DUAL-RETRIEVAL MODELS AND METAMEMORY IN YOUNGER AND OLDER ADULTS

A Thesis
Presented to the Faculty of the Graduate School
of Cornell University
In Partial Fulfillment of the Requirements for the Degree of
Master's in Human Development

by
Minyu Chang
December 2019

ABSTRACT

Accumulating evidence has suggested that older adults display a weaker correspondence between their retrospective confidence judgment and actual memory accuracy than younger adults, suggesting age-related declines in metamemory. However, the underlying mechanism for such age differences remains an open question. The primary goal of the study was revealing to what extent age-differences in verbatim- and gist-related memory processes could explain the age differences in metamemory accuracy. In this study, the verbatim and gist features of DRM word lists were factorially manipulated during encoding. Younger and older adults completed three consecutive free recall tests after encoding, and rated confidence for their recall accuracy. Fuzzy-trace theory's dual-retrieval model was used to estimate the two verbatim-related retrieval processes (direct verbatim access and forgetting) and two gist-related retrieval processes (reconstruction and familiarity judgments). The results show that metamemory accuracy was predicted by direct verbatim access, reconstruction and familiarity judgment at individual level for both younger and older adults. However, only direct access mediated the effects of age on metamemory accuracy, while other processes did not. This suggests that age-related decline in metamemory depends heavily on age-related difference in verbatim memory.

Keywords: Metamemory; aging; fuzzy-trace theory; dual-retrieval model

BIOGRAPHICAL SKETCH

The author, Minyu Chang, was born on October 1, 1995 in Hunan Province, China. She received a Bachelor of Social Science in Psychology from the University of Hong Kong in 2017, with a first-class honor. Currently, she is completing a MA/PhD program in the Human Development Department at Cornell University.

ACKNOWLEDGMENTS

I would like to express the deepest appreciation to my committee chair, Dr. Charles Brainerd, for his patient guidance and constructive suggestions for this research work. I would also like to thank my committee members: Dr. Valerie Reyna and Dr. Corinna Loeckenhoff, for their enthusiastic encouragement and helpful advice. Additionally, I would like to thank the undergraduate research assistants in the Memory and Neuroscience Lab, for their wonderful help in data collection and organization. Finally, a special thanks to my family members, who have supported me in everything along the way.

TABLE OF CONTENTS

CHAPTERS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1. Aging and metamemory

Metamemory refers to the knowledge of one's memory abilities, and the process of monitoring the contents of one's own memory (Flavell & Wellman, 1977; Nelson, 1990; Lovelace, 1990). In daily life, metamemory constantly influences people's choices and behaviors. For example, students spend less time reviewing materials that they have already remembered well, and drivers may turn off GPS if they are confident in their familiarity with the route. Metamemory is especially valuable for older adults, because they have more limited cognitive resources, and hence, they have a pressing need to strategically use those resources. If older adults can effectively monitor their memory accuracy, they will be able to adopt targeted mnemonic strategies for materials they cannot remember well, and thus optimize their memory performance (Dunlosky, Kubat-Silman, & Hertzog, 2003; Thiede, 1999).

There seems to be good news in that many studies have suggested that metamemory is largely preserved in aging. For example, memory complaints were shown to be a good predictor for age-related memory impairment in older adults, indicating that older adults had considerable self-awareness of their own memory status (Jonker, Geerlings, & Schmand, 2000). Also, older adults performed as well as younger adults in some prospective metamemory tasks, such as "judgment of learning" (JOL), which requires people to predict how likely they are to remember certain learning materials in later memory tests (Hertzog, Kidder, Powell-Moman, & Dunlosky, 2002 ; Hertzog & Dunlosky, 2011). Moreover, older adults could make comparable or even more accurate assessments for their accuracy on general knowledge tests relative to younger adults (Dahl, Allwood, & Hagberg, 2009; Pliske & Mutter, 1996).

However, some other studies have demonstrated that older adults are less adept at making retrospective confidence judgment for episodic memory accuracy than younger adults. Accumulating evidence for such age decline in retrospective metamemory has been found in cued-recall tests (Dodson, Bawa, & Krueger, 2007; Jacoby, Bishara, Hessels, & Toth, 2005; Kelley & Sahakyan, 2003; Meade & Roediger, 2006), recognition tests (Chua, Schacter, & Sperling, 2009; Fandakova, Shing, & Lindenberger, 2013; Gallo, Foster, & Johnson, 2009; Shing, Werkle-Bergner, Li, & Lindenberger, 2009; Wong, Cramer & Gallo, 2012), source memory tests (Dodson et al., 2007; Dodson, Bawa, & Slotnick, 2007), and misinformation paradigms (Karpel, Hoyer, & Toglia, 2001; Mitchell, Johnson, & Mather, 2003). Typically, many studies have observed that older adults have a greater tendency to assign high confidence judgments to false memories than younger adults (Dodson et al., 2007; Fandakova, et al., 2013; Jacoby et al., 2005; Mitchell et al., 2003; Shing et al., 2009).

If older adults do have more difficulty monitoring their own memory, this could cause serious consequences. For instance, it could provoke controversies in the trustworthiness of elderly eyewitnesses' testimony, considering that jurors tend to associate eyewitnesses' confidences with their actual memory accuracy (Wells, Ferguson, & Lindsay, 1981). Also, if older adult cannot accurately assess their own memory capacities, they may miss the opportunity for early detection of dementia. Therefore, it is crucial to understand why older adults have poorer retrospective metamemory than younger adults.

1.2. A dual-process perspective on age differences in metamemory

So far, the reasons for age differences in retrospective metamemory are still unclear. In order to better understand the underlying mechanism, it may be helpful to adopt a dual-process perspective. According to dual-process theories of memory, the ability to recollect specific details from previous experience (recollection) is impaired in aging, but the non-recollective operations, such as reconstructing experienced events based on gist information (Brainerd, Reyna, & Howe, 2009) or general feelings of familiarity (Jacoby & Rhodes, 2006; Koen & Yonelinas, 2014; Parkin & Walter, 1992), are relatively preserved.

Recently, some researchers have found that age-related changes in prospective metamemory mirrored the dual-process pattern found in objective memory performance. Moreover, some behavioral and neuroimaging research provided preliminary evidence that age decline in retrospective memory might also be associated with age differences in either recollective processes or non-recollective processes or both. Below I briefly review findings that are relevant to these two topics.

1.2.1. Dual processes and aging in prospective metamemory

Recently, some researchers have taken a dual-process approach to investigate age differences in prospective metamemory (i.e. JOL), and their findings nicely follow the dual-process pattern found in objective memory performance, including age-related deficits in recollection and age-related sparing of non-recollective processes (Daniels, Toth, & Hertzog, 2009; Kuhlmann & Undorf, 2018; Soderstrom, McCabe, & Rhodes, 2012).

In the study of Daniels et al. (2009), participants studied a series of words and rated JOL immediately after studying each word. Later, they took a recognition test and made recollect/familiarity judgments for each test cue, in which they responded "recollect" for words that they could remember specific details about and responded "familiar" for words that they felt familiar with but could not remember specific details about. The results showed that the difference in mean JOLs between recollected and familiar words was smaller in older adults than in younger adults, suggesting that older adults' JOLs were less able to discriminate recollected and familiar words than younger adults'. Also, although there were no age differences in the gamma correlation between JOLs and overall recognition accuracy, the JOL/accuracy gamma correlation for recollected items only was weaker in older adults than in younger adults. This reflects an age-related deficit specifically in recollection monitoring rather than overall monitoring.

Kuhlmann and Undorf (2018) reached the same conclusion when examining age differences in the relative JOL prediction accuracy using a process-dissociation procedure. Here, the process-dissociation procedure was based on Jacoby's (1991) inclusion/exclusion tests. As an illustration, in Kuhlmann and Undorf's (2018) Experiment 1, the to-be-remembered words were presented either auditorily or visually during the study phase. Later, in the inclusion test, participants were instructed to accept both *heard* and *seen* words, whereas in the exclusion test, they were required to accept the *heard* words only. In the former case, both recollection and familiarity support correct acceptance for the *seen* words, but in the latter case, recollection supports correct rejection of the *seen* words while familiarity supports incorrect acceptance of the *seen* words. Thus, by subtracting the probability of accepting the *seen* words in the exclusion test from that in the inclusion test, separate parameter estimates for recollection and familiarity could be obtained. The relative JOL prediction accuracy for recollection or familiarity was measured by the linear slope of recollection or familiarity parameters across different JOL levels. Kuhlmann and Undorf (2018) found that although younger and older adults had comparable relative JOL prediction accuracy for familiarity, the relative JOL prediction accuracy for recollection was lower in older adults than in younger adults. This echoes with the finding of Daniels et al. (2009) that familiarity monitoring remains intact in aging, but recollection monitoring declines with age.

To sum up, some recent findings have suggested that age differences in prospective metamemory may be closely associated with age differences in one of the dual processes: Recollection. This motivates further research on extending the application of dual-process theories from objective memory performance to metamemory performance.

1.2.2. Recollective processes and aging in retrospective metamemory

In the research for retrospective metamemory, some preliminary evidence converged on a view that age-related declines in retrospective metamemory are associated with the age-related deficiency in recollection (Dodson et al., 2007; Kelley & Sahakyan, 2003; Wong et al., 2012). In the studies reviewed below, retrospective metamemory accuracy is mainly measured by calibration error score, which is calculated by the absolute discrepancy between actual memory accuracy and the respective confidence level, weighted by the number of responses receiving that given confidence level (Dodson et al., 2007; Kelley & Sahakyan, 2003).

Dodson et al. (2007) found that older adults had lower metamemory accuracy than younger adults in source memory tests and cued-recall tests, but not in “old-new” item recognition or general knowledge tests. Noticeably, both source memory and cued-recall tests demand recollection of specific details about the to-be-remembered materials. Source memory tests require identification of the specific source of an item if it is judged to be old (e.g. who said what), and cued-recall tests required participants to recall the exact target word that was previously paired with the cue word. In contrast, correct responses on old-new item recognition or general knowledge tests do not necessitate recollection of specific details. The general knowledge test was essentially an old-new item recognition test for well-learned information, in which participants can simply base their old-new recognition decision on whether an item exceeds their familiarity threshold or not (Malmberg, 2008). Therefore, it seems that younger adults only outperform older adults in metamemory when the memory tests demand recollection of specific details.

Wong et al. (2012) manipulated the levels of recollective processing by using either pictures or words during the encoding phase, with pictures eliciting more detailed-oriented recollection than words. In the retrieval phase, they used a two-alternative forced choice (2FAC) recollection test, which required participants to identify the specific features of a studied item, such as whether an item was presented in a colored-picture or a line-drawing, or whether a word was presented in red or blue font. Their results showed that older adults had lower metamemory accuracy than younger adults. Younger and older adults' metamemory accuracies were both higher in the picture condition than in the word condition, and they were both positively correlated with accuracy in the recollection test. Based on the results, the authors argued that age-related declines in metamemory were strongly associated with the age-related declines in recollective capacity.

In addition, Kelley and Sahakyan (2003) found that older adults in a full-attention condition and younger adults in a divided-attention condition had comparable metamemory accuracy when their memory performance on a cued recall test was equated. Moreover, both groups had lower metamemory accuracy than younger adults in a full attention condition. The findings could be interpreted as supporting the recollective explanation of age differences in metamemory accuracy. Younger adults in a divided-attention condition lost their age-related advantage in metamemory, probably because their divided attention at encoding restricted the amount of details they recollected, and hence impaired their access to specific details of prior presentation at retrieval.

1.2.3. Non-recollective processes and aging in retrospective metamemory

Besides the growing evidence for the recollective explanation for age declines in metamemory, some other researchers showed evidence demonstrating that non-recollective processes also played an important role (Dennis et al., 2008; Jacoby, Debner, & Hay, 2001; Zheng, Lang, Wang, Xiao, & Li, 2019). For instance, in a functional magnetic resonance imaging (fMRI) study of Dennis et al. (2008), they found that older adults displayed higher activation in the middle temporal gyrus for high confidence false recognition than younger adults. Noticeably, the middle temporal gyrus is commonly seen as a semantic gist processing region (e.g., Thompson-Schill, Kan, & Oliver, 2006). Meanwhile, older adults also had lower activation in the hippocampus and higher activation in the retrosplenial cortex for high confidence true recognition than younger adults. Both hippocampus and retrosplenial cortex are recollection-related areas (Daselaar, Fleck, & Cabeza, 2006), which suggests that older adults' additional recruitment of retrosplenial cortex could be a compensation for age-related deficit in recollection. In conclusion, Dennis et al.'s (2008) findings indicate that older adults' worse metamemory could be associated with age-related increase in reliance on semantic gist as well as age-related declines in recollection.

In addition, Zheng et al. (2019) used the event-related potentials (ERP) in a categorized picture paradigm to examine the age-related increases in high confidence errors. They found that older adults displayed a late posterior negativity for false recognition, which is an ERP marker of reconstructing episodic details from prior experience, but younger adults did not. Meanwhile, they found no early frontal or early parietal old/new effects for false recognition in older adults, but they found early frontal old/new effects for false recognition in younger adults. Early frontal old/new effects are usually associated with general familiarity processes, and parietal old/new effects are an index of recollection for item-specific features (Rugg & Curran, 2007). In summary of Zheng et al.’s (2019) findings, familiarity processes were involved in younger adults’ false recognition, whereas reconstruction was involved in older adults’ false recognition. Since neither recollection nor familiarity was involved in older adults’ false recognition, reconstruction would be the most likely contributor to older adults’ high confidence false recognition.

1.2.4. Summary

Researchers who took a dual-process approach to study prospective metamemory in aging found that age-related declines in prospective metamemory was specifically constrained to recollection monitoring and that monitoring for non-recollective operations was relatively spared in aging. Such findings mirrored the age differences found in objective memory performance, and hence laid the foundation for applying dual-process theories to research on aging and metamemory.

Meanwhile, a series of behavioral and neuroimaging findings suggest that age-related declines in retrospective episodic metamemory are associated with age-related changes in recollective and/or non-recollective processing. These findings encourage more systematic investigations of age differences in metamemory using a dual-process approach. Noticeably, most studies reviewed above lack either separate assessment or factorial manipulation of recollective and non-recollective processes, which limits researchers’ abilities to draw causal inference about the roles of recollective and/or non-recollective operations in retrospective metamemory. Thus, it remains an open question to what extent the age differences in recollective and non-recollective processes can actually explain the age differences in retrospective metamemory.

1.3. Fuzzy-trace theory and dual-retrieval models

Fuzzy-trace theory (Reyna & Brainerd, 1995), as a dual-process theory for memory and reasoning, may help shed lights on the question raised above. According to Fuzzy-trace theory, people separately encode and store two types of memory traces: Verbatim traces of the detailed surface form of experience, and gist traces of the general meaning content. Verbatim traces support errorless recollection, which increases true memory and suppress gist-consistent false memory, whereas gist traces support non-recollective processing, which increases both true memory and gist-consistent false memory.

Based on fuzzy-trace theory, a family of dual-retrieval models was developed, which derives separate quantitative estimations for verbatim-related (recollective) and gist-related (non-recollective) retrieval processes from simple consecutive recall tests (Brainerd et al., 2009). Before this modeling approach, the traditional methods of dissociating recollective and non-recollective processes are usually based on recognition tests. Typical examples include the remember/know procedure (Tulving, 1985), inclusion/exclusion tests (Jacoby, 1991), confidence judgments accompanied by ROC curves (Yonelinas, 1994) and so on.

Critically, the traditional methods may not be appropriate for examining age differences in metamemory from a dual-retrieval respective. Firstly, most of them are recognition-based methods, which may not be able to effectively dissociate recollective and non-recollective processes, as Malmberg (2008) argued that recognition performance could be based on a single familiarity process alone. Since all metacognitive judgments are made based on the recognition performance, this would threat the validity of the method as a dual-process measurement. Secondly, many traditional methods, such as remember/know and confidence judgment procedures, involve self-reported metacognitive judgment. Such subjective report component can make age comparisons difficult, because different age groups may base their subjective reports on different processes. For example, it has been reported that older adults base their "remember" judgments more on emotional aspects or general familiarity, whereas younger adults rely more on perceptual details (Comblain, D'Argembeau, Van der Linden, & Aldenhoff, 2004; Norman & Schacter, 1997). More problematically, older adults are more likely to simply base their confidence judgment on information related to the recognition test cue, such as processing fluency or familiarity of cues (Busey, Tunnicliff, Loftus, & Loftus, 2000).

Dual-retrieval models, instead, do not have the limitations mentioned above, because they are pure recall-based techniques and require no subjective report. After recall data from at least three test cycles is collected, a two-stage Markov chain will be fit to the recall data to calculate the model parameters, which delivers quantitative estimates for verbatim- and gist-related retrieval processes (Gomes, Brainerd, Nakamura, & Reyna, 2014). Here, to get a better sense of the relationship between the model parameters and the underlying retrieval processes, please refer to Table 4 for explanations for dual-retrieval model parameters. It is also worth mentioning that, as prior research suggests, dual-retrieval models have consistently offered good fit to recall data, indicating high validity and reliability (Brainerd et al., 2014; Brainerd, Wright, & Reyna, 2012; Gomes et al., 2014).

In all versions of the dual-retrieval models, it is assumed that items are either recalled via a recollective process, called direct access, or a non-recollective process called reconstruction. Direct access refers to the process of directly accessing the verbatim traces of previously presented items. Such process guarantees perfectly correct recall, since people have the exact representation of studied items reinstated in their mind. Reconstruction occurs when there is no direct access to verbatim traces, and people need to actively reconstruct the studied items based on partially identifying information, typically semantic gist. Reconstruction is relatively more error-prone than direct access, because people may reconstruct gist-consistent non-studied items, which leads to false recall. In addition, reconstruction is always accompanied by a slave process called familiarity judgment, which allows people to decide whether to output a reconstructed item based on whether it has exceeded certain familiarity threshold or not (Brainerd et al., 2014). The mathematic machinery of the dual-retrieval model will be explained in greater depth in the Results section.

1.4. The current study

The current study aims to examine to what extent verbatim- and gist-related retrieval processes can explain age differences in metamemory accuracy. For the first time in this line of research, I factorially manipulated the verbatim and gist content of Deese/Roediger/McDermott (DRM; Deese, 1959; Roediger & McDermott, 1995) lists for both younger and older adults. In the verbatim manipulation, the list words were presented in either one standard font or multiple atypical fonts, which changed the levels of difficulty in accessing the verbatim traces for these words. In the gist manipulation, the words from different lists were either fully randomized or semantically blocked, which changed the levels of difficulty in processing the semantic relations among the words.

Another special feature of the current study is that I applied a mathematical modeling design (fuzzy-trace theory's dual-retrieval model; Gomes et al, 2014) to quantify the contributions of verbatim- and gist-related retrieval processes to metamemory accuracy. In such a design, participants first studied a series of word lists, and then completed three consecutive free recall tests. The recall data was then fitted to the dual-retrieval model to calculate separate parameters for verbatim- and gist-related retrieval processes. Meanwhile, after the three recall tests, participants were asked to make confidence judgments for their own recall, so that I could calculate the correspondence between confidence judgment and actual recall accuracy as the measure of metamemory accuracy.

In addition, I also measured two factors that can potentially impact metamemory performance: Depression and working memory, and both factors were treated as covariates in the data analyses. Depression is a prevalent mental disorder in older adults, and it has often been associated with increasing risk of dementia or cognitive decline (Geerlings et al., 2000; Jorm, 2000). Moreover, depression has been shown to increase memory complaints and mediate the relationship between memory complaints and actual memory performance (Kahn, Zarit, Hilbert, & Niedehrehe, 1975; Williams, Little, Scates, & Blockman, 1987), suggesting that it may influence metamemory accuracy. Meanwhile, given the well-established finding that working memory tends to decline with age (e.g., Daniels, Toth, & Jacoby, 2006), Wong et al. (2012) pointed out that age-related declines in working memory can potentially reduce older adults' accuracies in retrospective confidence judgments, because they may be less capable of comparing a recollected item with the other items stored in memory. To sum up, due to the potential effects of depression and age-related declines in working memory on metamemory judgments, it is important to control them in aging and metamemory research,

Four hypotheses were tested in the current study, which include: (a) older adults have lower metamemory accuracy than younger adults; (b) age differences in metamemory accuracy are smaller in conditions that enhance verbatim reliance and/or inhibit gist reliance; (c) verbatim and/or gist-related retrieval processes can predict metamemory accuracy at the individual level for both younger and older adults; (d) age differences in verbatim and/or gist-related retrieval processes could account for age differences in metamemory accuracy.

2. METHODS

2.1. Participants

Participants were 24 younger adults (aged 18 – 23 years, *M* = 19.63, *SD* = 1.28, 13 females) and 24 older adults (aged 66 – 96 years, *M* = 78.21, *SD* = 8.36, 18 females). The younger participants were undergraduate students recruited through the online SONA system at Cornell University, and the older participants were recruited from a participant pool retained by the Cornell Institute for Translational Research on Aging (CITRA), and from senior apartments and senior activity centers in the Tompkins County. Before the experiment began, screening tests were conducted over the telephone to exclude older participants who had cognitive impairment. In addition, all participants completed a depression scale and a questionnaire that helped us extract demographic information. In terms of compensation, the younger participants received extra course credits for their participation, and older participants received monetary reimbursement of 20 dollars.

2.2. Measures and Materials

2.2.1. Screening test for cognitive impairment. The Modified Telephone Interview for Cognitive Status (TICS-M; Brandt, Spencer, & Folstein, 1988; Welsh, Breitner, & Magruder-Habib, 1993) was used as a screening test for cognitive impairment during telephone recruitment. Prior research demonstrated that TICS-M had excellent sensitivity (82.4%) and specificity (87%) to the cognitive impairment diagnosis in older adults (Cook, Marsiske, & McCoy, 2009). The TICS-M consists of 12 items that assess orientation, learning, attention, calculation and language (Welsh et al., 1993). After the screening test, research assistants calculated each older participant's TICS-M score based on the scoring criteria. For the purpose of selecting participants that are cognitively normal, I chose a cutoff score of 32 (out of 50), as recommended by Knopman et al. (2010). Namely, any older adults who scored lower than 32 would not be recruited. A copy of the TICS-M is provided in Appendix A.

2.2.2. Measure for depression. The Geriatric Depression Scale Short Form (GDS; Brink et al., 1982) was used as a measure of depression for both younger and older adults. The reliability for the GDS short form has been well established in the literature for both older adults (e.g., Hermann et al., 1996) and younger adults (e.g., Ferrago & Chelminski, 1996). The form consists of 15 questions that inquire people's feelings over the past week. For each question, participants responded with either yes or no. A copy of the GDS is provided in Appendix B.

2.2.3. Demographic questionnaire. A short demographic questionnaire was designed to extract information about age, gender, race and education history (i.e., "How many years of formal education have you received?"). A copy of the demographic questionnaire is provided in Appendix C.

2.2.4. DRM word lists. Twenty 15-word DRM lists with the highest mean backward association strength (MBAS) were selected from the Appendix of Roediger, Watson, McDermott, & Gallo (2001), which should induce significant levels of false memory. Also, the selectin was done with the restriction that there is no overlap in words between different lists. The DRM lists are currently the most commonly used materials in false memory research. In such lists, all the list words are semantically related to a critical unpresented word (critical distractor). For example, the list for *chair* includes *table, sit, legs, seat, couch desk, recliner, sofa, wood, cushion, swivel, stool, sitting, rocking, bench*.

2.2.5. One-back task. The one-back task used in this study was adapted from the one used in Kane, Conway, Miura, and Colflesh (2007), which was programmed and administered in Psychopy2. This task functioned as a buffer task between the study and test phase and between the three test cycles, as well as a measure for working memory. Each one-back task contained 25 trials, among which five letters (B, K, M, Q, R) were presented, one at a time. Thus, each letter was presented for five times. In all the trials, a letter was visually presented on the computer screen (in 72-point white font against the grey background) for 0.5 second, and there was a 2-second interval between consecutive trials. In every trial after the first trial, participants were required to judge if the letter presented was the same as the letter one trial ago. If the letter was the same as the letter one trial ago, participants should press the "S" key for "same", otherwise they should press the "D" key for "different". In 40% of the trials, the letter presented was the same as the letter one trial ago (targets), and in the other 60% of the trials, the letter presented was different from the letter one trial ago (foils).

2.2.6. Recall and confidence judgment sheets. In each of the four conditions, participants completed three recall tests and one confidence judgment. Thus, they were provided with 12 recall sheets and four confidence judgment sheets in total. Each recall sheet consisted of 75 blanks for participants to write down the recalled words, and all blanks were numbered sequentially from 1 to 75. In correspondence to the recall sheets, each confidence judgment sheet contained 75 spaces that are numbered from 1 to 75. Also, in each space, there was a six-point scale (from 1 to 6) for confidence judgment.

2.3. Procedure

A 2 (verbatim: single, fonts) × 2 (gist: random, blocked) × 2 (age: younger, older) factorial design was used, with age being the between-subject factor, and verbatim and gist manipulations being the within-subject factor. The factorial manipulation of verbatim and gist gave rise to four conditions in total: Single-blocked, fonts-blocked, single-random and fonts-random conditions. The order of the four conditions was randomized across participants. Each condition comprised two phases: Study phase and test phase. In the study phase, participants studied five 15-word lists. In the test phase they completed three consecutive free recall tests, with each test preceded by a buffer task (one-back test), and then rated confidence for each recall on the third recall test.

The four conditions only differed in the five word lists presented in the study phase. In the single-blocked condition, all words were presented in a single standard font (Arial). Also, the words on the same DRM list were blocked together. Namely, the 15 words that comprised one DRM list were presented consecutively. In the fonts-blocked condition, the only difference from the single-blocked condition was that the 15 words on each DRM list were presented in 15 atypical fonts (an illustration can be found in Appendix D). In the single-random condition, all the 75 words were presented in a single standard font (Arial). Also, the 75 words (five DRM lists × 15 words/list) were fully randomized and regrouped into five new word lists, with the restriction that no more than 4 words from the same DRM list would be on a same new list. In the fonts-random condition, the 75 words were also fully randomized and regrouped into five new word lists, and the 15 words on each list were presented in 15 atypical fonts. To sum up, if we take the single-random condition as a baseline, the fonts manipulation should increase distinctiveness of each word and enhance verbatim processing (Arndt & Rreder, 2003), and the blocked manipulation should increase semantic association among list words and enhance gist processing (Payne, Elie, Blackwell, & Neuschatz, 1996; Toglia, 1999).

In the study phase of each condition, 75 words were visually presented on the computer screen (centered in 72-point white font against the grey background), with a 10-s interval between the consecutive word lists. The order of the lists was randomized across participants, while the order of the words within a list was fixed for all participants. As an attempt to control the recall accuracy between younger and older adults, I adopted a slower presentation rate for older adults (2000 ms/word) than for younger adults (700 ms/word), because it has been found that true memory increases and false memory decreases with longer presentation duration beyond 250 ms (McDermott & Watson, 2001).

After all the 75 words were presented in the study phase, participants entered the test phase. There were three cycles in the test phase. In each cycle, they first completed a buffer task (one-back task), and then a free recall test. To make sure participants understand the instructions of the one-back task, the first one-back task in this experiment was a practice, in which participants got feedback in every trial. In all of the subsequent one-back tasks, no feedback was provided. After the buffer task, participants were provided with a recall sheet, and they were instructed that they had at most three minutes to write down as many words as they can recall from the five lists they just studied. Before the recall test began, they were explicitly told that correct spellings were not necessary. After participants completed the three test cycles, they were provided with a confidence sheet along with their recall sheet for the third recall test, so they can refer to the words they recalled on the third test when making confidence judgment for each word recalled. Participants were instructed to rate their confidence by circling a number from 1 to 6 (1 = extremely unconfident, 2 = moderately unconfident, 3 = somewhat unconfident, 4 = somewhat confident, 5 = moderately confident, 6 = extremely confident). Participants were given at most one minute for the confidence judgment.

3. RESULTS

In what follows, I first report the demographics data, and then discuss the results from the analysis of variance (ANOVAs), model analyses, bivariate correlation analyses, regression analyses and mediation analyses. First of all, three-way ANOVA analyses were conducted for participants' objective memory performance, confidence judgments, and metamemory accuracy. The ANOVA for metamemory accuracy was meant to test the first two hypotheses, namely whether there were age differences in metamemory accuracy, and whether such age differences could be moderated by verbatim and/or gist manipulations. Then, I conducted model analyses to obtain quantitative estimations for verbatim- and gist-related retrieval processes and used the individual-level parameter estimations in the correlation, regression and mediation analyses. The correlation analyses revealed the relationships between individual parameters and metamemory accuracy. Based on that, the regression analyses were intended to test the third hypothesis: Whether the verbatim- and gist-related retrieval processes could predict metamemory accuracy at individual level, and the mediation analyses were used to test the fourth hypothesis: Whether they could explain age differences in metamemory accuracy.

3.1. Demographic data

The demographic data for younger and older adults were summarized in Table 1. Two-sample *t* tests showed that the younger and older adults did not differ in years of formal education received, $t(46) = -.39$, $p = .70$, or in self-reported depression scores, $t(39) = .26$, $p = .79$, whereas younger adults had significantly higher one-back accuracy than older adults, $t(26) = 2.72$, $p = .01$. Chi-tests showed that the there was no age difference in gender, $X^2 = 1.46$, $p = .22$, but the race composition of older adult sample (95.8% White, 4.2% unknown) was significantly different from that of younger adults (41.7% White, 41.7% Asian, 16.6% Hispanic or Latino), $X^2 = 20.12$, $p = .001$.

Table 1
Basic Demographic Information

	Age	Gender ratio	Education	Cognitive status	Working memory	Depression
Younger	19.63 (1.28)	1.18:1	14.16 (.96)	--	.93 (.05)	3.46 (2.86)
Older	78.21 (8.36)	3:1	16.71 (3.25)	38.02 (4.22)	.83 (.16)	3.17 (4.56)

Notes. Standard deviations are shown in the parentheses. Gender ratio is the ratio of females to males. Education is measured by the years of formal education received. Cognitive status was obtained from the Modified Telephone Interview for Cognitive Status (TICS-M), which was used as a screening test for cognitive impairment for older adults. Working memory was measured by the accuracy of one-back task. Depression was assessed with the Geriatric Depression Scale (GDS) short form.

3.2. ANOVA results

The dependent variables in ANOVA analyses were mean recall accuracy, mean intrusion of critical distractors (CDs), mean confidence, and calibration error scores. Because the confidence judgments were only obtained for recalls on the third recall test, for consistency, all the variables mentioned here refer to those for the third recall test only. Mean recall accuracy was calculated as the proportion of actual correct recall out of total possible correct recall (the number of words actually presented). Mean intrusion from critical distractors was calculated as the number of critical distractors falsely recalled divided by the total number of critical distractors. As mentioned above, the calibration error score is the weighted average of the deviation between actual recall accuracy and the respective confidence level (Dodson et al., 2007; Kelley & Sahakyan, 2003). Thus, a calibration error score of 0 would indicate perfect calibration, while higher calibration error scores would suggest poorer calibration. For convenience of interpretation, I adopted a similar method as Dodson et al. (2007) and transformed the 1-6 confidence scale into a 6-point predicted accuracy scale (0, 20%, 40%, 60%, 80%, 100%).

The descriptive statistics for mean recall accuracy, mean intrusion from critical distractors, mean confidence and mean calibration error scores are reported in Table 2. A visual inspection of Table 2 suggests that younger adults seemed to have higher mean recall accuracy, higher intrusion rate and lower calibration error scores than older adults, which was confirmed in the ANOVA results reported below.

Table 2

Mean Recall Accuracy, Mean Intrusions from Critical Distractors, Mean Confidence, and Mean Calibration Error Scores as a Function of Age and Condition

		Mean recall accuracy	Mean intrusion from CDs	Mean confidence	Mean calibration error scores
Younger adults					
	Single-blocked	.26 (.11)	.48 (.28)	5.24 (.48)	.16 (.11)
	Fonts-blocked	.28 (.11)	.45 (.32)	5.32 (.53)	.16 (.08)
	Single-random	.22 (.10)	.23 (.22)	5.21 (.48)	.15 (.07)
	Fonts-Random	.17 (.08)	.29 (.21)	5.11 (.42)	.17 (.09)
Older adults					
	Single-blocked	.13 (.09)	.31 (.26)	4.88 (1.24)	.26 (.16)
	Fonts-blocked	.20 (.11)	.27 (.25)	5.30 (1.05)	.22 (.11)
	Single-random	.10 (.07)	.12 (.17)	5.12 (1.15)	.26 (.16)
	Fonts-Random	.10 (.09)	.17 (.21)	5.12 (1.13)	.22 (.17)

Note. Standard deviations are shown in the parentheses. CDs = critical distractors

In addition, the overall distribution of confidence judgment was displayed in Table 3, which was meant to examine whether there were age differences in polarization of confidence judgment. Namely, whether younger or older adults had a higher tendency to give extreme values (Dodson et al., 2007). Two-sample *t* tests showed that the older adults made more extremely low confidence judgment of 1 than younger adults, $t(98) = 2.56$, $p = .01$, while they made less moderately low confidence judgment of 2 than younger adults, *t*(187) = -3.02, *p* = .003. However, younger and older adults did not differ in the overall use of the other level of confidence judgments (3, 4, 5, or 6), *p*s > .05. To sum up, older adults tended to give more extreme confidence judgment on the low end than younger adults, while younger adults gave more moderately low confidence judgment than older adults, but they did not differ in the tendency to give high confidence judgment.

Table 3
Mean Proportion of Trials in Each Confidence Level

	Confidence judgment					
	1	2	3	4	5	6
Younger	.01 (.03)	.05 (.07)	.08 (.08)	.10 (.09)	.09 (.10)	.67 (.17)
Older	.07 (.22)	.02 (.06)	.06 (.11)	.10 (.13)	.09 (.12)	.66 (.30)

Notes. Standard deviations are shown in the parentheses.

For each of the four dependent variables mentioned above, a 2 (verbatim: single, fonts) × 2 (gist: random, blocked) × 2 (age: younger, older) mixed ANOVA and post hoc tests were conducted. Before each ANOVA analysis, a visual inspection of the diagnostic plots was conducted to examine whether the assumptions for ANOVA analyses were fulfilled. Also, all results for post hoc tests were adjusted for multiple comparisons using the Tukey's honest significant difference test (HSD). The ANOVA and post hoc analyses were conducted using the lme4 package (Bates, Maechler & Bolker, 2012) and lmerTest package (Kuznetsova, Brockhoff, & Christensen, 2017) in R.

3.2.1. Recall accuracy

As for main effects, there was a gist manipulation (blocked vs. random) effect, *F*(1, 138) = 82.82, *MSE* = .23, partial $\eta^2 = .38$, $p < .001$, and an age effect, *F*(1, 46) = 16.51, *MSE* = .05, partial $\eta^2 = .54$, $p < .001$. The blocked conditions (M = .22, SD = .12) increased recall accuracy compared to the random conditions (M = .15, SD = .10), and younger adults (M = .23, SD = .11) had higher recall accuracy than older adults (M = .14, SD = .10).

In addition, there was a Verbatim × Gist interaction, *F*(1, 138) = 20.13, *MSE* = .06, partial η^2 = .13, *p* < .001, and a Verbatim × Age interaction, *F*(1, 138) = 10.57, *MSE* = .03, partial η^2 = .07, *p* = .001. Post hoc tests revealed that the effects of verbatim manipulation (fonts vs. single) was moderated by the gist effects and age effects. The difference between the fonts-blocked and single-blocked conditions (*M* = .24, *SD* = .12 vs. *M* = .20, *SD* = .12) was significant, *p* < .001, while that between the fonts-random and single-random conditions (*M* = .14, *SD* = .09 vs. *M* = .16, *SD* = .10) was not, *p* = .22. Also, older adults had higher recall accuracy in the fonts condition (*M* = .15, *SD* = .11) than in the single conditions (*M* = .12, *SD* = .08), *p* = .003, but younger adults' accuracy did not differ between the fonts and single conditions (*M* = .23, *SD* = .11 vs. *M* = .24, *SD* = .10), *p* = .72. These results suggest that the verbatim effects were more robust with the blocked lists than with the randomized lists, and they were stronger for older adults than for younger adults.

3.2.2. Intrusion from critical distractors

A main effect of gist manipulation was found, F(1, 138) = 39.03, MSE = 1.47, partial η^2 = .38, *p* < .001, and so was a main effect of age, F(1, 46) = 8.20, MSE = .31, partial η^2 = .54, *p* = .006. Post hoc tests revealed that older adults (M = .22, SD = .21) had less intrusions from critical distractors than young adults (M = .35, SD = .26), and the blocked conditions (M = .38, SD = .29) elicited higher intrusions than the random conditions (M = .20, SD = .21). This suggests that older adults had less semantic false memory than younger adults in the current study. This finding was unexpected, since many prior studies have found that older adults have higher recall for semantically related distractors than younger adults (for a review, see Gallo, 2006). In addition, the finding that false memory is higher in the blocked conditions than in the random conditions suggests that the gist manipulation is effective in enhancing gist processing.

3.2.3. Confidence

The visual inspection of the diagnostic plots revealed an obvious deviation from homoscedasticity: The residuals of confidence became more spread out as the fitted values decreased. In this case, data transformation was needed before conducting the ANOVA analysis. However, in this special case, the common log transformation would only make the situation worse. Thus, I divided confidence scores by six (which restricted confidence to a range between 0 and 1) and conducted a logit transformation, which successfully rectified the heteroscedasticity problem. The logit transformation was conducted using the car package in R (Fox et al., 2013), which automatically restricted the pre-transformed values to a range of .025 and .975. Thus, data points with the confidence judgment of six were preserved for the later ANOVA analysis.

The ANOVA results showed that there was neither a main effect of age nor a main effect of verbatim or gist manipulation. However, there was a Verbatim × Gist interaction, $F(1, 138) = 4.37$, $MSE = 1.93$, partial $\eta^2 = .03$, $p = .04$. Post hoc tests suggested that the mean confidence difference between the fonts-blocked condition and the single-blocked condition ($M = 5.31$, $SD = .82$ vs. $M = 5.06$, $SD = .95$) was approaching significance, $p = .06$, whereas the difference between the fonts-random condition and the single-random condition ($M = 5.12$, $SD = .84$ vs. $M = 5.17$, $SD = .87$) was not statistically significant at all, $p = .98$.

3.2.4. Calibration error scores

The visual inspection of the diagnostic plots revealed that there was again a violation of homoscedasticity assumption: The residuals of calibration error scores became more spread out as the fitted values increased. Accordingly, I conducted a commonly used log transformation and fixed the problem.

The results show that there was a main effect of age, *F*(1, 46) = 10.70, *MSE* = .11, partial $\eta^2 = .18$, *p* = .002. Post hoc tests revealed that younger adults (*M* = .16, *SD* = .09) had lower calibration error scores than older adults (*M* = .24, *SD* = .15), *p* = .002, suggesting that younger adults overall had higher metamemory accuracy than older adults. This result demonstrates that older adults had worse metamemory accuracy than younger adults. However, no other main effect or interaction was found, which suggests that the verbatim and gist manipulations in this study did not moderate the age differences in metamemory accuracy.

3.3. Modeling results

In this section, I first explain the five expanded versions of the dual-retrieval model, which are referred to as error, success, hybrid, learning and forgetting models respectively in the following text. Then, I report the modeling results, which includes five parts: Goodness-of-fit tests, parameter estimates, hypothesis testing on parameter estimates, recall decomposition, and hypothesis testing on recall compositions. Firstly, the goodness-of-fit tests were conducted for each of the five models. Secondly, based on goodness-of-fit results, I chose the model that delivered the best fit and conducted parameter estimate analyses using the given model. Both goodness-of-fit tests and parameter estimates analyses were conducted using the MPTinR package in R (Singmann & Kellen, 2013). Once parameter estimates were done, I examined whether there were age and manipulation effects on the model parameters. After that, I used the parameter estimates to conduct recall decomposition, which separated recall performance into proportions of recollective and non-recollective recall. Last, I tested whether there were age and manipulation effects on different compositions of recall.

3.3.1. Explanations of the dual-retrieval models

In dual-retrieval models, it is assumed that an item can occupy either an unlearned state (U), a non-recollective partially learned state (including a sub-state P_C for successful output after reconstruction and another sub-state P_E for failed output after reconstruction), or a recollective learned state (L). When there is only one test trial, an item could occupy different states with different probabilities: (a) the probability of occupying the L state is D; (b) the probability of occupying the P_C state, in which the item is reconstructed and successfully recalled, is $(1-D)RJ$; (c) the probability of occupying the P_E state, in which the item is reconstructed but not successfully recalled, is $(1-D)R(1-J)$; (d) the probability of occupying the U state is $(1-D)(1-R)$. Here, D represents direct access to verbatim traces, R represents reconstruction based on gist traces, and J represents familiarity judgment for reconstructed memory (Gomes et al., 2014). When there is more than one test trial, there could be transitions between these states across trials. Different versions of dual-retrieval models have different assumptions for such transitions.

The error, success and hybrid models all contain six parameters: $D_1, D_2, R_1, R_2, J_1, J_2$. The subscripts denote which trial the parameters are referring to. For example, D_1 is the probability that the verbatim traces of an item's prior presentation can be directly accessed on the first trial, and D_2 is the probability the verbatim traces of an item's prior presentation can be directly accessed on the second or third trial if they were not directly accessed on the first trial. The same logic applies to R_1 and R_2 as well as J_1 and J2. The error model was constructed under the assumption that transition to L in one trial can happen only after error in the previous trial. The success model, instead, suggests that transition to L can occur only after success in the previous trial, and the hybrid model suggests that transition to *L* can happen after either success or error in the prior trial (for a review, see Gomes et al., 2014).

The learning model contains five parameters: D_1, D_2, D_3, R, J. Noticeably, in this model there are three direct access parameters ($D_1, D_2,$ and D_3) separately for the first, second, and third trial, but there are only one overall reconstruction parameter (*R*) and one overall familiarity judgment parameter (*J*) for all three trials. The mathematical machinery for the learning model can be found in the Appendix of Brainerd, Payne, Wright, & Reyna (2003). The forgetting model contains six parameters: D, R, F, J_1, J_2, J_3. *D* and *R* represent the probabilities of direct access and reconstruction in all three trials. Importantly, a forgetting parameter (*F*) is added, which indicates that forgetting of verbatim traces (i.e., a transition from *L* to *U*) is possible on the second and/or third trial.

3.3.2. Goodness-of-fit tests and parameter estimates

To test the model fits and calculate parameter estimates, I first counted frequencies for the eight possible error-success sequences across the three recall tests, namely CCC, CCE, CEC, CEE, ECC, ECE, EEC, EEE (C = correct, E = erroneous). After the frequency data was obtained, I tested model fits for the five versions of dual-retrieval models using goodness-of-fit tests, and conducted parameter estimates using the model that delivered the best fit. The model fit tests and parameter estimations were carried out at two levels: Group level and individual level. For the group level, the frequency data was aggregated across subjects. For the individual level, the frequency data was calculated within each participant. Thus, compared to group-level analyses, individual-level analyses allow for different parameter values across participants.

As mentioned above, there are eight empirical probabilities, and the error, success, hybrid and forgetting models have six parameters, whereas the learning modes has five parameters.

Therefore, I was able to do goodness-of-fit tests with one degree of freedom for the error, success, hybrid and forgetting model, and with two degrees of freedom for the learning model. Thus, the critical value of the G^2 statistic to reject the null hypothesis of fit at the .05 level was 3.84 for the first four models, and 5.99 for the last model. The results showed that only the forgetting model gave a satisfactory account for the data at group level, with a mean $G^2(1) = 2.92$ when averaged across all conditions and both age groups. Furthermore, the forgetting model also gave an excellent account for the data at individual level. The mean $G^2(1)$ was 0.42 when averaged across all conditions and all participants. The full details for this model are explained in Appendix E, and this model's six parameters are defined in Table 4. The model fit statistics and parameter estimates for the group-level model analyses are summarized in Table 5.

Table 4
Definitions for the Dual-Retrieval Model's Parameters

Parameters	Definitions
D	Direct access/recollection: The probability that the verbatim trace of an item's presentation can be directly accessed on a recall test
R	Reconstruction: The probability that an item can be reconstructed on a recall test when the verbatim trace of the item's presentation cannot be accessed.
F	Forgetting for direct access: The probability that the direct access works in the first recall test but fails in the following recall tests
J_1, J_2, J_3	Familiarity judgment: The probability that a reconstructed item is judged to be familiar enough to output. J_1, J_2, J_3 represent the familiarity judgment for the first, second and third recall test.

Table 5

Estimates for the Dual-Retrieval Model's Parameters at Group Level

	Single-blocked		Fonts-blocked		Single-random		Fonts-random	
	Younger	Older	Younger	Older	Younger	Older	Younger	Older
$G^2(1)$	.13	7.74	.03	2.48	4.35	.28	1.70	2.27
D	.20	.09	.23	.13	.20	.08	.14	.08
R	.14	.12	.11	.16	.06	.07	.06	.07
F	.10	.3	.08	.12	.08	.27	.08	.16
J_1	.61	.6	.56	.6	.42	.75	.53	.63
J_2	.72	.77	.73	.64	.56	.75	.73	.67
J_3	.77	.62	.84	.70	.78	.69	.73	.62

3.3.3. Age effects and manipulation effects on model parameters

Firstly, I conducted an omnibus test for all the individual-level model parameters, using a 2 (verbatim: single, fonts) × 2 (gist: random, blocked) × 2 (age: younger, older) multivariate analysis of variance (MANOVA). The MANOVA results reveal an effect of age, $F(6, 41) = 9.67$, $p < .001$, an effect of gist manipulation $F(6, 133) = 18.27$, $p < .001$, a Verbatim × Gist interaction, $F(6, 133) = 4.71$, $p < .001$, and a Verbatim × Age interaction, $F(6, 133) = 2.47$, $p < .05$.

Secondly, to examine how each underlying retrieval process varies with verbatim/gist manipulations and age, a 2 (verbatim) × 2 (gist) × 2 (age) mixed ANOVA was conducted for each individual-level model parameter. The ANOVA results are reported as follows.

Direct access (D). There was an age main effect, $F(1, 46) = 29.30$, $MSE = .09$, partial $\eta^2 = .47$, $p < .001$, a gist manipulation main effect, $F(1, 138) = 15.86$, $MSE = .05$, partial $\eta^2 = .10$, p

< .001, and a Verbatim × Gist interaction, *F*(1, 138) = 15.03, *MSE* = .05, partial η^2 = .10, *p* < .001. Post hoc tests revealed that the *D* parameter was significantly higher for younger adults (*M* = .17, *SD* = .08) than for older adults (*M* = .09, *SD* = .07), *p* < .001, suggesting that younger adults had better verbatim memory than older adults. Also, the *D* parameter ordering for the four conditions was fonts-blocked (*M* = .16, *SD* = .10) > single-blocked (*M* = .13, *SD* = .09) = single-random (*M* = .13, *SD* = .08) = fonts-random (*M* = .10, *SD* = .07). Here, we can see the *D* parameter was overall higher in the blocked conditions than the random conditions, which was mainly because it was the highest in the fonts-blocked condition among the four conditions. This result is consistent with the prior ANOVA results for recall accuracy.

Reconstruction (R). There was a gist main effect, *F*(1, 138) = 50.47, *MSE* = .19, partial η^2 = .27, *p* < .001, and a Verbatim × Age interaction, *F*(1, 138) = 4.19, *MSE* = .02, partial η^2 = .001, *p* = .04. Post hoc tests revealed that the *R* parameter was higher for the blocked lists (*M* = .23, *SD* = .09) than for the randomized lists (*M* = .17, *SD* = .05). Such results suggest that the gist manipulation (blocked vs. random) was effective in enhancing and reducing gist processing, which was consistent with the prior ANOVA results for intrusions from critical distractors. Meanwhile, the *R* parameter between the fonts and single conditions in older adults (*M* = .21, *SD* = .10 vs. *M* = .19, *SD* = .08) was slightly larger than that in younger adults (*M* = .19, *SD* = .06 vs. *M* = .19, *SD* = .08). However, such difference was not significant in either younger or older adults.

Forgetting (F). There was an age main effect, *F*(1, 46) = 26.58, *MSE* = 1.76, partial η^2 = .17, *p* < .001, a verbatim manipulation main effect, *F*(1, 138) = 4.95, *MSE* = .33, partial η^2 = .04, *p* = .03, and a Verbatim × Age interaction, *F*(1, 138) = 4.19, *MSE* = .02, partial η^2 = .001, *p* = .04. Post hoc tests revealed that older adults (*M* = .31, *SD* = .34) had higher *F* parameter than younger adults (*M* = .12, *SD* = .13), $p < .001$. Also, the *F* parameter was higher in the single conditions (*M* = .26, *SD* = .31) than in the fonts conditions (*M* = .17, *SD* = .24). However, although older adults had higher *F* parameter in the single conditions (*M* = .39, *SD* = .37) than in the fonts conditions (*M* = .23, *SD* = .30), the *F* parameter did not differ between the single (*M* = .12, *SD* = .13) and fonts (*M* = .12, *SD* = .14) conditions for younger adults. The results suggest that older adults had a higher tendency to forget items which they could initially get direct access to. Also, older were more forgetful in the single conditions than in the fonts conditions, probably because the different fonts made the words more distinctive from each other, and hence less forgettable. However, the fonts condition did not lower younger adults' *F* parameters compared to the single condition, probably because younger adults already had very low forgetting rate in the single condition.

Familiarity judgments (J_1, J_2, J_3). No significant main effect or interaction was found for the J_1 parameter. For the J_2 parameter, a gist manipulation main effect was found, $F(1, 138) = 4.30$, $MSE = .09$, partial $\eta^2 = .03$, $p = .04$. For the J_3 parameter, there was an age main effect, $F(1, 46) = 4.74$, $MSE = .07$, partial $\eta^2 = .04$, $p = .03$, and a gist manipulation main effect, $F(3, 138) = 6.19$, $MSE = .09$, partial $\eta^2 = .04$, $p = .01$. Post hoc tests revealed that the J_2 and J_3 parameters were both higher for the two blocked conditions (*M* = .60, *SD* = .15 and *M* = .61, *SD* = .14, respectively) than the two random conditions (*M* = .56, *SD* = .13 and *M* = .57, *SD* = .11, respectively), *p*s < .05. Also, the J_3 parameter was significantly higher for younger adults (*M* = .61, *SD* = .13) than older adults (*M* = .57, *SD* = .12). These results demonstrate that compared to older adults, younger adults were more likely to successfully output a word if they had reconstructed it based on gist, and both younger and older adults enrolled in more familiarity judgments when gist processing was enhanced.

3.3.4. Recall decomposition

Based on the model parameter estimates, recall performance can be decomposed to recollective recall, in which recalled items occupied the L state, and non-recollective recall, in which the recalled items occupied the P_C state (Gomes, 2013). Via algebraically transforming the Equations A1 – A9 in Appendix E, it is easy to know that the proportions of recollective recall on test 1, test 2, and test 3 can be calculated as D, $D(1 - F)$ and $D(1 - F)$, respectively, and the proportions of non-recollective recall on test 1, test 2, and test 3 could be calculated as $(1 - D)RJ_1$, $(1 - D)RJ_2$, $(1 - D)RJ_3$, respectively.

3.3.5. Age and manipulation effects on recollective and non-recollective recall

Similar to the analyses above, I used 2 (verbatim: single, fonts) × 2 (gist: random, blocked) × 2 (age: younger, older) ANOVAs for proportions of recollective and non-recollective recall on each test and also the mean proportion recollective and non-recollective recall.

Proportion of recollective recall on test 1. Because the expression for the proportion of recollective recall on test 1 is simply D, readers can refer to the ANOVA results for the D parameter in Section 3.3.3.

Proportion of recollective recall on test 2 and test 3. Because the expressions for the proportion of recollective recall on test 2 and 3 are the same, the results are only reported once here. There was a main effect of age, $F(1, 46) = 31.23$, $MSE = .10$, partial $\eta^2 = .47$, $p < .001$, a main effect of gist manipulation, $F(1, 138) = 11.75$, $MSE = .04$, partial $\eta^2 = .08$, $p < .001$, and a Verbatim × Gist interaction, $F(1, 138) = 14.12$, $MSE = .04$, partial $\eta^2 = .09$, $p < .001$. Post hoc tests revealed that the proportions of recollective recall on test 2 and test 3 were higher for younger adults ($M = .16$, $SD = .08$) than for older adults ($M = .07$, $SD = .07$), $p < .001$. Also, the recollective recall ordering for the four conditions was fonts-blocked ($M = .15$, $SD = .10$) >

single-blocked (*M* = .11, *SD* = .09) = single-random (*M* = .11, *SD* = .08) = fonts-random (*M* = .09, *SD* = .07). Consistent with the ANOVA result for the *D* parameter, the results suggest that the verbatim manipulation only elevated the recollective recall with the blocked lists but not with the randomized lists.

Proportion of non-recollective recall on test 1. There was an age main effect, $F(1, 46) = 4.38$, $MSE = .01$, partial $\eta^2 = .04$, $p = .04$, and a gist main effect, $F(1, 138) = 20.81$, $MSE = .04$, partial $\eta^2 = .13$, $p < .001$. Older adults (*M* = .10, *SD* = .05) had more non-recollective recall on test 1 than younger adults (*M* = .08, *SD* = .04), and both younger and older adults had more non-recollective recall on test 1 in the blocked conditions (*M* = .11, *SD* = .05) than in the random conditions (*M* = .08, *SD* = .03).

Proportion of non-recollective recall on test 2. Only a gist main effect was found, $F(1, 138) = 25.08$, $MSE = .06$, partial $\eta^2 = .15$, $p < .001$, as there was more non-recollective recall in the blocked conditions (*M* = .12, *SD* = .06) than in the random conditions (*M* = .08, *SD* = .04).

Proportion of non-recollective recall on test 3. Again, only a gist main effect was found, $F(1, 138) = 33.30$, $MSE = .06$, partial $\eta^2 = .19$, $p < .001$. The proportion of non-recollective recall was higher in the blocked conditions (*M* = .12, *SD* = .06) than in the random conditions (*M* = .08, *SD* = .03).

Mean proportion of recollective recall. The results for the mean proportion of recollective recall are the same as the results for test 1, 2 and 3. There was a main effect for age, $F(1, 46) = 30.76$, $MSE = .09$, partial $\eta^2 = .48$, $p < .001$, for gist manipulation, $F(1, 138) = 13.33$, $MSE = .04$, partial $\eta^2 = .09$, $p < .001$, and a Verbatim × Gist interaction, $F(1, 138) = 14.70$, $MSE = .04$, partial $\eta^2 = .10$, $p < .001$. Older adults (*M* = .07, *SD* = .07) had less recollective recall than younger adults (*M* = .16, *SD* = .08), while both younger and older adults had more recollective recall in the fonts-blocked condition than the other three conditions.

Mean proportion of non- recollective recall. The results for the mean proportion of recollective recall are the same as the results for test 2 and 3. There was a gist main effect, *F*(1, 138) = 28.96, *MSE* = .05, partial $\eta^2 = .17$, $p < .001$, with more non-recollective recall in the blocked conditions (*M* = .11, *SD* = .05) than in the random conditions (*M* = .08, *SD* = .03).

3.3.6. Summary of modeling results

The goodness-of-fits test suggests that only the forgetting version of the dual-retrieval models delivered a satisfactory account for the current data. The parameter estimates indicate that the reconstruction (*R*) parameter and the familiarity (J_2 and J_3) parameters were significantly higher in the blocked conditions than in the random conditions, and so was the proportion of non-recollective recall. This suggests that the gist manipulation was very successful. Also, the forgetting (*F*) parameter was higher in the single conditions than the fonts conditions, while the direct access (*D)* parameter and the proportion of recollective recall were both the highest in the fonts-blocked condition among the four conditions. This indicates that the verbatim manipulation was more effective with the blocked lists than with the randomized lists. In addition, older adults had lower *D* and J_3 parameter but higher *F* parameter than younger adults, which indicates that older adults had less direct verbatim access and were less likely to use familiarity judgments to successfully recall a word after reconstruction, and they were more likely to forget the direct access after the first recall test. Consistent with such results, relative to younger adults, older adults also had lower proportion of recollective recall in all tests and higher proportion of non-recollective recall on test 1.

3.4. Bivariate Correlation results

The bivariate correlation analyses were carried out using the Hmisc package in R (Harrel & Dupont, 2006). All correlation coefficients for the relationships between error calibration scores, mean recall accuracy, mean confidence judgment, mean proportion of recollective recall, mean proportion of non-recollective recall, and the six model parameters are displayed in Table 6. Firstly, recall accuracy and confidence were positively correlated, whereas calibration error score was negatively correlated with both recall accuracy and confidence. This suggests that participants who had higher recall accuracy and confidence also tend to have lower calibration error (i.e., higher metamemory accuracy). Next, recall accuracy and confidence were both positively correlated with direct access (*D*) parameter, reconstruction (*R*) and familiarity judgment (J_3) parameter, *p*s < .05, and they were both negatively correlated with the forgetting (*F*) parameter, *p*s < .001. Participants who had better verbatim memory, better gist memory and less forgetting tended to have higher recall accuracy as well as confidence. However, although recall accuracy was positively correlated with the both the mean proportion of recollective recall and mean proportion of non-recollective recall, confidence was only correlated with the former. This suggests that non-recollective processing helped boost the recall accuracy without increasing participants' subjective confidence. Lastly, calibration error score was negatively correlated with the *D* parameter, *R* parameter and J_3 parameter, *p*s < .05, and positively correlated with *F* parameter, $p < .001$. Also, calibration error score was negatively correlated with the mean proportion of recollective recall, but not with the mean proportion of non-recollective recall, indicating metamemory accuracy was more dependent on recollective recall than non-recollective recall.

Table 6
Bivariate Correlation Matrix

	1	2	3	4	5	6	7	8	9	10	11
1. Calibration	--										
2. Mean recall	-.46**	--									
3. Mean confidence	-.38**	.28**	--								
4. Mean recollective recall	-.44**	.87**	.25**	--							
5. Mean non-recollective recall	-.12	.40**	.11	-.09	--						
6. *D*	-.44**	.86**	.24**	.99**	-.08	--					
7. *F*	.25**	-.38**	-.22*	-.52**	.26**	-.44**	--				
8. *R*	-.20*	.63**	.15*	.25**	.83**	.24**	.02	--			
9. J_1	-.04	-.07	.03	-.27**	.45**	-.28**	.16*	.10	--		
10. J_2	-.02	.07	.04	-.18*	.53**	-.14*	.34**	.13	.38*	--	
11. J_3	-.22*	.39**	.15*	.11	.49**	.14	.10	.21*	.31*	.67*	--

Note. * for $p < .05$, ** for $p < .001$. Calibration = calibration error scores. Mean recollective recall = mean proportion of recollective recall. Mean non-recollective recall = mean proportion of non-recollective recall.

3.5. Regression results

The multiple regression analyses were conducted using the lme4 package (Bates et al., 2012) and lmerTest package in R (Kuznetsova et al., 2017). I performed three separate linear mixed effects analyses, using mean recall accuracy, mean confidence judgment and calibration error scores as outcome variables, respectively. The most important regression analysis is the third one, which tested the hypothesis that verbatim- and/or gist-related retrieval processes can predict metamemory accuracy at individual level. In all of the three models, I included the six retrieval parameters (D, R, F, J_1, J_2, J_3), age (younger vs. older), condition (single-blocked, fonts-blocked, random, fonts-random), depression (GDS scores), and working memory (one-back accuracy) as fixed effects, and also added the by-participant random slope as random effect. Based on the preliminary regression analyses, I kept the interaction terms that reached significance in the models. For the model where recall accuracy was the outcome variable, I included three additional interaction terms: D × Age, F × Age and J_3 × Age. For the model where confidence was the outcome variable, no interaction term was included. For the model where calibration error score was the outcome variable, I included an interaction term J_3 × Age. For convenience of comparison, all the numeric predictors were standardized in the three models. The significant fixed effects of the three regression models were displayed in Figure 1.

3.5.1. Recall accuracy

Before running the regression model, I visually inspected the diagnostic plots and did not spot any apparent deviation from homoscedasticity or normality. The results show that the fixed effect variables explained 98.0% of the variance in recall accuracy (marginal $R^2 = .98$). All the model parameters, except for J_2, significantly predicted recall accuracy. The coefficient estimates are as follows: The direct access (D) parameter, $\beta = .08$, $t(150) = 36.56$, $p < .001$; the reconstruction (R) parameter, $\beta = .05$, $t(164) = 35.66$, $p < .001$; the forgetting (F) parameter $\beta = -.03$, $t(172) = -7.64$, $p < .001$; and the familiarity judgment (J_3) parameter, $\beta = .02$, $t(172) = 10.04$, $p < .001$. In addition, there were a significant D × Age interaction, $\beta = -.009$, $t(137) = -3.02$, $p = .003$, a F × Age interaction, $\beta = .02$, $t(172) = 4.90$, $p < .001$, and a J_3 × Age interaction, $\beta = .009$, $t(172) = 3.58$, $p < .001$. These results suggest that both verbatim- and gist-related processes could predict recall accuracy. This is completely consistent with dual-retrieval model's theoretical assumption, since all the parameter estimates are derived from recall accuracy data. Furthermore, the interactions indicate that compared to younger adults, older adults' recall accuracy was less dependent on verbatim-related processes (*D* and *F*), but more dependent on gist-related process (J_3).

3.5.2. Confidence

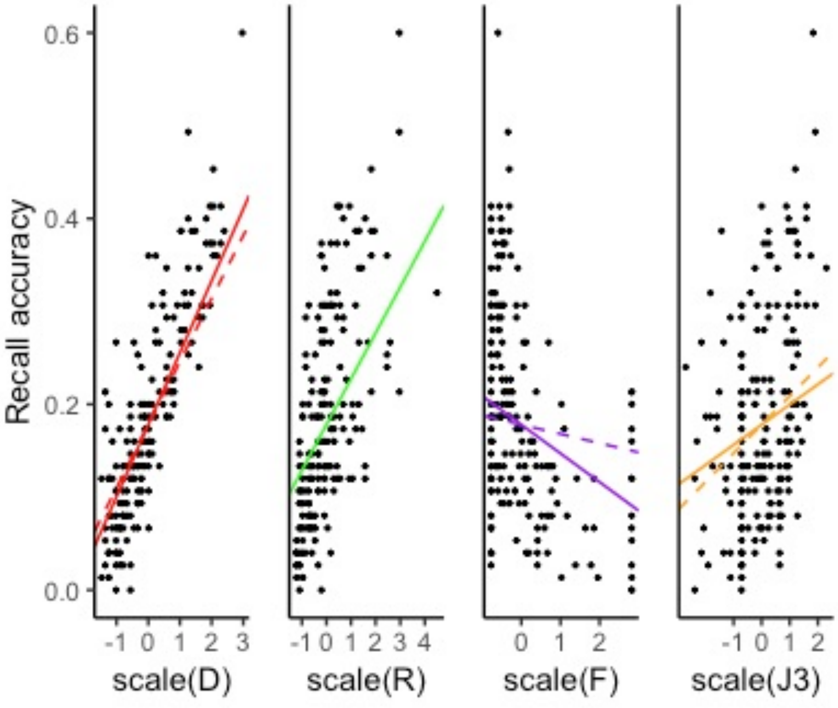
As explained in the ANOVA analysis (Section 3.2.3), I conducted the same logit transformation for mean confidence to handle the violation of homoscedasticity assumption. The results showed that the fixed effect variables explained 10% of the variance in confidence judgment (marginal $R^2 = .10$). The *D* parameter significantly predicted confidence, $\beta = .21$, $t(174) = 2.24$, $p = .03$), and the *F* parameter also significantly predicted confidence, $\beta = -.14$, $t(144) = -2.16$, $p = .03$. Therefore, individuals with better direct verbatim access and less forgetting were expected to have higher confidence judgments for their own recall accuracy. In other words, the level of confidence depends heavily on how good an individual's verbatim memory is.

3.5.3. Calibration error scores

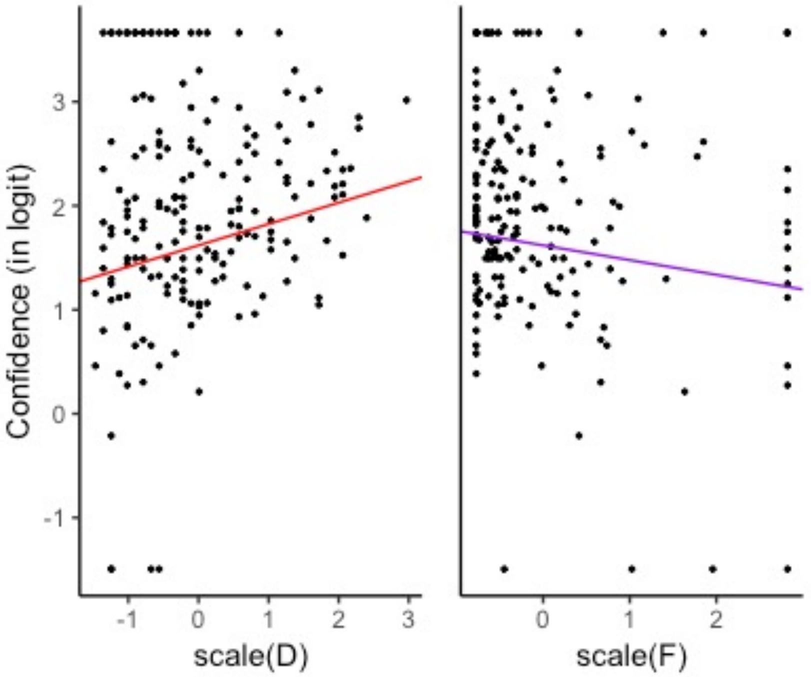
To fix the violation of homoscedasticity assumption, a log transformation was applied to calibration error scores, just as in the ANOVA analysis (Section 3.2.4). The results suggest that the fixed effect variables explained 39% of the variance in calibration error scores (marginal $R^2 = .39$). The *D* parameter significantly predicted calibration error scores, $\beta = -..25$, $t(164) = -4.59$, $p < .001$, so did the *R* parameter, $\beta = -1.04$, $t(158) = -3.09$, $p < .001$; and the J_1 parameter, $\beta = -.89$, $t(157) = -2.44$, $p = .03$. Also, a significant interaction between judgment (J_3) parameter and age was found, $\beta = -.17$, $t(156) = -2.40$, $p = .02$. The J_3 parameter only significantly predicted calibration error scores for older adults, but not for younger adults. Interestingly, age did not significantly predict calibration error scores when controlling for the other predictors (*p* = .40), suggesting that the age difference in overall calibration error scores (as previously found in the ANOVA analyses) could potentially be explained by the individual differences in the *D, R*, J_1 and J_3 parameters.

Figure 1

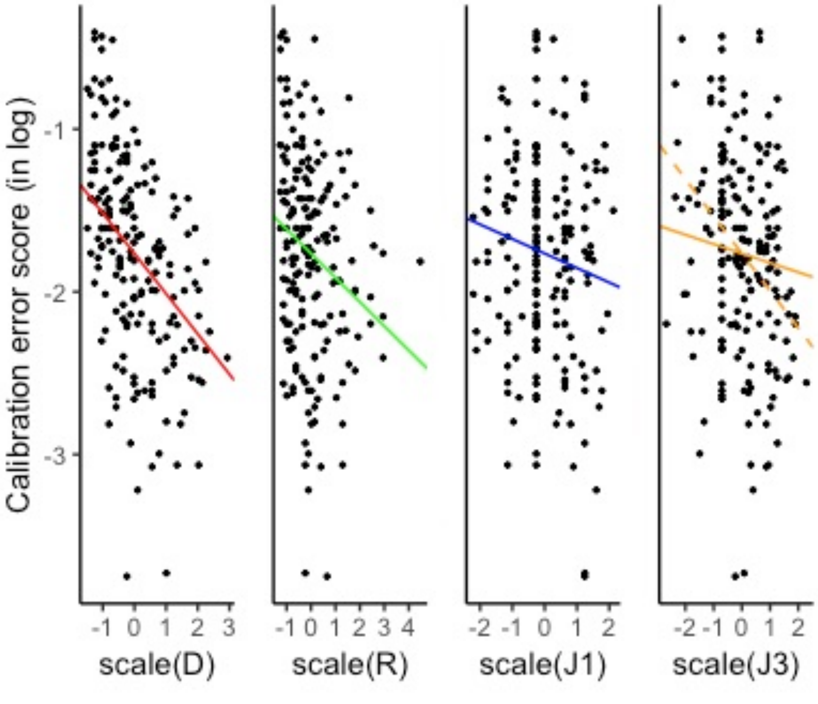
Recall Accuracy, Confidence and Calibration Error Scores as a Function of the Retrieval Parameters



(a)



(b)



(c)

Note. Figure 1(a), 1(b), 1(c) show mean recall accuracy, mean confidence, mean calibration error scores as a function of the retrieval parameters, respectively. *D* = direct access, *R* =

reconstruction, F = forgetting, J_1 = familiarity judgment for trial 1, J_3 = familiarity judgment for trial 3. All retrieval parameters have been standardized for the ease of comparison (e.g. scale(D)). When there was no significant interaction between a parameter and age, the solid line represents aggregated data for younger and older adults. When there was a significant interaction between a parameter and age, the solid line represents data for younger adult, while the dashed line represents data for older adults.

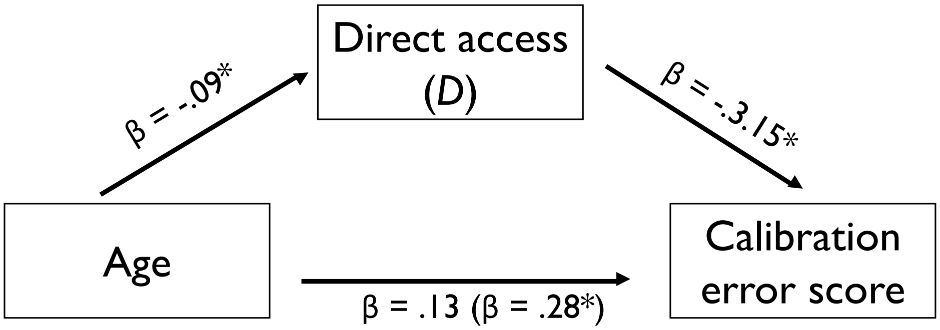
3.6. Mediation results

To further explore which retrieval process can explain the age difference in metamemory accuracy, I conducted mediation analyses using the lavaan package in R (Rosseel, 2010). With the mediation analyses, I examined whether the age effects on metamemory accuracy were mediated by verbatim- and/or gist related retrieval processes. Based on the regression results, the candidate mediators should be direct access (D), familiarity judgment 1 (J_1) or familiarity judgment 3 (J_3). The reconstruction (R) parameter was not included here because it did not differ significantly between younger and older adults (see Section 3.3.3 for more details).

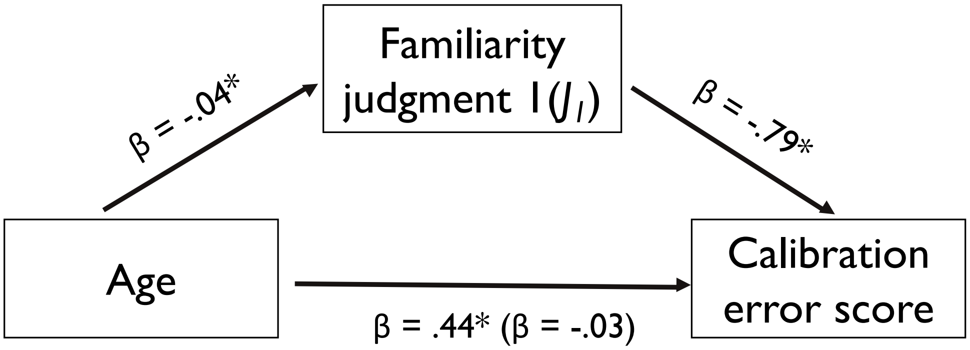
As illustrated in Figure 2a, the regression coefficient between age and the D parameter was significant (β = -.09, SE = .01, p < .001), and so was the regression coefficient between the D parameter and calibration error scores (β = - 3.15, SE = .46, p < .001). The indirect effect was computed for 1000 bootstrapped samples. The bootstrapped indirect effect was significant (β = .28, SE = .05, p < .001), whereas the bootstrapped direct effect was insignificant (β = .13, SE = .09, p = .14). This result indicates that the effects of age on calibration error scores were fully mediated by direct access (D), namely, age differences in verbatim direct access could account for the age-related declines in metamemory accuracy.

However, as illustrated in Figure 2b and 2c, the effects of age on calibration error scores were not mediated by familiarity judgments (J_1 and J_3). The regression coefficient between age and J_1 parameter was significant (β = -.04, SE = .02, p =.02), and so was the regression coefficient between J_1 parameter and calibration error scores (β = - .80, SE = .38, p = .04). The bootstrapped indirect effect via J_1 parameter was insignificant (β = -.03, *SE* = .02, *p* = .14), whereas the bootstrapped direct effect was significant (β = .44, *SE* = .09, *p* < .001). Similarly, the regression coefficient between age and J_3 parameter was significant (β = -.04, *SE* = .02, *p* =.02), and so was the regression coefficient between J_3 parameter and calibration error scores (β = - .84, *SE* = .33, *p* = .01). The bootstrapped indirect effect via J_3 parameter was insignificant (β = .04, *SE* = .02, *p* = .09), but the bootstrapped direct effect was significant (β = .37, *SE* = .09, *p* < .001). These results suggest that the age differences in familiarity judgments could not account for the age decline in overall metamemory accuracy.

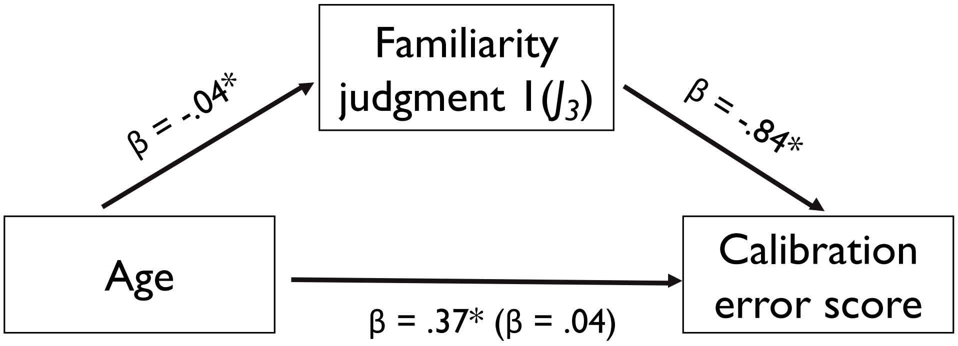
Figure 2
Regression Coefficients for the Relationship between Age and Calibration Error Scores as Mediated by Direct Access or Familiarity Judgments.



(a)



(b)



(c)

Note. Figure 1a shows the relationship between age and calibration error scores as mediated by the direct access (*D*) parameter. Figure 1b and 1c show the relationship between age and calibration error scores as mediated by the familiarity judgement (J_1 and J_3, respectively) parameters. The regression coefficients between age and calibration error scores controlling for direct access or familiarity judgment are show in parentheses. **p* < .05.

Concerns could be raised that the J_1 or J_3 parameter was estimated using data from only one trial, while the *D* parameter was estimated based on data from all three trials, which led to lower statistical reliability for the J_1 and J_3 parameter estimations than the *D* parameter estimation. Accordingly, I run a separate model with a pooled *J* parameter estimate, in which the J_1, J_2, and J_3 parameters in the original model were replaced with one single *J* parameter. Then, I used the pooled *J* parameter estimation as the mediator between age and calibration error scores. However, the regression coefficient between age and the *J* parameter was insignificant (β = .009, *SE* = .02, *p* =.54), which led to an insignificant bootstrapped indirect effect via *J* parameter (β = -.009, *SE* = .02, *p* = .59). This result confirms the previous finding that the age differences in familiarity judgment could not account for the age-related declines in metamemory accuracy.

In addition to the mediation results reported above, I also found that the mean proportion of recollective recall fully mediated the age effect on calibration error scores, while the mean proportion of non-recollective recall did not. The regression coefficient between age and recollective recall (β = -.26, *SE* = .03, *p* < .001) and that between calibration error scores and recollective recall (β = -1.0, *SE* = .16, *p* < .001) were both significant. The indirect effect via recollective recall (β = .29, *SE* = .05, *p* < .001) was significant, with the direct effect being insignificant (β = .12, *SE* = .10, *p* = .19). However, the regression coefficient between age and non-recollective recall was insignificant (β = .03, *SE* = .02, *p* =.16), leading to an insignificant indirect effect (β = -.02, *SE* = .02, *p* =.23). This again supports the conclusion that the age differences in metamemory accuracy was mainly explained by age differences in verbatim memory.

4. DISCUSSION

The current study adopts a dual-process approach to examine the underlying mechanism for age differences in retrospective metamemory accuracy. Firstly, I hypothesized that older adults would have lower metamemory accuracy for free recall tests than younger adults. As expected, I found that older adults had higher calibration error scores than younger adults, which suggests that older adults were less accurate in assessing their own recall accuracy. Such a result is consistent with many prior findings about age-related declines in metamemory accuracy (e.g., Chua, et al., 2009; Dodson et al., 2007; Kelley & Sahakyan, 2003; Meade & Roediger, 2006; Shing, et al., 2009; Wong, et al., 2012). Noticeably, while most prior studies in this line of research used recognition or cued recall tests, the current study extends the findings to free recall tests.

More importantly, I found that the verbatim-related retrieval process (direct access; *D*) and gist-related retrieval process (reconstruction; *R*) could predict metamemory accuracy for both younger and older adults. As the *D* parameter or the *R* parameter increased, calibration error scores were expected to decrease, indicating better correspondence between the confidence judgment and actual recall accuracy. This suggests that if people were more likely to directly access the verbatim details of an item's prior presentation, or reconstruct the item based on the partial-identifying gist information, they would be better at judging their own memory accuracy. Meanwhile, the familiarity judgment (J_3) only significantly predicted metamemory accuracy for older adults but not for younger adults. This means that older adults' metamemory accuracy was more dependent on familiarity judgment than younger adults. This was an interesting finding, because the modeling results reveal that given that there was no age difference in the *R* parameter, the J_3 parameter was lower for older adults than for younger adults. Taken together, these results indicate that younger and older adults were equally capable of reconstructing items based on gist, and the reconstructed items seemed more familiar to younger adults than to older adults. Nevertheless, younger adults relied less on familiarity judgment for their metamemory accuracy than older adults. This paradox suggests that younger adults relied overwhelmingly on direct verbatim access as the primary contributor to metamemory accuracy, possibly because they had superior verbatim memory. Meanwhile, older adults based their metamemory accuracy on both direct verbatim access and familiarity judgment, which echoes with the idea that older adults tended to recruit additional processing as compensation for age-related cognitive decline (e.g., Park, & Reuter-Lorenz, 2009; Spaniol & Grady, 2012).

Based on the finding that both verbatim- and gist-related retrieval parameters could predict metamemory accuracy, a further mediation analysis was conducted to test which parameters were responsible for age differences in metamemory accuracy. As revealed by the mediation analysis, only the D parameter can account for the effects of age on calibration error scores, while the J_1 or J_3 parameter cannot. This demonstrates that age-related declines in metamemory accuracy is mainly attributed to the age-related deficiency in verbatim memory, rather than age-related heavier reliance on gist memory. This is conceptually consistent with the findings of prior dual-process research on prospective metamemory, which indicates that JOL predictive accuracy of recollection declines with age whereas that of familiarity is relatively spared in aging (Daniels et al., 2009; Kuhlmann & Undorf, 2018; Soderstrom et al., 2012).

Although I initially hypothesized that age differences in metamemory accuracy would be smaller in conditions that encourage verbatim processing and/or that inhibit gist processing, the results show that there was neither an Age × Verbatim interaction nor an Age × Gist interaction for metamemory accuracy. Given that the mediation analyses reveal that age differences in metamemory accuracy is mainly explained by age-related declines in verbatim memory, it is not surprising that the gist manipulation did not moderate age difference in metamemory accuracy. However, the absence of the Age × Verbatim interaction was unexpected. This null result is most probably due to the fact that the verbatim manipulation was not robust enough to modulate age differences in metamemory accuracy. As shown in the group-level modeling analyses (see Section 3.3.3), the 2 (verbatim) 2 (gist) × 2 (age) ANOVA for the *D* parameter revealed a Verbatim × Gist interaction. The post hoc tests suggested that significant difference in the *D* parameter was only found between the fonts-blocked condition and the other three conditions, suggesting that the different atypical fonts only enhanced verbatim processing with the blocked DRM lists. In addition, as can be seen in the second row of Table 5, the differences in the *D* parameter between fonts-blocked condition and other conditions were fairly small (ranging from .01 to .04 in younger adults and from .01 to .05 in older adults), especially when compared to the age differences in direct access (ranging from .10 to .12 among different conditions). Thus, the variations in verbatim processing across conditions may not be large enough to induce differences in metamemory accuracy. Accordingly, future studies are recommended to administer stronger verbatim manipulations, such as full attention vs. divided attention (Kelley & Sahakyan, 2003), which was shown to successfully moderate the age differences in metamemory accuracy.

Although an age-related decline was found in both mean recall accuracy and metamemory accuracy, no age effect was found in mean confidence judgment. However, when further examining the proportion of high-confidence responses within true and false recall, I found that younger and older adults had qual proportion of high confidence responses for true recall, but older adults (*M* = .81, *SD* = .32) had higher proportion of high confidence false recall than younger adults (*M* = .63, *SD* = .34), *F*(1,46) = 6.04, *MSE* = .41, partial η^2 = .09, *p* = .02. Also, older adults' high confidence errors were not due to a higher tendency to assign extremely high confidence, because t-tests showed that younger and older adults did not differ in the assignment of high confidence judgment of 4, 5, or 6 (see Table 3). Such result conceptually replicates the prior findings that older adults tended to make more high confidence errors than younger adults (Dodson et al., 2007; Fandakova et al., 2013; Gallo et al., 2009; Jacoby et al., 2005; Shing et al., 2009).

How can the current findings explain the age-related increases in high confidence error? Recall that in the regression model for recall accuracy, there were interactions between age and the two verbatim-related processes (direct access and forgetting). The slopes for the two verbatim-related processes were flatter for older adults than for younger adults, which suggests that older adults relied less on verbatim memory in the recall test than younger adults did. However, in the regression model for confidence, there was no interactions between age and any of the two verbatim-related process, suggesting that younger and older adults relied equally heavily on verbatim memory during confidence judgment. Such discrepancy means that the age-related dissociation between objective memory performance and subjective confidence judgment may be attributed to older adults' disproportionate reliance on verbatim memory when making confidence judgment. When older adults' worse direct verbatim access led to lower recall accuracy than younger adults, their confidence did not effectively reflect such age-related verbatim deficiency. Accordingly, older adults' confidence judgment would be less discriminative of true versus false recall, leading to their higher tendency of committing high-confidence errors.

Last but not least, two limitations of the current study should be noted. Firstly, both younger adults and older adults recruited in this study were highly educated. The mean years of formal education for younger adults is 14.16 years (*SD* = .96 years), and that for older adults is 16.71 years (*SD* = 3.25 years). As Szajer and Murphy (2013) suggested, older adults with 17 years of education or more had significantly higher retrospective metamemory accuracy than those with 12 years of education or less. Thus, the current sample may not be representative of the general public. Secondly, as discussed in the Results section, only one measure of metamemory accuracy (calibration error scores) was used. Other potential measures, such as gamma correlation coefficients or adjusted normalized discrimination index (ANDI; Koriat & Goldsmith, 1996), were not appropriate for the current study, because they were not calculable for some participants when they had no variance in either confidence judgment or recall accuracy. Such cases typically occurred when participants consistently chose one single level of confidence for all items recalled. Thus, due to the concern of losing data and decreasing power, I did not use gamma correlation or ANDI in the current study. To cope with this limitation, I will try to make more detailed instructions in the follow-up studies, such as asking participants to carefully finely grade their confidence judgments. Also, I may use recognition or cued-recall test in the follow-up studies, which should allow for more variability in the accuracy and confidence judgment, and hence I should be able to use multiple measures of metamemory accuracy.

5. Conclusion

The main goal of the current study was to investigate whether and how age differences in verbatim-related retrieval process (direct verbatim access and forgetting) and gist-related retrieval process (reconstruction and familiarity judgment) could account for age differences in metamemory accuracy. As many prior studies, I found that metamemory accuracy was lower in older adults than in younger adults. Importantly, the direct verbatim access and reconstruction could predict metamemory accuracy at individual level for both younger and older adults, while familiarity judgment could only predict metamemory accuracy for older adults. Moreover, only direct verbatim access significantly mediated the age effects on metamemory accuracy, whereas other retrieval processes did not, suggesting that age differences in metamemory accuracy were greatly dependent on age-related declines in verbatim memory.

References

Arndt, J., & Reder, L. M. (2003). The effect of distinctive visual information on false recognition. *Journal of Memory and Language, 48*(1), 1-15.

Bates, D., Maechler, M., and Bolker, B. (2012). *lme4: Linear mixed-effects models using S4 classes. R package version 0.999999-0*. Retrieved from http://cran.r-project.org/web/packages/lme4/index.html.

Brainerd, C. J., Payne, D. G., Wright, R., & Reyna, V. F. (2003). Phantom recall. *Journal of Memory and Language, 48*(3), 445-467.

Brandt, J., Spencer, M., and Folstein, M. (1988). The telephone interview for cognitive status. *Neuropsychiatry Neuropsychol. Behav. Neurol. 1*, 111–117.

Brainerd, C. J., Reyna, V. F., & Howe, M. L. (2009). Trichotomous processes in early memory development, aging, and neurocognitive impairment: a unified theory. *Psychological Review, 116*(4), 783.

Brainerd, C. J., Reyna, V. F., Gomes, C. F., Kenney, A. E., Gross, C. J., Taub, E. S., & Spreng, R. N. (2014). Dual-retrieval models and neurocognitive impairment. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(1), 41.

Brainerd, C. J., Wright, R., Reyna, V. F., & Payne, D. G. (2002). Dual-retrieval processes in free and associative recall. *Journal of Memory and Language, 46*(1), 120-152.

Brink, T. L., Yesavage, J. A., Lum, O., Heersema, P. H., Adey, M., & Rose, T. L. (1982). Screening tests for geriatric depression. *Clinical gerontologist, 1*(1), 37-43.

Busey, T. A., Tunnicliff, J., Loftus, G. R., & Loftus, E. F. (2000). Accounts of the confidence-accuracy relation in recognition memory. *Psychonomic Bulletin & Review, 7*(1), 26-48.

Chua, E. F., Schacter, D. L., & Sperling, R. A. (2009). Neural basis for recognition confidence in younger and older adults. *Psychology and Aging, 24*(1), 139.

Comblain, C., D'Argembeau, A., Van der Linden, M., & Aldenhoff, L. (2004). The effect of ageing on the recollection of emotional and neutral pictures. *Memory, 12*(6), 673-684.

Cook, S. E., Marsiske, M., & McCoy, K. J. (2009). The use of the Modified Telephone Interview for Cognitive Status (TICS-M) in the detection of amnestic mild cognitive impairment. *Journal of Geriatric Psychiatry and Neurology, 22*(2), 103-109.

Dahl, M., Allwood, C. M., & Hagberg, B. (2009). The realism in older people's confidence judgments of answers to general knowledge questions. *Psychology and Aging, 24*(1), 234.

Daniels, K. A., Toth, J. P., & Hertzog, C. (2009). Aging and recollection in the accuracy of judgments of learning. *Psychology and Aging, 24*(2), 494.

Daniels, K., Toth, J., & Jacoby, L. (2006). The aging of executive functions. In Bialystok, E and Craik, F.I.M (Eds). *Lifespan cognition: Mechanisms of change* (pp. 96-111). New York, NY: Oxford University Press.

Daselaar, S. M., Fleck, M. S., & Cabeza, R. E. (2006). Triple dissociation in the medial temporal lobes: Recollection, familiarity, and novelty. *Journal of Neurophysiology, 96*, 1902–1911.

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

Dennis, N. A., Kim, H., & Cabeza, R. (2008). Age-related differences in brain activity during true and false memory retrieval. *Journal of Cognitive Neuroscience, 20*(8), 1390-1402.

Dodson, C. S., Bawa, S., & Krueger, L. E. (2007). Aging, metamemory, and high-confidence errors: A misrecollection account. *Psychology and Aging, 22*(1), 122.

Dodson, C. S., Bawa, S., & Slotnick, S. D. (2007). Aging, source memory, and misrecollections. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(1), 169.

Dunlosky, J., Kubat-Silman, A. K., & Hertzog, C. (2003). Training monitoring skills improves older adults' self-paced associative learning. *Psychology and Aging, 18*(2), 340.

Fandakova, Y., Shing, Y. L., & Lindenberger, U. (2013). High-confidence memory errors in old age: The roles of monitoring and binding processes. *Memory, 21*(6), 732-750.

Ferraro, F. R., & Chelminski, I. (1996). Preliminary normative data on the Geriatric Depression Scale-Short Form (GDS-SF) in a young adult sample. *Journal of Clinical Psychology, 52*(4), 443-447.

Flavell, J. H. & Wellman, H. M. (1977). Metamemory. In R. V. Kail & J. W. Hagen (Eds.), *Perspectives on the development of memory and cognition* (pp. 3-33). Hillsdale, NJ: Erlbaum.

Fox, C., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., Firth, D., Friendly, M., Gorjanc, G., Graves, S., Heiberger, R., Laboissiere, R.. Monette, G., Murdoch, D., Nilsson, H., Ogle, D., Ripley, B., Venables, W., and Zeileis, A. (2013), *Package car*, available at: http://cran.r.project.org/web/packages/car/car.pdf (accessed 6 November 2013).

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

Gallo, D. A., Foster, K. T., & Johnson, E. L. (2009). Elevated false recollection of emotional pictures in young and older adults. *Psychology and aging, 24*(4), 981.

Geerlings, M. I., Bouter, L. M., Schoevers, R. A., Beekman, A. T. F., Jonker, C. E. E. S., Deeg, D. J. H., ... & Schmand, B. (2000). Depression and risk of cognitive decline and Alzheimer's disease: Results of two prospective community-based studies in The Netherlands. *The British Journal of Psychiatry, 176*(6), 568-575.

Gomes, C. F. A. (2013). *On the relation between memory and metamemory in free recall: The effects of list length and word frequency on dual processes* (Unpublished master's thesis). Cornell University.

Gomes, C. F. A., Brainerd, C. J., Nakamura, K., & Reyna, V. F. (2014). Markovian interpretations of dual retrieval processes. *Journal of Mathematical Psychology, 59*, 50-64.

Harrell Jr, F. E., & Dupont, C. (2008). Hmisc: Harrell miscellaneous. *R package version, 3*(2).

Herrmann, N., Mittmann, N., Silver, I. L., Shulman, K. I., Busto, U. A., Shear, N. H., & Naranjo, C. A. (1996). A validation study of the Geriatric Depression Scale short form. *International Journal of Geriatric Psychiatry, 11*(5), 457-460.

Hertzog, C., & Dunlosky, J. (2011). Metacognition in later adulthood: Spared monitoring can benefit older adults' self-regulation. *Current Directions in Psychological Science, 20*(3), 167-173.

Hertzog, C., Kidder, D. P., Powell-Moman, A., & Dunlosky, J. (2002). Aging and monitoring associative learning: Is monitoring accuracy spared or impaired?. *Psychology and Aging, 17*(2), 209.

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

Jacoby, L. L., Bishara, A. J., Hessels, S., & Toth, J. P. (2005). Aging, subjective experience, and cognitive control: Dramatic false remembering by older adults*. Journal of Experimental Psychology: General, 134*(2), 131.

Jacoby, L. L., Debner, J. A., & Hay, J. F. (2001). Proactive interference, accessibility bias, and process dissociations: Valid subject reports of memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(3), 686.

Jacoby, L. L., & Rhodes, M. G. (2006). False remembering in the aged. *Current Directions in Psychological Science, 15*(2), 49-53.

Jonker, C., Geerlings, M. I., & Schmand, B. (2000). Are memory complaints predictive for dementia? A review of clinical and population-based studies. *International Journal of Geriatric Psychiatry, 15*(11), 983-991.

Jorm, A. F. (2000). Is depression a risk factor for dementia or cognitive decline?. *Gerontology, 46*(4), 219-227.

Kahn, R. L., Zarit, S. H., Hilbert, N. M., & Niederehe, G. (1975). Memory complaint and impairment in the aged: The effect of depression and altered brain function. *Archives of General Psychiatry, 32*(12), 1569-1573.

Kane, M. J., Conway, A. R., Miura, T. K., & Colflesh, G. J. (2007). Working memory, attention control, and the N-back task: a question of construct validity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(3), 615-622.

Karpel, M. E., Hoyer, W. J., & Toglia, M. P. (2001). Accuracy and qualities of real and suggested memories: Nonspecific age differences. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 56*(2), P103-P110.

Kelley, C. M., & Sahakyan, L. (2003). Memory, monitoring, and control in the attainment of memory accuracy. *Journal of Memory and Language, 48*(4), 704-721.

Knopman, D. S., Roberts, R. O., Geda, Y. E., Pankratz, V. S., Christianson, T. J., Petersen, R. C., & Rocca, W. A. (2010). Validation of the telephone interview for cognitive status-modified in subjects with normal cognition, mild cognitive impairment, or dementia. *Neuroepidemiology, 34*(1), 34-42.

Koen, J. D., & Yonelinas, A. P. (2014). The effects of healthy aging, amnestic mild cognitive impairment, and Alzheimer's disease on recollection and familiarity: A meta-analytic review. *Neuropsychology review, 24*(3), 332-354.

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. Psychological review, 103(3), 490.

Kuhlmann, B. G., & Undorf, M. (2018). Is all metamemory monitoring spared from aging? A dual-process examination. *Psychology and aging, 33*(8), 1152.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: tests in linear mixed effects models. *Journal of Statistical Software, 82*(13).

Lovelace, E. A. (1990). Aging and metacognitions concerning memory function. In E. A. Lovelace (Ed.), *Aging and cognition: Mental processes, self-awareness and interventions* (pp. 157-187). North Holland: Elsevier Science.

Malmberg, K. J. (2008). Recognition memory: A review of the critical findings and an integrated theory for relating them. *Cognitive Psychology, 57*(4), 335-384.

Meade, M. L., & Roediger III, H. L. (2006). The effect of forced recall on illusory recollection in younger and older adults. *The American Journal of Psychology, 119*(3), 433-462.

McDermott, K. B., & Watson, J. M. (2001). The rise and fall of false recall: The impact of presentation duration. *Journal of Memory and Language, 45*(1), 160-176.

Mitchell, K. J., Johnson, M. K., & Mather, M. (2003). Source monitoring and suggestibility to misinformation: Adult age-related differences. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition, 17*(1), 107-119.

Nelson, T. O. (1990). Metamemory: A theoretical framework and new findings. In *Psychology of learning and motivation* (Vol. 26, pp. 125-173). Academic Press.

Norman, K. A., & Schacter, D. L. (1997). False recognition in younger and older adults: Exploring the characteristics of illusory memories. *Memory & Cognition, 25*(6), 838-848.

Park, D. C., & Reuter-Lorenz, P. (2009). The adaptive brain: aging and neurocognitive scaffolding. *Annual Review of Psychology, 60,* 173-196.

Parkin, A. J., & Walter, B. M. (1992). Recollective experience, normal aging, and frontal dysfunction. *Psychology and aging, 7*(2), 290.

Payne, D. G., Elie, C. J., Blackwell, J. M., & Neuschatz, J. S. (1996). Memory illusions: Recalling, recognizing, and recollecting events that never occurred. *Journal of Memory and Language, 35*, 261–285.

Pliske, R. M., & Mutter, S. A. (1996). Age differences in the accuracy of confidence judgments. *Experimental Aging Research,* 22(2), 199-216.

Reyna, V. F., & Brainerd, C. J. (1995). Fuzzy-trace theory: An interim synthesis. *Learning and Individual Differences, 7*(1), 1-75.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(4), 803.

Roediger, H. L., Watson, J. M., McDermott, K. B., & Gallo, D. A. (2001). Factors that determine false recall: A multiple regression analysis. *Psychonomic Bulletin & Review, 8*(3), 385-407.

Rosseel, Y. (2010). lavaan: An R package for structural equation modeling and more Version 0. 3-1. *Word Journal of The International Linguistic Association, 1*, 1–32.

Rugg, M. D., & Curran, T. (2007). Event-related potentials and recognition memory. *Trends in Cognitive Sciences, 11*, 251–257.

Shing, Y. L., Werkle-Bergner, M., Li, S. C., & Lindenberger, U. (2009). Committing memory errors with high confidence: Older adults do but children don't. *Memory, 17*(2), 169-179.

Singmann, H., & Kellen, D. (2013). MPTinR: Analysis of multinomial processing tree models in R. *Behavior Research Methods, 45*(2), 560-575.

Soderstrom, N. C., McCabe, D. P., & Rhodes, M. G. (2012). Older adults predict more recollective experiences than younger adults. *Psychology and Aging, 27*(4), 1082.

Spaniol, J., & Grady, C. (2012). Aging and the neural correlates of source memory: over-recruitment and functional reorganization. *Neurobiology of aging, 33*(2), 425-e3.

Szajer, J., & Murphy, C. (2013). Education level predicts retrospective metamemory accuracy in healthy aging and Alzheimer's disease. *Journal of Clinical and Experimental Neuropsychology, 35*(9), 971-982.

Thiede, K. W. (1999). The importance of monitoring and self-regulation during multitrial learning. *Psychonomic Bulletin & Review, 6*(4), 662-667.

Thompson-Schill, S. L., Kan, I. P., & Oliver, R. T. (2006). Functional neuroimaging of semantic memory. In R. Cabeza & A. Kingstone (Eds.), *Handbook of functional neuroimaging of cognition* (pp. 149–190). Cambridge, MA: MIT Press.

Toglia, M. P. (1999). Recall accuracy and illusory memories: When more is less. *Memory, 7*(2), 233-256.

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

Undorf, M., Böhm, S., & Cüpper, L. (2016). Do judgments of learning predict automatic influences of memory?. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*(6), 882.

Wells, G. L., Ferguson, T. J., & Lindsay, R. C. (1981). The tractability of eyewitness confidence and its implications for triers of fact. *Journal of Applied Psychology, 66*(6), 688.

Welsh, K. A., Breitner, J. C., & Magruder-Habib, K. M. (1993). Detection of dementia in the elderly using telephone screening of cognitive status. *Neuropsychiatry, Neuropsychology, & Behavioral Neurology, 6*(2), 103-110.

Williams, J. M., Little, M. M., Scates, S., & Blockman, N. (1987). Memory complaints and abilities among depressed older adults. *Journal of Consulting and Clinical Psychology, 55*(4), 595.

Wong, J. T., Cramer, S. J., & Gallo, D. A. (2012). Age-related reduction of the confidence–accuracy relationship in episodic memory: Effects of recollection quality and retrieval monitoring. *Psychology and Aging, 27*(4), 1053.

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

Zheng, Z., Lang, M., Wang, W., Xiao, F., & Li, J. (2019). Episodic reconstruction contributes to high-confidence false recognition memories in older adults: Evidence from event-related potentials. *Brain and Cognition, 132*, 13-21.

APPENDIX A

Modified Telephone Interview for Cognitive Status (TICS-M)

Scoring: one point for each correct answer unless otherwise specified.

	Instructions	Score
1.	Please tell me you full name: First name ______________ Last name ______________	/ 2
2.	What is today's date? Month __________________ Day ____________________ Year ____________________ Day of week ______________ Season___________________	/ 5
3.	Where are you right now? (1 point each for house number, street, state, zip code)	/ 4
4.	Please count backwards from 20 to 1. Trial 1: (Circle each correct response): 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1 (2 points for the correct response in trial 1). Trial 2: (Administer only if the participant did not correctly complete trial 1): "Now, let's try that again. I would like for you to count backwards from 20 to 1: 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1 (1 point for the correct response in trial 2; otherwise, 0 point).	/ 2
5.	"I am going to read a list of 10 words. Please listen carefully and try to remember them. When I am done, tell me as many words as you can, in any order. Ready? The words are:" cabin _______ theatre _______ pipe _______ watch _______ elephant _______ whip _______ chest _______ pillow _______ silk _______ giant _______ "Now, tell me all the words you can remember." (1 point for each correct response above)	/ 10
6.	"Now I'd like you to subtract 7 from 100. Then keep subtracting 7 from each answer until I tell you to stop. What is 100 take away 7?" 93 86 79 72 65	/ 5

(Record the participant’s responses in the boxes. 1 point for the correct answer in each box)

7. “What do people usually use to cut paper?” (1 point for scissors or shears) “How many things are in a dozen?” (1 point for 12) “What do you call the prickly green plant that lives in the dessert” (1 point for cactus) “What animal does wool comes from?” (1 point for sheep or lamb)	/ 4
8. “Say this, ‘No ifs, ands or buts’.” “Say this, ‘Methodist Episcopal’.” (1 point for each completed repetition on the first trial. Repeat only if poorly presented)	/ 2
9. “Who is the President of the United States right now?” (1 point for correct first and last name only) “Who is the Vice President?” (1 point for correct first and last name only)	/ 2
10. “With your finger, tap 5 times on the part of the phone you speak into.” (2 points if 5 taps are heard; 1 point if participant taps more or less than 5 times)	/ 2
11. “I’m going to give you a word and I want you to give me its opposite. For example, the opposite of hot is cold. What is the opposite of ‘west’?” (1 point for east) What is the opposite of ‘generous’?” (1 point for selfish, greedy, stingy, tight, cheap, mean, meager, skimpy or other good antonyms)	/ 2
12. “Earlier I read a long list of words to you. Please tell me all of the words that you can remember from that list.” cabin _______ theatre _______ pipe _______ watch _______ elephant _______ whip _______ chest _______ pillow _______ silk _______ giant _______	/ 10

Total score = / 50

APPENDIX B

Geriatric Depression Scale (short form)

Instruction: circle the answer that best describe how you felt over the past week.

1. Are you basically satisfied with your life? Yes / No
2. Have you dropped many of your activities and interests? Yes / No
3. Do you feel that your life is empty? Yes / No
4. Do you often get bored? Yes / No
5. Are you in good spirits most of the time? Yes / No
6. Are you afraid that something bad is going to happen to you? Yes / No
7. Do you feel happy most of the time? Yes / No
8. Do you often feel helpless? Yes / No
9. Do you prefer to stay at home, rather than going out and doing things? Yes / No
10. Do you feel that you have more problems with memory than most? Yes / No
11. Do you think it is wonderful to be alive now? Yes / No
12. Do you feel worthless the way you are now? Yes / No
13. Do you feel full of energy? Yes / No
14. Do you feel that your situation is hopeless? Yes / No
15. Do you think that most people are better off than you are? Yes / No

Total Score

____

Scoring scheme:
One point for each of these answers:

1. No
2. Yes
3. Yes
4. Yes
5. No
6. Yes
7. No
8. Yes
9. Yes
10. Yes
11. No
12. Yes
13. No
14. Yes
15. Yes

APPENDIX C

1. What is your age?

   ____
2. What is your gender?
   - Female
   - Male
   - Other
3. Do you consider yourself to be Hispanic or Latino?
   - Yes
   - No
   - Not sure
4. Which of the follow do you consider yourself to be?
   - White/Black or African American
   - American Indian or Alaska Native
   - Asian
   - Native Hawaiian or Pacific Islander
   - Other
5. How many years of formal education (including primary school and beyond) do you have?

   ____

APPENDIX D

bed	bed
rest	rest
awake	Awake
tired	tired
dream	dream
wake	wake
snooze	snooze
blanket	BLANKET
doze	Doze
slumber	slumber
snore	snore
nap	nap
peace	peace
yawn	yawn
drowsy	DROWSY

APPENDIX E

The dual-retrieval model used in the current study is:

$$p(CCC) = D(1 - F) + (1 - D)RJ_1J_2J_3 \quad (A1)$$

$$p(CCE) = (1 - D)RJ_1J_2(1 - J_3) \quad (A2)$$

$$p(CEC) = (1 - D)RJ_1(1 - J_2)J_3 \quad (A3)$$

$$p(CEE) = DF + (1 - D)RJ_1(1 - J_2)(1 - J_3) \quad (A4)$$

$$p(ECC) = (1 - D)R(1 - J_1)J_2J_3 \quad (A5)$$

$$p(ECE) = (1 - D)R(1 - J_1)J_2(1 - J_3) \quad (A6)$$

$$p(EEC) = (1 - D)R(1 - J_1)(1 - J_2)J_3 \quad (A7)$$

$$p(EEE) = (1 - D)R(1 - J_1)(1 - J_2)(1 - J_3) + (1 - D)(1 - R) \quad (A8)$$

where D is the probability that the verbatim trace of an item's presentation can be directly accessed on a recall test, R is the probability that an item can be reconstructed on a recall test when the verbatim trace of the item's presentation cannot be accessed, F is the probability that the direct access works in the first recall test but fails in the following recall tests, and J_1, J_2 and J_3 are the probabilities that a reconstructed item is judged to be familiar enough to output on test 1, test 2 and test 3, respectively.

The likelihood function for the data predicted by the dual-retrieval model is:

$$L_6 = \Pi(p_i)^{N(i)} \quad (A9)$$

where p_i is the predicted recall probabilities on the left side of Equations (A1) – (A8), and the $N_{(i)}$ is actual observed data counts. Because six parameter estimates are obtained with the model, one empirical probability is free to vary. Namely, there is one degree of freedom for L_6.

In order to establish goodness of fits, one should compare the likelihood in Equation (A9) to the likelihood of the same data when all seven empirical probabilities are free to vary. The goodness-of-fit test is:

$G^2 = -2\ln[L_6/L_7]$ (A10)

where L_6 is the likelihood of the data predicted by the model, and L_7 is the likelihood of the same data when all empirical probabilities are free to vary. G^2 is distributed asymptotically as the χ^2. Thus, the critical value of rejecting null hypothesis at the .05 confidence level is 3.84.