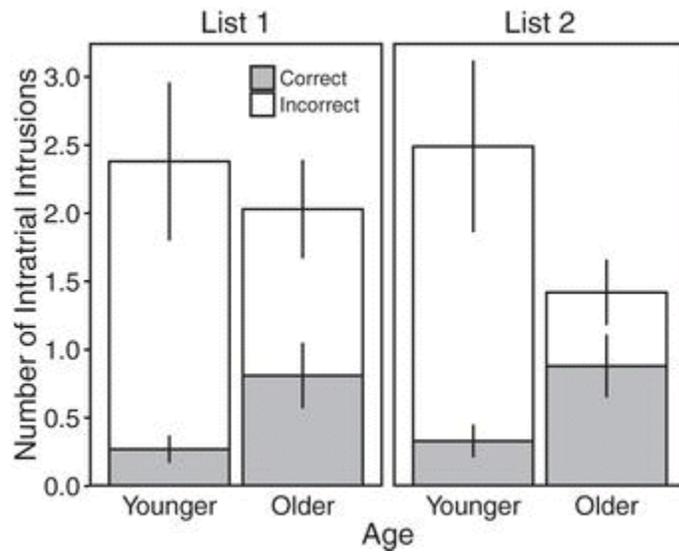


*Figure 2. Mean number of intratrial intrusions per trial as a function of age, trial, and accuracy classification. Error bars are 95% confidence intervals.*
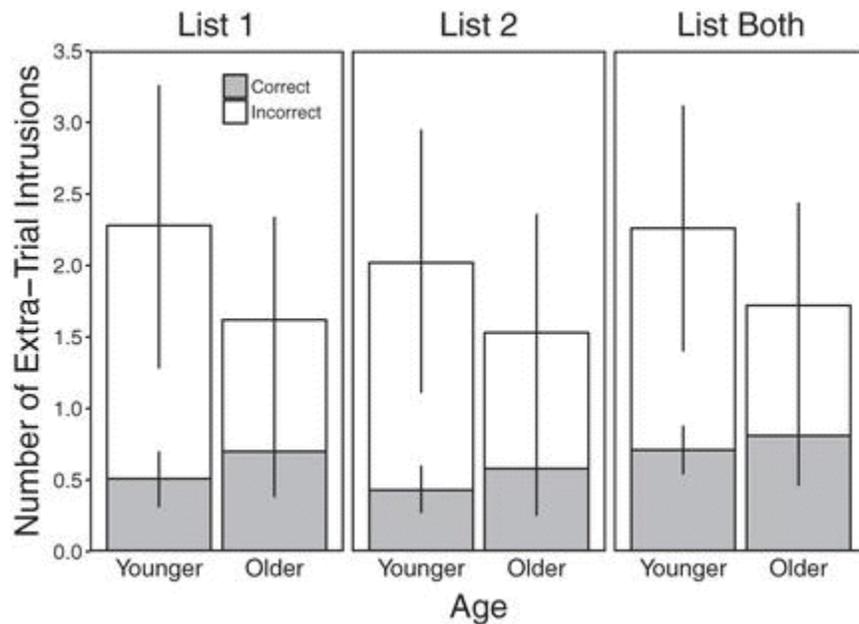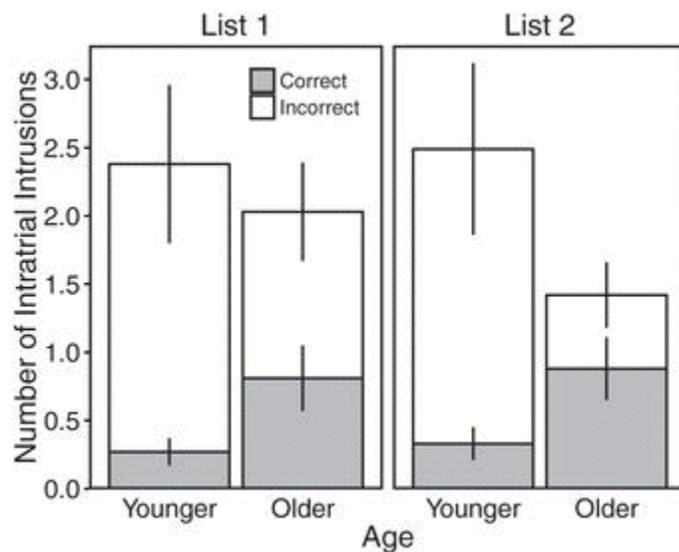
*Figure 3. Mean number of extratrial intrusions per trial as a function of age, trial, and accuracy classification. Error bars are 95% confidence intervals.*



*Figure 1. Mean number of correct recalls per trial as a function of age, trial, and accuracy classification. Note that the total number of possible correct recalls for the List 1 and List 2 conditions (10) was lower than for the List Both condition (20). Error bars are 95% confidence intervals.*

*Figure 2. Mean number of intratrial intrusions per trial as a function of age, trial, and accuracy classification. Error bars are 95% confidence intervals.*
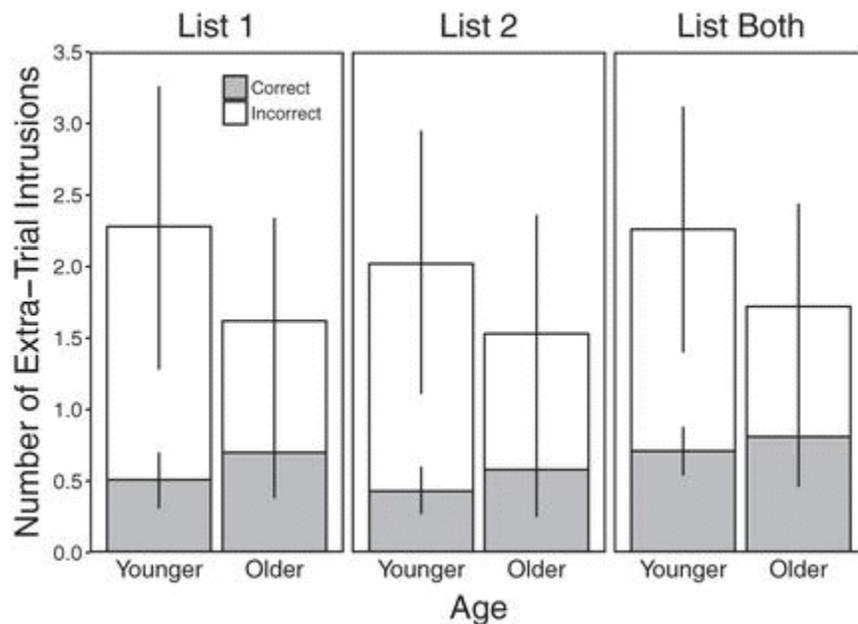


*Figure 3. Mean number of extratrial intrusions per trial as a function of age, trial, and accuracy classification. Error bars are 95% confidence intervals.*

## Correct recall

Figure 1 displays correct recall response frequencies for younger and older adults in all trial conditions. A 2 (Age: Younger vs. Older) × 3 (Trial: List 1 vs. List 2 vs. List Both) ANOVA for total correct recalls produced revealed significant effects of age, $F(1, 58) = 132.13$, $p < .001$, $\eta p^2 = .70$; and trial, $F(2, 116) = 178.16$, $p < .001$, $\eta p^2 = .75$; and a significant Age × Trial

interaction, $F(2, 116) = 20.24$, $p < .001$, $\eta p^2 = .26$. The interaction showed a production advantage for younger adults that did not differ between the List 1 and List 2 conditions (as shown by a nonsignificant 2 (Age) × 2 (Trial: List 1 vs. List 2) interaction, $F(1, 58) = 1.46$, $p = .23$, $\eta p^2 = .03$), but was larger in the List Both than List 2 condition (as shown by a significant 2 (Age) × 2 (Trial: List 2 vs. List Both) interaction, $F(1, 58) = 30.19$, $p < .001$, $\eta p^2 = .34$). A 2 (Age: Younger vs. Older) × 3 (Trial: List 1 vs. List 2 vs. List Both) ANOVA for the relative proportion of correct recalls classified as incorrect revealed significant effects of age, $F(1, 58) = 10.60$, $p = .002$, $\eta p^2 = .15$; and trial, $F(2, 116) = 11.99$, $p < .001$, $\eta p^2 = .17$; and a significant Age × Trial interaction, $F(2, 116) = 4.44$, $p = .01$, $\eta p^2 = .07$. These effects showed that older adults incorrectly classified proportionally more correct recalls than younger adults, younger adults' misclassifications did not differ among trial conditions, *largest* $t(29) = 1.49$, $p = .15$, $d = 0.41$, and older adults misclassified proportionally more correct recalls in the List 1 than List 2 and List Both conditions, *smallest* $t(29) = 2.98$, $p = .006$, $d = 0.61$.
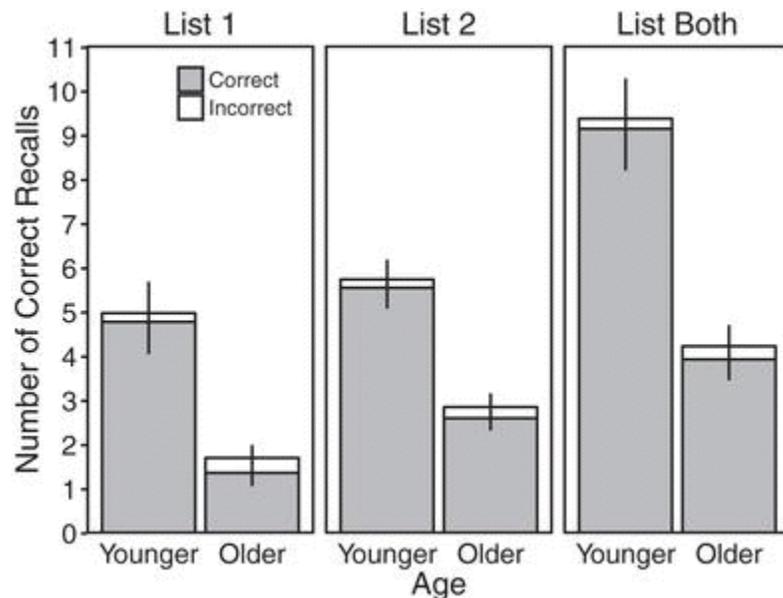


*Figure 1. Mean number of correct recalls per trial as a function of age, trial, and accuracy classification. Note that the total number of possible correct recalls for the List 1 and List 2 conditions (10) was lower than for the List Both condition (20). Error bars are 95% confidence intervals.*

**Intratrial intrusions**

Figure 2 displays intratrial intrusion response frequencies for younger and older adults in the List 1 and List 2 conditions (intratrial intrusions could not occur in the List Both condition). A 2 (Age: Younger vs. Older) × 2 (Trial: List 1 vs. List 2) ANOVA for total intratrial intrusions produced revealed significant effects of age, $F(1, 58) = 5.27$, $p = .03$, $\eta p^2 = .08$; and trial, $F(1, 58) = 4.45$, $p = .04$, $\eta p^2 = .07$; and a significant Age × Trial interaction, $F(1, 58) = 9.05$, $p = .004$, $\eta p^2 = .14$. These effects showed that younger and older adults did not differ in their production of intratrial intrusions in the List 1 condition, $t(58) = 1.06$, $p = .29$, $d = 0.27$, but younger adults

produced significantly more intratrial intrusions than older adults in the List 2 condition, $t(58) =$ 3.25, $p = .002$, $d = 0.84$. A 2 (Age: Younger vs. Older) × 2 (Trial: List 1 vs. List 2) ANOVA for the relative proportion of intratrial intrusions classified as correct revealed significant effects of age, $F(1, 58) = 34.84$, $p < .001$, $\eta p^2 = .38$; and trial, $F(1, 58) = 6.17$, $p = .02$, $\eta p^2 = .10$; and a significant Age × Trial interaction, $F(1, 58) = 9.01$, $p = .004$, $\eta p^2 = .13$. These effects showed that: older adults incorrectly classified proportionally more intratrial intrusions than younger adults, younger adults did not differ in their relative proportions of misclassifications between trial conditions, $t(29) = 0.40$, $p = .69$, $d = 0.08$; and older adults misclassified proportionally more intratrial intrusions in the List 2 than List 1 condition, $t(29) = 3.59$, $p = .001$, $d = 0.66$.
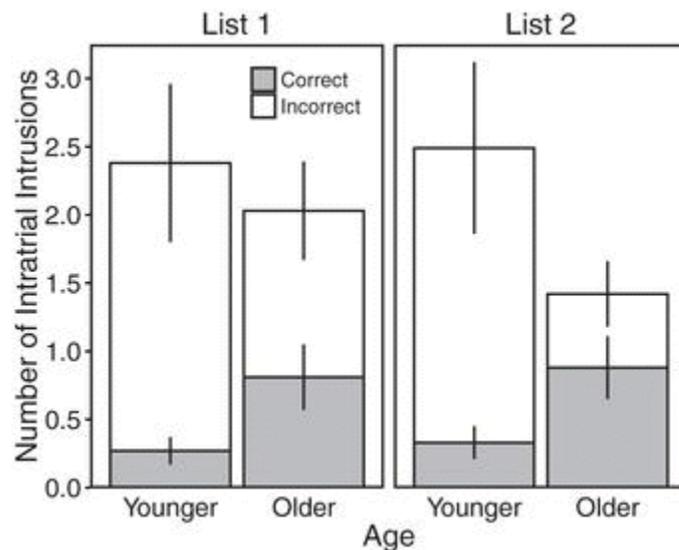


*Figure 2. Mean number of intratrial intrusions per trial as a function of age, trial, and accuracy classification. Error bars are 95% confidence intervals.*

**Extra-trial intrusions**

Figure 3 displays extratrial intrusion response frequencies for all trial conditions. A 2 (Age: Younger vs. Older) × 3 (Trial: List 1 vs. List 2 vs. List Both) ANOVA for all extratrial intrusions produced revealed no significant effect of age, $F(1, 58) = 1.59$, $p = .21$, $\eta p^2 = .03$; a marginal effect of trial, $F(2, 116) = 2.49$, $p = .09$, $\eta p^2 = .04$; and no significant Age × Trial interaction, $F(2, 116) = 0.34$, $p = .71$, $\eta p^2 < .01$. These results showed a slight tendency for participants to produce the most extratrial intrusions in the List Both condition. A 2 (Age: Younger vs. Older) × 3 (Trial: List 1 vs. List 2 vs. List Both) ANOVA for the relative proportion of extratrial intrusions classified as correct revealed no significant effects of age, $F(1, 53) =$ 1.43, $p = .24$, $\eta p^2 = .03$; trial, $F(2, 106) = 1.46$, $p = .24$, $\eta p^2 = .03$; and no significant Age × Trial interaction, $F(2, 106) = 2.07$, $p = .13$, $\eta p^2 = .04$. These results showed no differences in the proportion of extratrial intrusions classified as correct. However, visual inspection of Figure 3 shows patterns similar to those obtained for intratrial intrusions (see Figure 2) suggesting that age differences may have been more difficult to detect for extratrial intrusions.
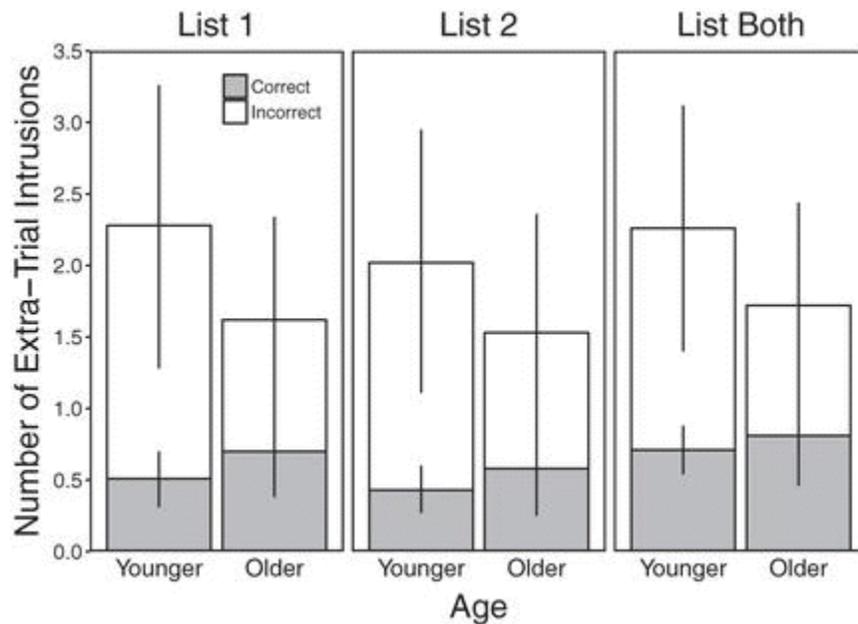
*Figure 3.* Mean number of extratrial intrusions per trial as a function of age, trial, and accuracy classification. Error bars are 95% confidence intervals.

## Confidence Judgments

Confidence judgment magnitudes for accuracy classifications were compared for correct recalls and all types of intrusions to examine age differences in monitoring ability (see Figure 4). We included all intrusions in these analyses to compare confidence judgments for correct and incorrect responses, which is typical for assessing monitoring accuracy (Dunlosky & Metcalfe, 2009). The following analyses were conducted only for the List 1 and List 2 conditions to test for differences in monitoring when participants were instructed to retrieve from a specific list under conditions of retroactive and proactive interference, respectively. As described in the Introduction, greater magnitude differences between accurate and inaccurate classifications were taken to indicate more effective monitoring. We report separate analyses for responses classified as correct and incorrect because many participants did not produce at least one response in every cell, and this approach maximized the number of participants that could be included.
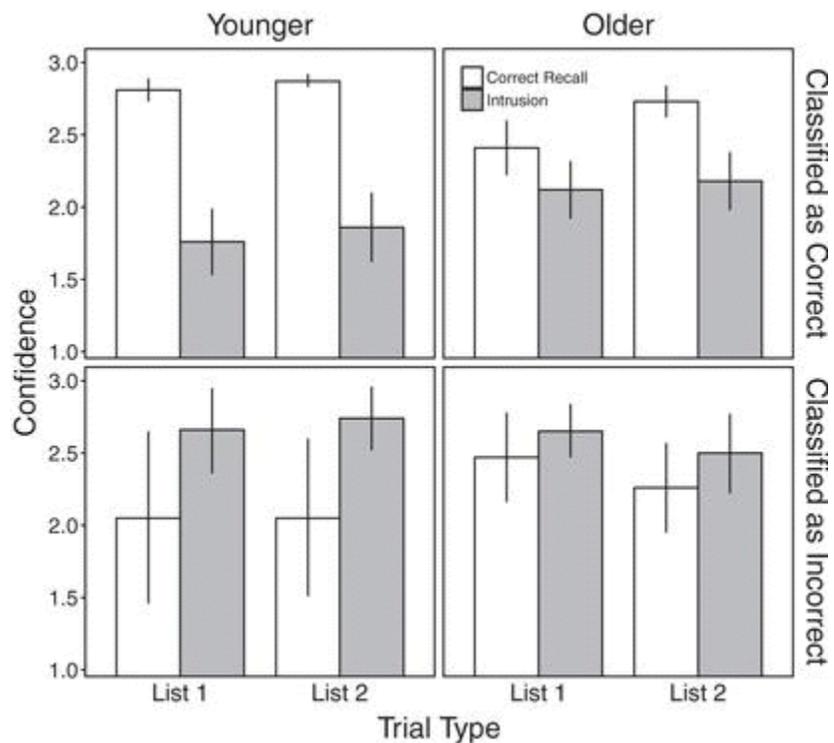
*Figure 4. Mean confidence judgments for accuracy classifications indicating that produced responses were correct as a function of age, classification, and trial type. Intrusions include both intra- and extratrial intrusions. Error bars are 95% confidence intervals.*

Confidence judgments for responses classified as correct (Figure 4, top panels) were first submitted to a 2 (Age: Younger vs. Older) × 2 (Trial: List 1 vs. List 2) × 2 (Response: Correct Recall vs. Intrusion) ANOVA. A significant Age × Response interaction, $F(1, 52) = 34.54$, $p < .001$, $\eta p^2 = .40$, showed that the difference between accurate and inaccurate classifications was greater for younger than older adults, thus indicating an age-related monitoring deficit. Although the Age × Trial × Response interaction was not significant, $F(1, 52) = 2.30$, $p = .14$, $\eta p^2 = .04$, visual inspection of the data suggested that these age differences depended on trial condition. We further explored these potential differences below.

We conducted separate 2 (List: List 1 vs. List 2) × 2 (Response: Correct Recall vs. Intrusion) ANOVAs for younger adults (top left panel) and older adults (top right panel). Younger adults showed no significant effect of list, $F(1, 24) = 0.99$, $p = .33$, $\eta p^2 = .04$; a significant effect of response, $F(1, 24) = 178.65$, $p < .001$, $\eta p^2 = .88$; and no significant List × Response interaction, $F(1, 24) = 0.06$, $p = .81$, $\eta p^2 < .01$. Older adults showed a significant effect of list, $F(1, 28) = 10.66$, $p = .003$, $\eta p^2 = .28$; a significant effect of response, $F(1, 28) = 35.74$, $p < .001$, $\eta p^2 = .56$; and a near-significant List × Response interaction, $F(1, 28) = 4.10$, $p = .05$, $\eta p^2 = .13$. These results showed that younger adults' monitoring accuracy was comparable in the List 1 and List 2 conditions, whereas older adults' monitoring deficit was greater in the List 1 than List 2 condition.

Confidence judgments for responses classified as incorrect (Figure 4, bottom panels) were first submitted to a 2 (Age: Younger vs. Older) × 2 (Trial: List 1 vs. List 2) × 2 (Response: Correct Recall vs. Intrusion) ANOVA. Similar to the analyses above, there was a marginal Age × Response interaction, $F(1, 22) = 3.89$, $p = .06$, $\eta p^2 = .15$, suggesting that younger adults' confidence judgments distinguished between correct recalls and intrusions to a greater extent. There was also a marginal Age × List interaction, $F(1, 22) = 3.51$, $p = .08$, $\eta p^2 = .14$, suggesting that older adults were more confident in the List 1 than List 2 condition, whereas younger adults did not differ in those conditions. As with the analyses above, we explored potential age differences in the effects of trial condition below.

Separate 2 (List: List 1 vs. List 2) × 2 (Response: Correct Recall vs. Intrusion) ANOVAs for younger adults (bottom left panel) and older adults (bottom right panel) revealed the following results. Younger adults showed a significant effect of response, $F(1, 9) = 12.13$, $p = .007$, $\eta p^2 = .57$; no significant effect of list, $F(1, 9) = 0.23$, $p = .65$, $\eta p^2 = .03$; and no significant List × Response interaction, $F(1, 9) = 0.26$, $p = .62$, $\eta p^2 = .03$. Older adults showed a significant effect of list, $F(1, 13) = 5.33$, $p = .04$, $\eta p^2 = .29$; no significant effect of response, $F(1, 13) = 2.44$, $p = .14$, $\eta p^2 = .16$; and no significant List × Response interaction, $F(1, 13) = 0.07$, $p = .79$, $\eta p^2 < .01$. Together, these results confirm older adults' monitoring deficit and show that they were more confident when classifying responses as incorrect in the List 1 than List 2 condition.

**Probability of First Recall**

The retrieval dynamics exhibited by younger and older adults for first-recalled items using the EFR-C procedure were examined to inform the issue of whether PFR curves obtained using standard recall instructions in earlier studies faithfully reflect the manner by which information comes to mind and the extent to which strategic reporting influences those functions. The idea here is that younger and older adults may sometimes differ in their reinstatement of context at the outset of recall, but these differences may be masked by selective reporting. If comparisons of younger and older adults' PFR curves in the present experiment are inconsistent with previous findings, this would suggest that these age groups initiate retrieval either more differently or more similarly than what has been concluded from the extant literature. These comparisons were conducted by conditionalizing first-recalled items from the List 1, List 2, and List Both conditions on the original input position from both study lists. These functions were smoothed by averaging across three adjacent positions for all except the first and last positions in each list. These data were submitted to separate 2 (Age: Younger vs. Older) × 20 (Position: 1 − 20) ANOVAs for each condition.

For the List 1 condition (see Figure 5), there were significant effects of age, $F(1, 58) = 4.14$, $p = .046$; and position, $F(19, 1102) = 21.70$, $p < .001$, $\eta p^2 = .27$; and a significant Age × Position interaction, $F(19, 1102) = 10.05$, $p < .001$, $\eta p^2 = .15$. These effects confirmed substantial differences in retrieval initiation patterns for younger and older adults. Specifically, younger adults initiated recall primarily from List 1 primacy positions and to a much lesser extent from the List 2 recency positions. In contrast, older adults initiated recall to a greater extent from List 2 recency than List 1 primacy positions. The pattern obtained for younger adults largely replicates the primacy effects in delayed recall shown earlier (e.g., Kahana et al.,

2002; Wahlheim & Huff, 2015), whereas the pattern obtained for older adults is remarkably distinct from earlier findings. These results suggest that younger adults were able to effectively reinstate the List 1 context at the outset of retrieval, whereas older adults were less able to shift their representations from the most recent list context to an earlier list context.
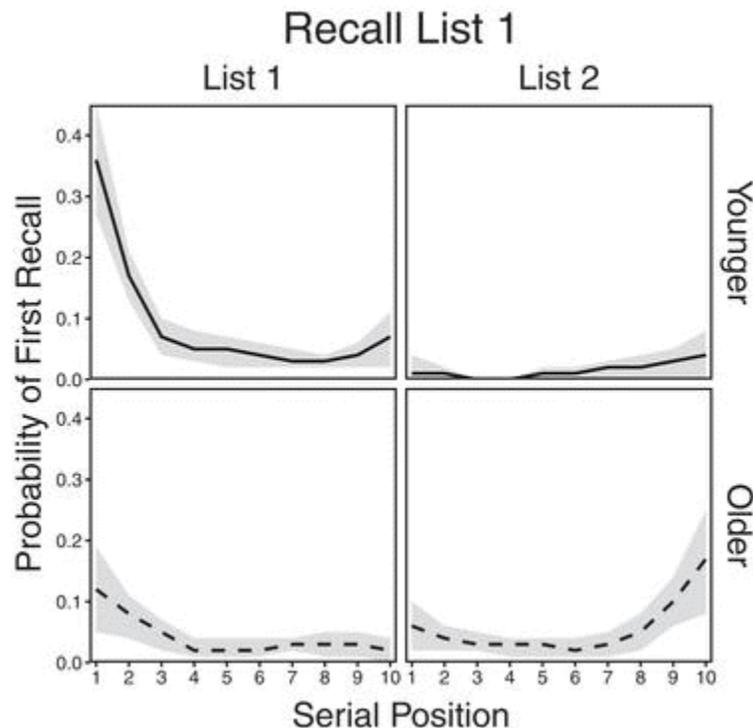


*Figure 5. Smoothed probability of first recall curves in the List 1 condition for younger and older adults. Shaded regions are 95% confidence intervals.*

For the List 2 condition (see Figure 6), there were significant effects of age, $F(1, 58) = 12.02$, $p = .001$, $\eta p^2 = .17$; and position, $F(19, 1102) = 94.54$, $p < .001$, $\eta p^2 = .62$; and a significant Age × Position interaction, $F(19, 1102) = 10.24$, $p < .001$, $\eta p^2 = .15$. These effects indicated that both younger and older adults initiated recall primarily from List 2 recency positions, but younger adults did so from earlier positions and also showed slight List 2 primacy effects. In addition, older adults initiated retrieval from List 1 primacy and recency on a few occasions, whereas younger adults never initiated retrieval from that list. The age differences in List 2 recency depart from earlier studies showing comparable effects on immediate recall tests (e.g., Kahana et al., 2002; Wahlheim & Huff, 2015). The present results suggest that younger adults were better able to use controlled processing to initiate retrieval from earlier positions in target lists when the recall period was initiated following a brief delay (3 s).
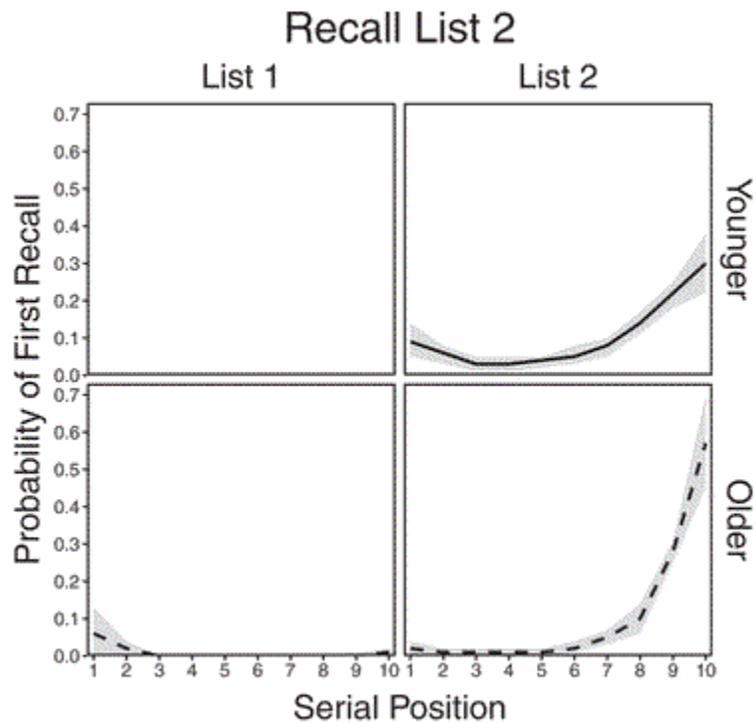
*Figure 6. Smoothed probability of first recall curves in the List 2 condition for younger and older adults. Shaded regions are 95% confidence intervals.*

For the List Both condition (see Figure 7), there was not a significant effect of Age, $F(1, 58) = 1.31$, $p = .26$, $\eta p^2 = .02$, but there was a significant effect of position, $F(19, 1102) = 67.14$, $p < .001$, $\eta p^2 = .54$; and a significant Age × Position interaction, $F(19, 1102) = 2.64$, $p < .001$, $\eta p^2 = .04$. These results indicated that both age groups initiated recall primarily from List 2, but younger adults did so from earlier list positions, as in the List 2 condition. Younger adults also showed a tendency to initiate recall from List 2 primacy positions, but to a lesser extent than observed by Wahlheim and Huff (2015). This difference in List 2 primacy between studies may have been because the break in context between lists here did not require active engagement by participants, whereas participants in the earlier study were required to press the space bar to begin studying List 2. Finally, both age groups sometimes initiated recall from List 1 primacy positions, but the extent to which older adults did so more often than younger adults was far less than shown earlier by Wahlheim and colleagues. These results show that the accessibility of responses when initiating recall from two lists is more similar between younger and older adults than earlier indicated by the stark differences shown under standard recall instructions.
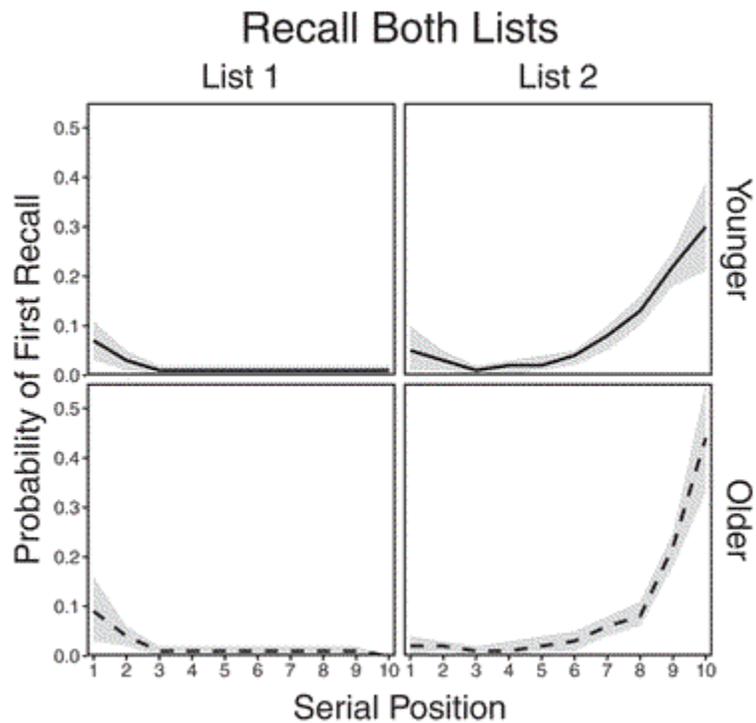
*Figure 7. Smoothed probability of first recall curves in the List Both condition for younger and older adults. Shaded regions are 95% confidence intervals.*

**Probability of First Recall (Classified as Correct)**

As noted above, younger and older adults typically show similar primacy effects on delayed tests in standard free recall (e.g., Kahana et al., 2002; Wahlheim & Huff, 2015). However, in the List 1 condition under EFR instructions in the present experiment, younger adults showed List 1 primacy effects (similar to patterns on delayed tests in standard recall), whereas older adults showed both List 1 primacy and List 2 recency effects. Together, these results suggest that PFR curves obtained in standard free recall may mask the extent to which older adults covertly edit List 2 responses that remain accessible as they attempt to reinstate List 1 context. To test this, we computed PFR classified as correct (PFR-C) curves for older adults in the List 1 condition (Figure 8, bottom panels) and compared them with the unconditionalized PFR curves in the same condition reported above (Figure 8, top panels). PFR-C curves were computed by conditionalizing the probabilities of first recalls classified as correct with medium or high confidence on input position.
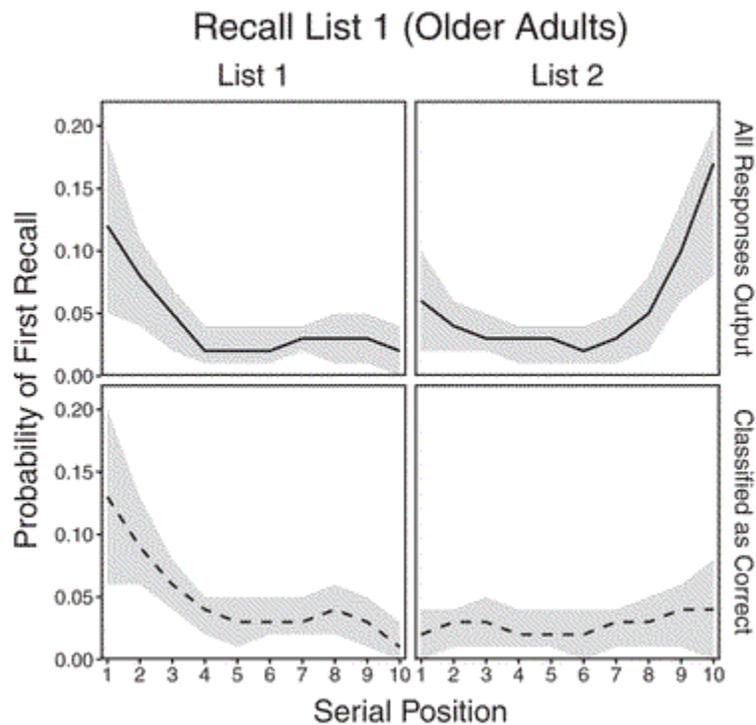
*Figure 8.* Smoothed probability of first recall curves for all responses (top panels) and smoothed probability of first recall classified as correct curves for responses given confidence judgments of 2 and 3 (bottom panels) in the List 1 condition for older adults only. Shaded regions are 95% confidence intervals.

Separate 2 (Measure: PFR vs. PFR-C) × 10 (Position: 1–10) ANOVAs were conducted for each list to determine whether older adults could identify that intratrial intrusions from List 2 were incorrect responses. The analysis of List 1 revealed a significant effect of position, $F(9, 522) = 11.21$, $p < .001$, $\eta p^2 = .16$; but there was neither a significant effect of measure, $F(1, 58) = 1.01$, $p = .32$, $\eta p^2 = .02$; nor a significant Measure × Position interaction, $F(9, 522) = 0.11$, $p = 1.00$, $\eta p^2 < .01$, showing that primacy effects did not differ between measures. In contrast, the analysis of List 2 revealed significant effects of measure, $F(1, 58) = 14.31$, $p < .001$, $\eta p^2 = .20$; and position, $F(9, 522) = 6.67$, $p < .001$, $\eta p^2 = .10$, along with a significant Measure × Position interaction, $F(9, 522) = 4.08$, $p < .001$, $\eta p^2 = .07$, showing that the recency effects obtained with the PFR measure were not obtained with the PFR-C measure. These results are consistent with the suggestion that older adults covertly edit List 2 responses that remain accessible when attempting to reinstate List 1 context in standard recall.

Note that comparable analyses of older adults' retrieval initiation were not conducted for the List 2 and List Both condition, because there was no need to edit the most accessible List 2 recency items. Thus, PFR-C curves would do little to further illuminate age differences in response production in the List 2 and List Both conditions.

**Output Profiles**

We extended our investigation of age differences in response output beyond the first-recalled responses by computing output probabilities across the entire recall period. We assume that response output under EFR instructions reveals age differences in the production and monitoring of responses that are masked by standard recall instructions. Although the design of the present experiment precluded a direct comparison between output profiles from EFR and standard recall instructions, we approximated this comparison by computing profiles for all responses output under EFR instructions and for only responses classified as correct with medium or high confidence (simulated standard recall). Output profiles were computed for each trial condition by averaging across participants the probabilities of producing List 1 responses, List 2 responses, and "Other" responses (extratrial intrusions and repeats of correct recalls) across output positions (Figures 9–11). Note that we collapsed repeats of earlier-output responses with extratrial intrusions because repeats occurred very infrequently (highest mean number per trial = 0.61). Separate 3 (Response: List 1 vs. List 2 vs. Other) × Position ANOVAs were conducted for each age group in each condition for each recall measure. The number of levels in the position variable differed across analyses based on when production appeared to end. The specific details of each ANOVA are listed below.
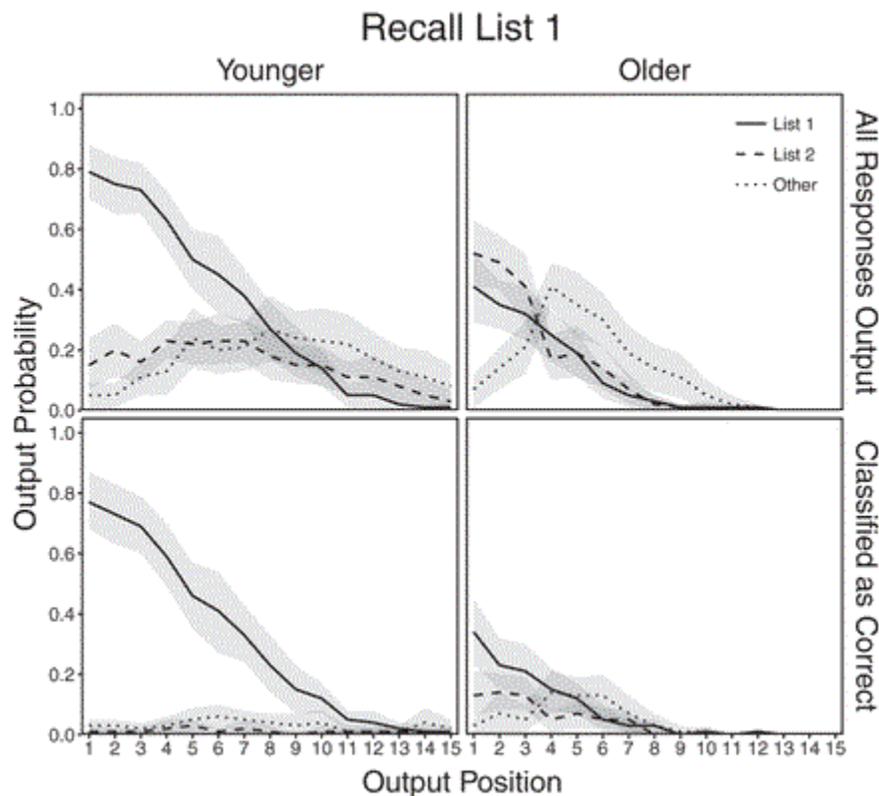


Figure 9. Mean probabilities of response output as a function of age, response type, and recall position for all responses output (top panels) and only responses classified as correct with confidence judgments or 2 or 3 (bottom panels) in the List 1 condition. Responses in the "Other" category include all extratrial intrusions and repetitions of earlier-output responses. Shaded regions are 95% confidence intervals.

## Recall List 2

*Figure 10.* Mean probabilities of response output as a function of age, response type, and recall position for all responses output (top panels) and only responses classified as correct with confidence judgments or 2 or 3 (bottom panels) in the List 2 condition. Responses in the "Other" category include all extratrial intrusions and repetitions of earlier-output responses. Shaded regions are 95% confidence intervals.
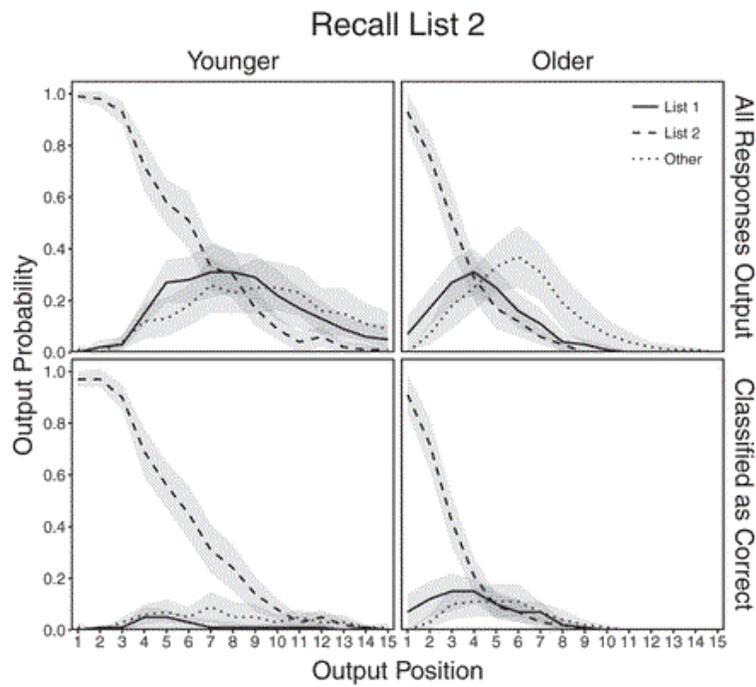


## Recall Both Lists

*Figure 11. Mean probabilities of response output as a function of age, response type, and recall position for all responses output (top panels) and only responses classified as correct with confidence judgments or 2 or 3 (bottom panels) in the List Both condition. Responses in the "Other" category include all extratrial intrusions and repetitions of earlier-output responses. Shaded regions are 95% confidence intervals.*
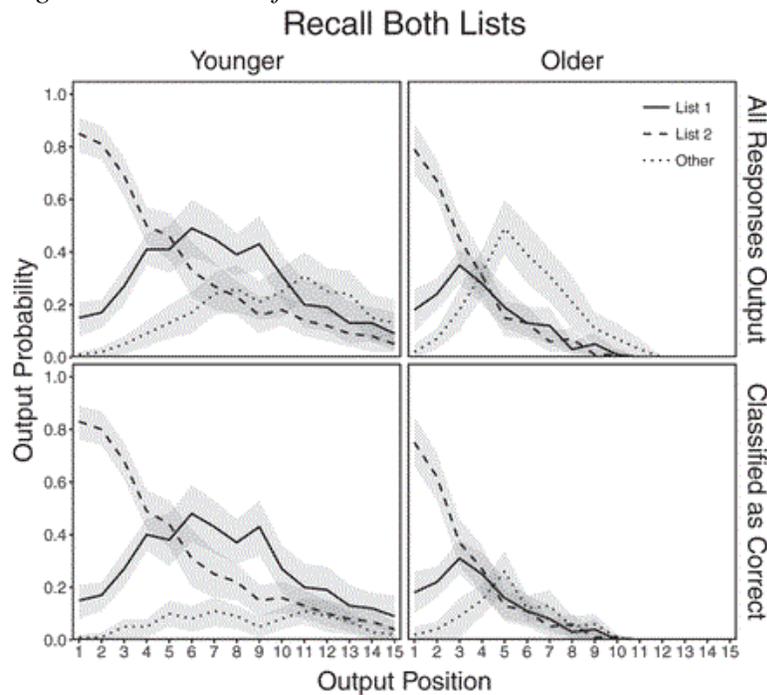


*Figure 9. Mean probabilities of response output as a function of age, response type, and recall position for all responses output (top panels) and only responses classified as correct with confidence judgments or 2 or 3 (bottom panels) in the List 1 condition. Responses in the "Other" category include all extratrial intrusions and repetitions of earlier-output responses. Shaded*

*Figure 10.* Mean probabilities of response output as a function of age, response type, and recall position for all responses output (top panels) and only responses classified as correct with confidence judgments or 2 or 3 (bottom panels) in the List 2 condition. Responses in the "Other" category include all extratrial intrusions and repetitions of earlier-output responses. Shaded
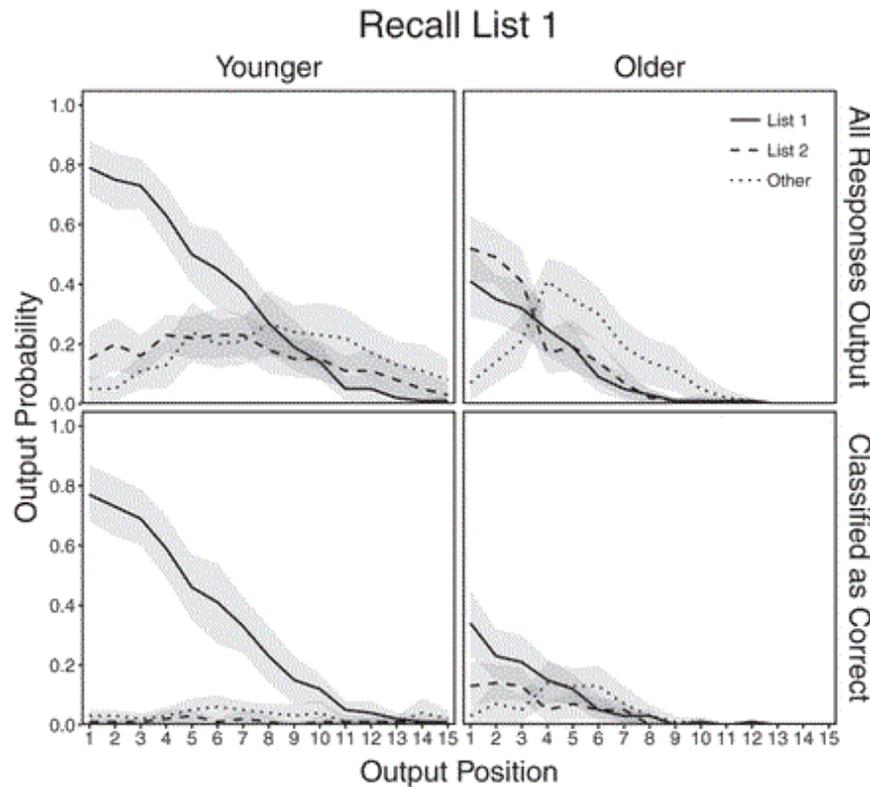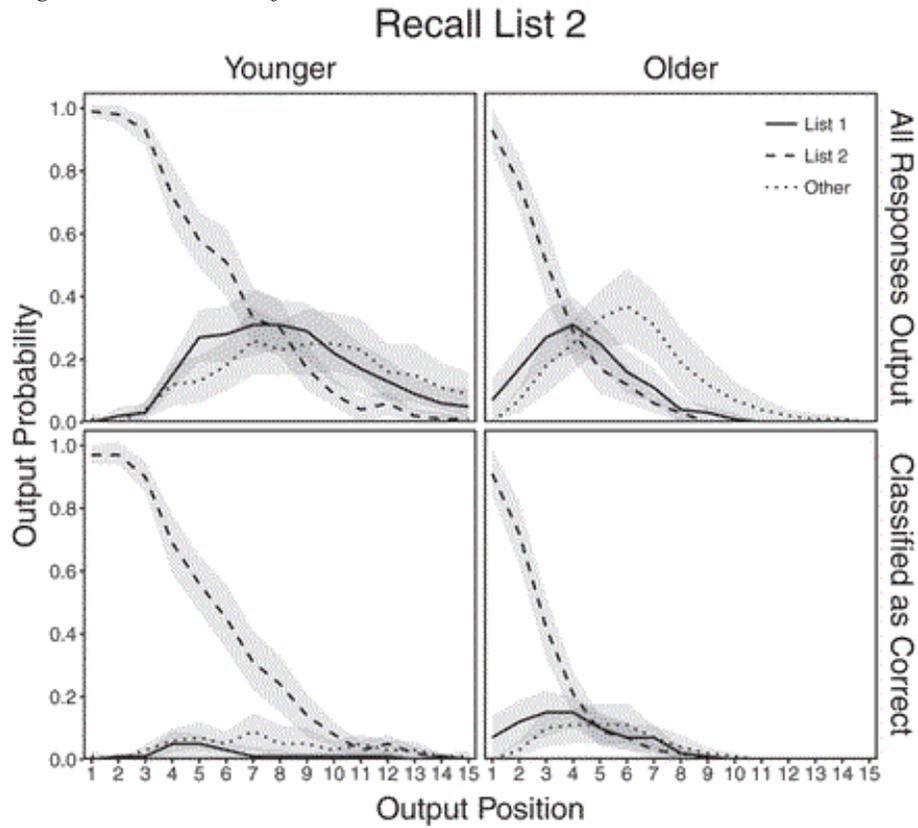
## Recall Both Lists



*Figure 11. Mean probabilities of response output as a function of age, response type, and recall position for all responses output (top panels) and only responses classified as correct with confidence judgments or 2 or 3 (bottom panels) in the List Both condition. Responses in the "Other" category include all extratrial intrusions and repetitions of earlier-output responses. Shaded regions are 95% confidence intervals.*

## List 1 condition

All responses produced by younger adults (Figure 9, top left panel) were examined using a 3 (Response) × 15 (Position) ANOVA. There were significant effects of response, $F(2, 58) = 16.99$, $p < .001$, $\eta p^2 = .37$; and position, $F(14, 406) = 104.42$, $p < .001$, $\eta p^2 = .78$, that were qualified by a significant Response × Position interaction, $F(28, 812) = 22.36$, $p < .001$, $\eta p^2 = .44$. These results show that correct recalls were produced most often during the initial portion of the recall period, intratrial intrusions were produced slightly more often than other responses during the initial recall period, and other responses were produced most often during the remaining portion of the recall period.
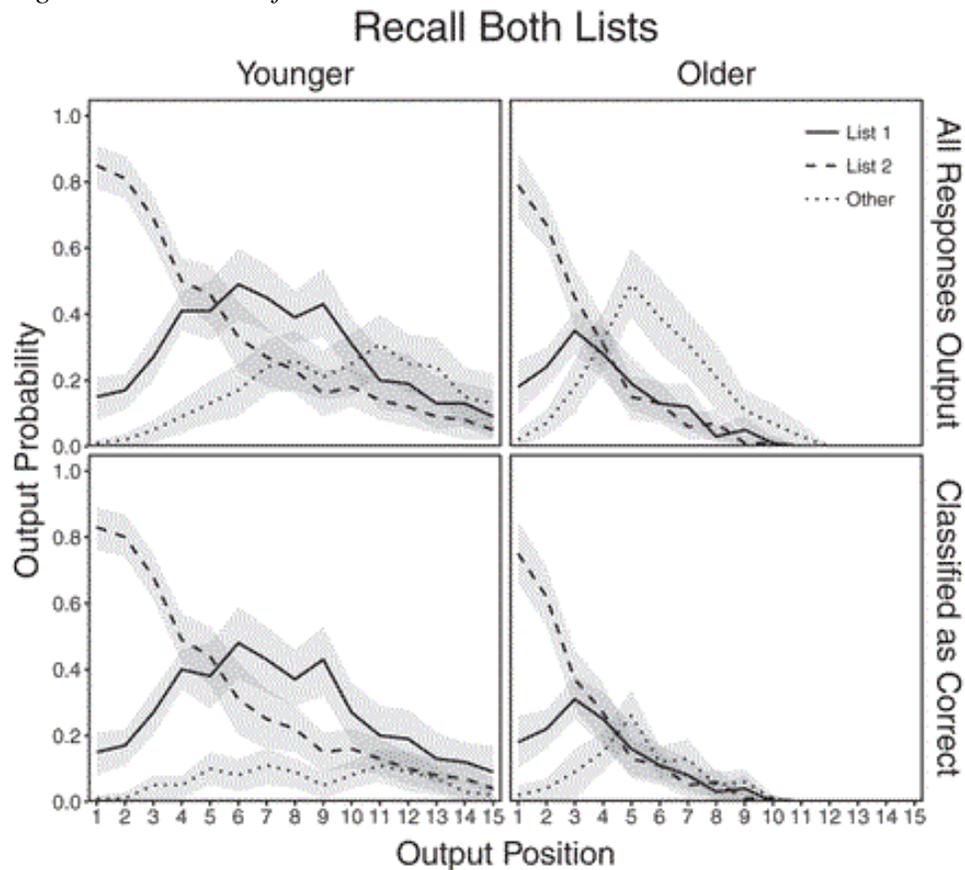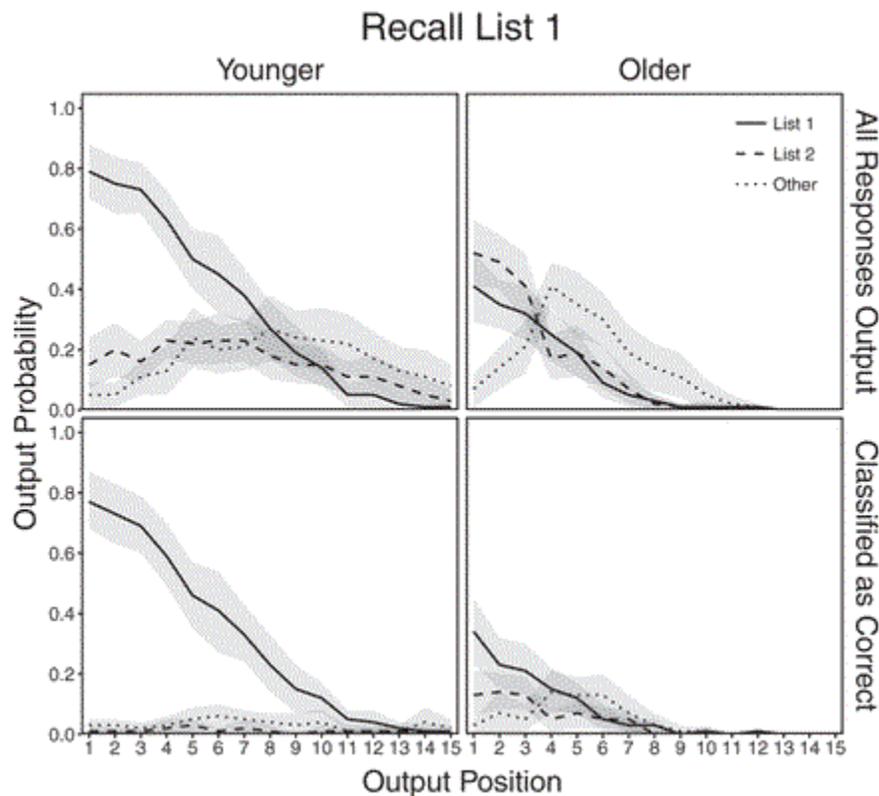
*Figure 9. Mean probabilities of response output as a function of age, response type, and recall position for all responses output (top panels) and only responses classified as correct with confidence judgments or 2 or 3 (bottom panels) in the List 1 condition. Responses in the "Other" category include all extratrial intrusions and repetitions of earlier-output responses. Shaded regions are 95% confidence intervals.*

Responses classified as correct by younger adults (Figure 9, bottom left panel) were examined using a 3 (Response) × 10 (Position) ANOVA. There were significant effects of response, $F(2, 58) = 136.05$, $p < .001$, $\eta p^2 = .82$; and position, $F(9, 261) = 44.53$, $p < .001$, $\eta p^2 = .61$, that were qualified by a significant Response × Position interaction, $F(18, 522) = 38.23$, $p < .001$, $\eta p^2 = .57$. These results show that most correct recalls were classified as such, whereas nearly all other responses were classified as incorrect.

All responses produced by older adults (Figure 9, top right panel) were examined using a 3 (Response) × 8 (Position) ANOVA. There was no significant effect of response, $F(2, 58) = 0.92$, $p = .41$, $\eta p^2 = .03$; but there was a significant effect of position, $F(7, 203) = 85.62$, $p < .001$, $\eta p^2 = .75$; and a significant Response × Position interaction, $F(14, 406) = 11.77$, $p < .001$, $\eta p^2 = .29$. In contrast to younger adults, these results showed that older adults were more likely to produce intratrial intrusions than correct recalls across the first several output positions before producing both of these responses types at similar declining rates throughout the remainder of the recall period. In addition, as the initial production of both types of intratrial responses

declined, the production of other responses increased sharply and were produced at higher rates than all other response types across the remainder of the recall period.

Responses classified as correct by older adults (Figure 9, bottom right panel) were examined using a 3 (Response) × 7 (Position) ANOVA. There were significant effects of response, $F(2, 58) = 8.09$, $p < .001$, $\eta p^2 = .82$; and position, $F(6, 174) = 11.22$, $p < .001$, $\eta p^2 = .28$, that were qualified by a significant Response × Position interaction, $F(12, 348) = 6.90$, $p < .001$, $\eta p^2 = .19$. These results show that despite greater production of intratrial intrusions than correct recalls during the first several output positions, older adults rejected most intratrial intrusions and classified most correct recalls as such. However, older adults showed poorer monitoring than younger adults as they rejected more correct recalls and output more of every other response type.

Taken with the results above, these results show that younger adults were better able to reinstate List 1 context and monitor production quality throughout the recall period. In addition, both age groups tended to produce intrusions from the more local intratrial context earlier during recall and intrusions from outside that context later in recall, showing that they relaxed their focus of retrieval across the recall period. Importantly, these results highlight the utility of the EFR approach for revealing qualitative age differences in response production in a retroactive interference situation that are masked under standard recall instructions.

**List 2 condition**

All responses produced by younger adults (Figure 10, top left panel) were examined using a 3 (Response) × 15 (Position) ANOVA. There were significant effects of response, $F(2, 58) = 42.34$, $p < .001$, $\eta p^2 = .59$; and position, $F(14, 406) = 102.95$, $p < .001$, $\eta p^2 = .78$, that were qualified by a significant Response × Position interaction, $F(28, 812) = 50.81$, $p < .001$, $\eta p^2 = .64$. As in the List 1 condition, these results show that correct recalls were produced most often throughout the initial recall period. However, relative to the List 1 condition response production for intratrial intrusions and other responses started later and increased more rapidly in the initial portion of recall. Similar to the List 1 condition, intratrial intrusions were produced more often than other responses earlier on, but this pattern showed a slight tendency to reverse as production declined throughout the remainder of the recall period.
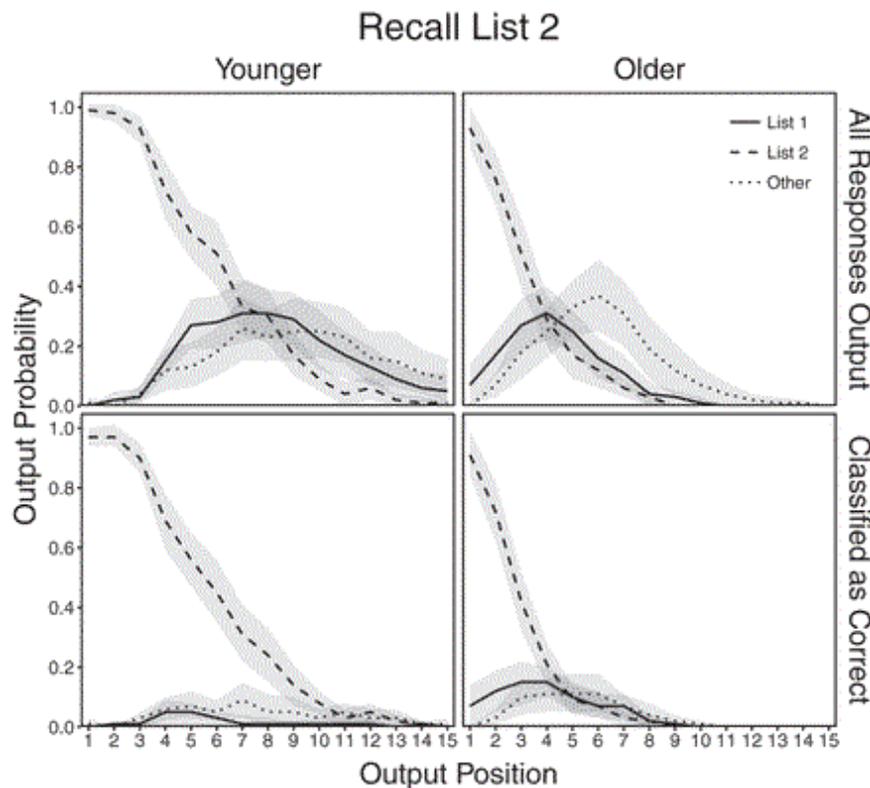
*Figure 10. Mean probabilities of response output as a function of age, response type, and recall position for all responses output (top panels) and only responses classified as correct with confidence judgments or 2 or 3 (bottom panels) in the List 2 condition. Responses in the "Other" category include all extratrial intrusions and repetitions of earlier-output responses. Shaded regions are 95% confidence intervals.*

Responses classified as correct by younger adults (Figure 10, bottom left panel) were examined using a 3 (Response) × 10 (Position) ANOVA. There were significant effects of response, $F(2, 58) = 457.61$, $p < .001$, $\eta p^2 = .82$; and position, $F(9, 261) = 111.68$, $p < .001$, $\eta p^2 = .79$, that were qualified by a significant Response × Position interaction, $F(18, 522) = 67.39$, $p < .001$, $\eta p^2 = .70$. These results show that, similar to the List 1 condition, younger adults classified nearly every correct recall as such and rejected nearly every other response type.

All responses produced by older adults (Figure 10, top right panel) were examined using a 3 (Response) × 8 (Position) ANOVA. There were significant effects of response, $F(2, 58) = 24.50$, $p < .001$, $\eta p^2 = .46$; and position, $F(7, 203) = 56.57$, $p < .001$, $\eta p^2 = .66$, that were qualified by a significant Response × Position interaction, $F(14, 406) = 34.79$, $p < .001$, $\eta p^2 = .55$. These results revealed a pattern similar to younger adults as correct recalls were produced most often in earlier output positions, whereas the production rate of intratrial intrusions and other responses were lower initially and increased across the early positions. Production of intratrial intrusions peaked after the first several positions and declined with correct recalls, whereas production of all other responses increased more slowly, peaked later than intratrial intrusions,

and remained higher than for both intratrial response types as they declined across the remainder of recall.

The production of responses classified as correct by older adults (Figure 10, bottom right panel) were examined using a 3 (Response) × 8 (Position) ANOVA. There were significant effects of response, $F(2, 58) = 88.77$, $p < .001$, $\eta p^2 = .75$; and Position, $F(7, 203) = 119.44$, $p < .001$, $\eta p^2 = .81$, that were qualified by a significant Response × Position interaction, $F(14, 406) = 48.66$, $p < .001$, $\eta p^2 = .63$. These results show that older adults could effectively reject most responses that were not correct recalls, but they still tended to accept more of these incorrect responses in early output positions than younger adults.

Together, these results show that younger and older adults were more comparable in their ability to reinstate the target-list context (List 2) across the recall period than in the List 1 condition, presumably due to its similarity with the time-of-test context. However, age-related deficits in this ability still remained and older adults relaxed their retrieval focus to a greater extent later in the recall period.

### List both condition

All responses produced by younger adults (Figure 11, top left panel) were examined using a 3 (Response) × 15 (Position) ANOVA. There were significant effects of response, $F(2, 58) = 23.45$, $p < .001$, $\eta p^2 = .45$; and position, $F(14, 406) = 57.14$, $p < .001$, $\eta p^2 = .66$, that were qualified by a significant Response × Position interaction, $F(28, 812) = 25.22$, $p < .001$, $\eta p^2 = .47$. These results show that when recalling from two lists, younger adults were most likely reinstate the most recent context (List 2) during the early portion of the recall period before shifting their focus to the List 1 context through the remainder of recall. Other responses were produced less often than both types of intratrial responses early in the recall period, but this pattern tended to reverse toward the end of recall.
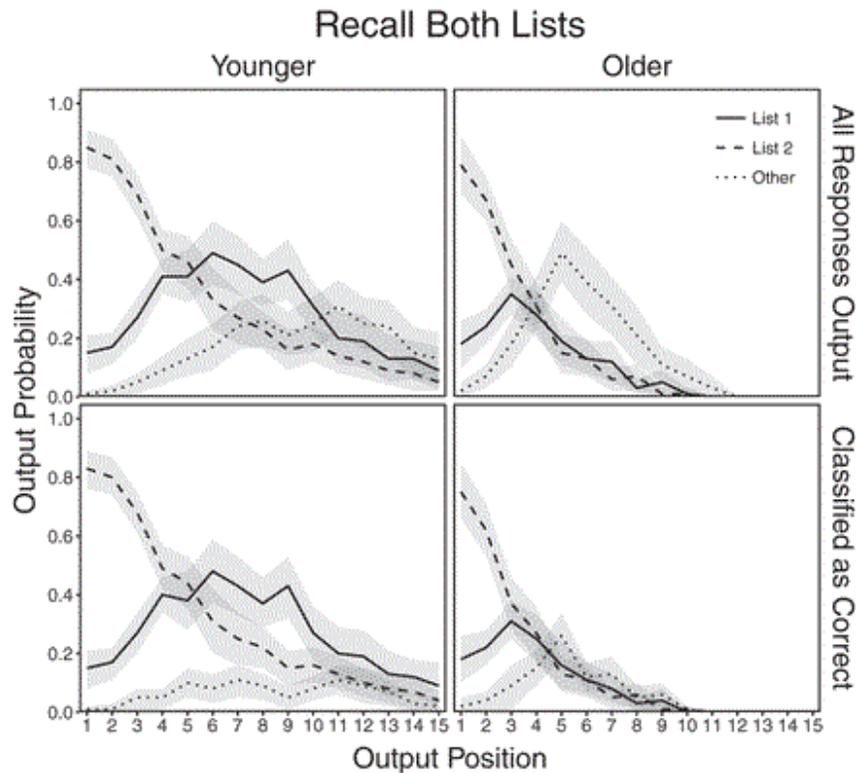
## Recall Both Lists



*Figure 11. Mean probabilities of response output as a function of age, response type, and recall position for all responses output (top panels) and only responses classified as correct with confidence judgments or 2 or 3 (bottom panels) in the List Both condition. Responses in the "Other" category include all extratrial intrusions and repetitions of earlier-output responses. Shaded regions are 95% confidence intervals.*

Responses classified as correct by younger adults (Figure 11, bottom left panel) were examined using a 3 (Response) × 15 (Position) ANOVA. There were significant effects of response, $F(2, 58) = 165.92$, $p < .001$, $\eta p^2 = .85$; and position, $F(14, 406) = 91.69$, $p < .001$, $\eta p^2 = .76$, that were qualified by a significant Response × Position interaction, $F(28, 812) = 23.90$, $p < .001$, $\eta p^2 = .45$. These results show that younger adults classified most correct recalls as such and rejected the majority of other responses.

All responses produced by older adults (Figure 11, top right panel) were examined using a 3 (Response) × 8 (Position) ANOVA. There were significant effects of response, $F(2, 58) = 10.20$, $p < .001$, $\eta p^2 = .26$; and position, $F(7, 203) = 52.08$, $p < .001$, $\eta p^2 = .64$, that were qualified by a significant Response × Position interaction, $F(14, 406) = 31.41$, $p < .001$, $\eta p^2 = .52$. These results show that similar to younger adults, older adults reinstated the recent list context (List 2) most often, but once they shifted their focus to List 1, production of both intratrial response types declined at the same rate. In contrast to younger adults, older adults' increasing production rate for other responses during the early portion of recall accelerated more rapidly, peaked earlier, and dropped more sharply. As in the List 1 and 2 conditions older adults were

also more likely to produce other responses than intratrial responses across the latter portion of recall.

Responses classified as correct by older adults (Figure 11, bottom right panel) were examined using a 3 (Response) × 9 (Position) ANOVA. There were significant effects of response, $F(2, 58) = 28.02$, $p < .001$, $\eta p^2 = .49$; and position, $F(8, 232) = 87.21$, $p < .001$, $\eta p^2 = .75$, that were qualified by a significant Response × Position interaction, $F(16, 464) = 27.06$, $p < .001$, $\eta p^2 = .48$. These results show that older adults classified most correct recalls as such, and rejected the majority of other responses, but the production rate of other responses was still comparable to rates for intratrial responses during the latter portion of recall.

## Discussion

The present experiment employed an EFR-C procedure in a dual-list free recall paradigm to provide a more complete characterization of adult age differences in response production and monitoring under conditions of proactive and retroactive interference. The following findings inform theoretical perspectives on the candidate processes proposed to underlie age-related differences in free recall. First, when participants reported all accessible responses while directing their retrieval to a specific list, older adults were impaired in their accuracy classifications, as they misclassified proportionally more responses than younger adults, especially intratrial intrusions in the List 2 condition. Second, older adults' confidence judgments in accuracy classifications discriminated more poorly between correct and incorrect responses, especially in the List 1 condition. Third, when reporting all productions, older adults initiated retrieval from nontarget lists more often than younger adults, especially when recalling from List 1 only. In addition, retrieval initiation in the List Both condition did not show large age differences, suggesting that the qualitative age differences shown in earlier studies by Wahlheim and colleagues were most likely due to strategic reporting. Finally, output profiles for all productions in the List 1 condition showed that List 2 context representations persisted across the first portion of recall to a greater extent for older than younger adults that was far more pronounced than shown in earlier studies. In addition, both groups relaxed their constraints across the recall period resulting in incorrect response production switching from intratrial to extratrial origins, and this occurred earlier and to a greater extent for older adults. We discuss these findings in turn below.

### Metacognitive Monitoring

Studies using both computational modeling and behavioral methods provide converging evidence that older adults' recall deficits in part reflect impaired monitoring (e.g., Healey & Kahana, 2016; Kahana et al., 2005; Wahlheim & Huff, 2015; Wahlheim et al., 2016). The present study was the first to show that older adults misclassified proportionally more correct recalls. In addition, the present results showing that older adults misclassified proportionally more intratrial intrusions than younger adults under conditions of proactive interference were consistent with earlier studies (e.g., Kahana et al., 2005; Wahlheim & Huff, 2015). However, the present experiment was the first to show that older adults misclassified proportionally more intratrial intrusions than younger adults in a retroactive interference situation. The age

differences in misclassifications were smaller in the List 1 than List 2 condition, suggesting that, contrary to our predictions, older adults were better able to reject intrusions under conditions of retroactive than proactive interference. However, examination of confidence judgments suggested that older adults had a greater overall monitoring deficit in the List 1 than List 2 condition, as their confidence in responses classified as correct distinguished more poorly between correct recalls and intrusions in the List 1 condition. This seemingly contradictory combination of results suggests that older adults were better able to reject intrusions in the List 1 than List 2 condition partly because they were more willing to classify productions in the List 1 condition as incorrect. Consistent with this, older adults also showed greater confidence in responses classified as incorrect in the List 1 than List 2 condition. Finally, more evidence for this classification bias was also shown by older adults misclassifying the most correct recalls (proportionally) in the List 1 condition. This bias may have resulted from older adults recalling few contextual details indicating List 1 membership and consequently being less likely to endorse responses as correct in that condition. Further research should examine the mechanisms underlying these differences in classification bias in greater depth.

A more general direction for future research would be to integrate perspectives from context-based models (e.g., Healey & Kahana, 2016; Lohnas, Polyn, & Kahana, 2015; Polyn et al., 2009) and the strategic regulation framework (e.g., Goldsmith, 2016; Halamish et al., 2012; Koriat & Goldsmith, 1996) to explain age differences in free recall. Both approaches propose that a preretrieval selection mechanism operates in the service of constraining retrieval to a target episode. However, context-based models do not specify the strategic role that metacognitive control processes play in selecting retrieval strategies and reinstating specific episodic elements in the same level of detail as the strategic regulation framework. Further, the strategic regulation framework offers a more nuanced description of postretrieval mechanisms that specify roles for cue-utilization in monitoring decisions and the ability to subsequently control the grain size of reporting. Considering these processes in the context of context-based models of free recall seems reasonable given that the contextual details that accompany recalls can vary in amount and quality and that individuals can flexibly adjust their report criteria.

**Retrieval Initiation**

Retrieval initiation patterns are often comparable for younger and older adults when participants recall items from individual lists (e.g., Kahana et al., 2002; Wahlheim & Huff, 2015). However, recent findings have shown qualitative age differences in retrieval initiation when participants recall from two distinct lists (Wahlheim & Huff, 2015; Wahlheim et al., 2016). These similarities and differences have been shown using standard recall instructions that allow participants to selectively report their retrievals. Consequently, extant characterizations of younger and older adults' patterns of retrieval initiation may reflect a combined influence of context reinstatement and strategic reporting decisions. The use of the EFR-C procedure here afforded the opportunity to examine the role of strategic reporting in these characterizations.

PFR curves for delayed tests often show comparable primacy effects for younger and older adults (e.g., Kahana et al., 2002; Wahlheim & Huff, 2015). This finding was replicated in younger adults EFR production in the List 1 condition in the present experiment. However, older

adults showed List 1 primacy and List 2 recency effects of similar magnitudes, suggesting that List 2 context representations persisted longer into their retrieval sequence than what could be inferred from standard recall results. Further evidence for this was found as older adults' PFR-C curves in the List 1 condition (which simulated standard recall) preserved List 1 primacy effects and eliminated List 2 recency effects, comparable to standard recall results. Together, these results provide a more comprehensive view of age differences in retrieval initiation in showing that older adults' deficit in context reinstatement was greater than originally inferred. These results also inform context-based computational models of interlist effects in free recall (e.g., Lohnas et al., 2015), as they point to the need to account for the effects of age on the rate of contextual drift from the most recent study list into the beginning of the recall period.

In contrast to delayed tests, PFR curves for immediate tests often show comparable recency effects for younger and older adults (e.g., Kahana et al., 2002; Wahlheim & Huff, 2015). In the present experiment, PFR curves in the List 2 condition showed recency effects akin to earlier studies, but younger adults initiated retrieval from earlier positions. This difference can be accommodated by a framework holding that individuals with greater working memory capacity can maintain access to more items at the end of a study list (Unsworth & Engle, 2007), as older adults generally show a working memory deficit (Bopp & Verhaegen, 2005). In addition, younger adults showed slight List 2 primacy effects, whereas older adults showed slight List 1 primacy and recency effects reflecting their impaired reinstatement of target-list context.

The qualitative age differences in PFR curves in the List Both condition in earlier studies (Wahlheim & Huff, 2015; Wahlheim et al., 2016) poses a theoretical challenge for extant models of free recall. Wahlheim and Huff (2015) originally found that younger adults had List 2 recency and smaller List 2 primacy effects, akin to findings from typical immediate tests, whereas older adults showed List 2 recency and List 1 primacy effects. These results were assumed to reflect older adults' broader context reinstatement, which was considered to parallel the more general retrieval orientation characteristic of individuals with executive control deficits (e.g., Burgess & Shallice, 1996; Moscovitch & Melo, 1997). However, further analyses along with results from two new experiments that replicated age differences in the variability of retrieval initiation (Wahlheim et al., 2016) suggested differences in strategic initiation.

The retrieval initiation patterns in the List Both condition in the present experiment confirm this suggestion, as the previous age differences in standard recall were not shown in EFR. Here, both age groups initiated their retrieval mostly from List 2 recency positions, and younger adults did so from earlier positions, presumably because of their working memory advantage. Both age groups also showed slight List 1 primacy effects, but the extent to which older adults did so was far less than observed in the earlier studies. Together, the results from this collection of studies shows that when older adults are required to alternate their retrieval among immediate and delayed tests across trials within an experiment, they sometimes strategically vary the list from which they begin reporting under standard recall instructions. In contrast, the EFR procedure revealed that the contextual representations for both age groups in the List Both condition during retrieval initiation are more similar than was earlier inferred. It is also noteworthy that the context breaks between study lists and between the study and test phase were not controlled by

participants in the present experiment, which may have diminished the salience of the breaks relative to the earlier studies. Overall, these results establish boundary conditions for the age differences in retrieval initiation in recall of hierarchically structured lists.

## Output Profiles

Examination of the entire retrieval sequence under EFR instructions provided a more complete characterization of response output when comparing retrieval dynamics for all responses produced with only those classified as correct with medium to high confidence (simulated standard recall). With the exception of the List 1 condition for older adults that showed a recalcitrance of List 2 context representations early during recall, the output profiles for all responses produced in the List 1 and List 2 conditions showed qualitatively similar patterns to those in Wahlheim and Huff (2015). However, the simulated standard profile for older adults in the List 1 condition also paralleled earlier results. Together, these findings bolster the validity of the EFR procedure for assessing production and monitoring operations.

The inclusion of "other" responses in output profiles further clarified age differences in context reinstatement. A finding common to the List 1 and List 2 conditions for both age groups was that intratrial intrusions were more accessible than extratrial intrusions across the initial portion of recall, whereas the reverse was true during the later portion. This finding suggests that participants' reinstatement of intratrial context diminished across recall, which may have resulted from self-initiated cue elaboration becoming less precise to increase the quantity of response candidates generated. Moreover, this pattern was especially pronounced for older adults, which could have reflected their greater attempt to increase production quantity.

## Limitations of Externalized Free Recall

Despite the obvious strengths of EFR for providing a clearer picture of response accessibility and covert editing in standard recall, some limitations should be considered. Most obvious, perhaps, is that requiring judgments between responses disrupts the natural organization of retrieval. Another limitation is that individuals and age groups may differ in their willingness and ability to report produced responses. This could result in adopting conservative report criteria to limit output of responses perceived as incorrect or adopting liberal report criteria to maximize memory quantity. Older adults may be more likely to exhibit these tendencies, perhaps in attempt to disconfirm stereotypes about age-related memory deficits. In addition, younger adults with low memory self-efficacy might adopt these reporting strategies to preserve the appearance of having socially acceptable memory abilities. Despite these concerns, output profiles were consistent across studies, suggesting that the present results validly inform the overall collection of results across studies.

## Conclusion

Age-related episodic memory deficits are especially pronounced in free recall under conditions of proactive and retroactive interference. The present experiment provided direct behavioral evidence that these deficits in part reflect older adults' impaired ability to produce and monitor retrievals. The retrieval initiation patterns and output profiles for all responses and simulated

standard recall provided more insight into role of strategic reporting in dual-list free recall, by suggesting that standard recall instructions partly mask the accessibility of responses in proactive and retroactive interference situations. Future studies should examine whether perspectives from context-based computational models and the strategic regulation framework can be integrated to provide a more comprehensive account of age-related deficits in free recall.

**References**

Anderson, J. R., & Bower, G. H. (1972). Recognition and retrieval processes in free recall. *Psychological Review*, *79*, 97–123. 10.1037/h0033773

Balota, D. A., Dolan, P. O., & Duchek, J. M. (2000). Memory changes in healthy young and older adults. In E.Tulving & F.Craik (Eds.), *Oxford handbook of memory* (pp. 395–410). Oxford, UK: Oxford University Press.

Bopp, K. L., & Verhaeghen, P. (2005). Aging and verbal memory span: A meta-analysis. *The Journals of Gerontology, Series B: Psychological Sciences and Social Sciences*, *60*, 223–233. 10.1093/geronb/60.5.P223

Bousfield, W. A., & Rosner, S. R. (1970). Free vs. uninhibited recall. *Psychonomic Science*, *20*, 75–76. 10.3758/BF03335608

Burgess, P. W., & Shallice, T. (1996). Confabulation and the control of recollection. *Memory*, *4*, 359–411. 10.1080/096582196388906

Ceci, S. J., & Tabor, L. (1981). Flexibility and memory: Are the elderly really less flexible?*Experimental Aging Research*, *7*, 147–158. 10.1080/03610738108259797

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology*, *33A*, 497–505. 10.1080/14640748108400805

Craik, F. I. M. (1968). Two components in free recall. *Journal of Verbal Learning and Verbal Behavior*, *7*, 996–1004. 10.1016/S0022-5371(68)80058-1

Craik, F. I. M. (1986). A functional account of age differences in memory. In F.Klix & H.Hagendorf (Eds.), *Human memory and cognitive capabilities, mechanisms and performance*(pp. 409–422). Amsterdam, the Netherlands: North-Holland and Elsevier.

Dodson, C. S., Koutstaal, W., & Schacter, D. L. (2000). Escape from illusion: Reducing false memories. *Trends in Cognitive Sciences*, *4*, 391–397. 10.1016/S1364-6613(00)01534-5

Dunlosky, J., & Metcalfe, J. (2009). *Metacognition*. Los Angeles, CA: Sage.

Epstein, W. (1969). Poststimulus output specification and differential retrieval from short-term memory. *Journal of Experimental Psychology*, *82*, 168–174. 10.1037/h0028045

Epstein, W. (1970). Facilitation of retrieval resulting from post-input exclusion of part of the input. *Journal of Experimental Psychology*, *86*, 190–195. 10.1037/h0029982

Goldsmith, M. (2016). Metacognitive quality-control processes in memory retrieval and reporting. In J.Dunlosky & S. K.Tauber (Eds.), *The Oxford handbook of metamemory* (pp. 357–385). New York, NY: Oxford University Press.

Halamish, V., Goldsmith, M., & Jacoby, L. L. (2012). Source constrained recall: Strategic control of production quality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*, 1–15. 10.1037/a0025053

Hartley, J. T., & Walsh, D. A. (1980). The effect of monetary incentive on amount and rate of free recall in older and younger adults. *Journal of Gerontology*, *35*, 899–905. 10.1093/geronj/35.6.899

Healey, M. K., & Kahana, M. J. (2016). A four-component model of age-related memory change. *Psychological Review*, *123*, 23–69. 10.1037/rev0000015

Hultsch, D. F. (1969). Adult age differences in the organization of free recall. *Developmental Psychology*, *1*, 673–678. 10.1037/h0028271

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. 10.1037/0278-7393.25.1.3

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

Jacoby, L. L., Shimizu, Y., Daniels, K. A., & Rhodes, M. G. (2005). Modes of cognitive control in recognition and source memory: Depth of retrieval. *Psychonomic Bulletin & Review*, *12*, 852–857. 10.3758/BF03196776

Jang, Y., & Huber, D. E. (2008). Context retrieval and context change in free recall: Recalling from long-term memory drives list isolation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 112–127. 10.1037/0278-7393.34.1.112

Kahana, M. J. (1996). Associative retrieval processes in free recall. *Memory & Cognition*, *24*, 103–109. 10.3758/BF03197276

Kahana, M. J., Dolan, E. D., Sauder, C. L., & Wingfield, A. (2005). Intrusions in episodic recall: Age differences in editing of overt responses. *The Journals of Gerontology, Series B: Psychological Sciences and Social Sciences*, *60*, 92–97. 10.1093/geronb/60.2.P92

Kahana, M. J., Howard, M. W., Zaromb, F., & Wingfield, A. (2002). Age dissociates recency and lag recency effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 530–540. 10.1037/0278-7393.28.3.530

Kelley, C. M., & Sahakyan, L. (2003). Memory, monitoring, and control in the attainment of memory accuracy. *Journal of Memory and Language*, *48*, 704–721. 10.1016/S0749-596X(02)00504-1

Keppel, G. (1968). Retroactive and proactive inhibition. In T. R.Dixon & D. L.Horton (Eds.), *Verbal behavior and general behavior theory* (pp. 172–213). Englewood Cliffs, NJ: Prentice Hall.

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review*, *103*, 490–517. 10.1037/0033-295X.103.3.490

Lohnas, L. J., Polyn, S. M., & Kahana, M. J. (2015). Expanding the scope of memory search: Modeling intralist and interlist effects in free recall. *Psychological Review*, *122*, 337–363. 10.1037/a0039036

Moscovitch, M., & Melo, B. (1997). Strategic retrieval and the frontal lobes: Evidence from confabulation and amnesia. *Neuropsychologia*, *35*, 1017–1034. 10.1016/S0028-3932(97)00028-6

Pansky, A., Goldsmith, M., Koriat, A., & Pearlman-Avnion, S. (2009). Memory accuracy in old age: Cognitive, metacognitive, and neurocognitive determinants. *The European Journal of Cognitive Psychology*, *2*, 303–329. 10.1080/09541440802281183

Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009). A context maintenance and retrieval model of organizational processes in free recall. *Psychological Review*, *116*, 129–156. 10.1037/a0014420

Raaijmakers, J. G. W. (2003). Spacing and repetition effects in human memory. *Cognitive Science*, *27*, 431–452. 10.1207/s15516709cog2703_5

Raaijmakers, J. G. W., & Shiffrin, R. M. (1980). SAM: A theory of probabilistic search of associative memory. In G. H.Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (pp. 207–262). New York, NY: Academic Press. 10.1016/S0079-7421(08)60162-0

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review*, *88*, 93–134. 10.1037/0033-295X.88.2.93

Rhodes, M. G., & Kelley, C. M. (2005). Executive processes, memory accuracy, and memory monitoring: An aging and individual difference analysis. *Journal of Memory and Language*, *52*, 578–594. 10.1016/j.jml.2005.01.014

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. 10.3758/BF03198437

Sahakyan, L., & Hendricks, H. E. (2012). Context change and retrieval difficulty in the list-before-last paradigm. *Memory & Cognition*, *40*, 844–860. 10.3758/s13421-012-0198-0

Schonfield, D., & Robertson, B. A. (1966). Memory storage and aging. *Canadian Journal of Psychology*, *20*, 228–236. 10.1037/h0082941

Shiffrin, R. M. (1970). Forgetting: Trace erosion or retrieval failure?*Science*, *168*, 1601–1603. 10.1126/science.168.3939.1601

Shipley, W. C. (1986). *Shipley Institute of Living Scale*. Los Angeles, CA: Western Psychological Services.

Stine, E. L., & Wingfield, A. (1987). Process and strategy in memory for speech among younger and older adults. *Psychology and Aging*, *2*, 272–279. 10.1037/0882-7974.2.3.272

Unsworth, N., Brewer, G. A., & Spillers, G. J. (2013). Focusing the search: Proactive and retroactive interference and the dynamics of free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 1742–1756. 10.1037/a0033743

Unsworth, N., & Engle, R. W. (2007). The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. *Psychological Review*, *114*, 104–132. 10.1037/0033-295X.114.1.104

Unsworth, N., Spillers, G. J., & Brewer, G. A. (2012). Evidence for noisy contextual search: Examining the dynamics of list-before-last recall. *Memory*, *20*, 1–13. 10.1080/09658211.2011.626430

Wahlheim, C. N., & Huff, M. J. (2015). Age differences in the focus of retrieval: Evidence from dual-list free recall. *Psychology and Aging*, *30*, 768–780. 10.1037/pag0000049

Wahlheim, C. N., Richmond, L. L., Huff, M. J., & Dobbins, I. G. (2016). Characterizing adult age differences in the initiation and organization of retrieval: A further investigation of retrieval dynamics in dual-list free recall. *Psychology and Aging*, *31*, 786–797. 10.1037/pag0000128

Ward, G., & Tan, L. (2004). The effect of the length of to-be-remembered lists and intervening lists on free recall: A reexamination using overt rehearsal. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 1196–1210. 10.1037/0278-7393.30.6.1196

Wixted, J. T., & Rohrer, D. (1994). Analyzing the dynamics of free recall: An integrative review of the empirical literature. *Psychonomic Bulletin & Review*, *1*, 89–106. 10.3758/BF03200763

Zacks, R. T., Hasher, L., & Li, K. Z. H. (2000). Human memory. In T.Salthouse & F.Craik (Eds.), *Handbook of aging and cognition* (2nd ed., pp. 293–357). Mahwah, NJ: Erlbaum.