Fall 1-10-2021

Exploring the Mechanisms that Underlie the Benefits of Retrieval Practice in Younger and Older Adults

Ruth A. Shaffer

WASHINGTON UNIVERSITY IN ST. LOUIS
Department of Psychological & Brain Sciences

Exploring the Mechanisms that Underlie the Benefits of Retrieval Practice in Younger and Older Adults

by
Ruth A. Shaffer

A thesis presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Master of Arts

January 2021
St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

Acknowledgments

I am deeply grateful to my advisor, Kathleen McDermott, for her guidance and support on this project and for her mentorship throughout my time at Wash U. Her commitment to the study of memory has been an example and inspiration for me and others in the Wash U community and beyond. I would also like to thank the other members of my committee, Roddy Roediger and Dave Balota, for their generous advice and encouragement on this project and throughout my time at Wash U in both lab and classroom settings. Wash U has been something of a home to me for the past many years and it is due in large part to the above three mentors—to whom I owe so much—that this is the case. I would also like to thank my present and former labmates, Chris Zerr, Nate Anderson, Thomas Spaventa, Hank Chen, and Adrian Gilmore, for their helpful feedback and advice on this and other projects over the years.

This work was supported by a Collaborative Activity Award from the James S. McDonnell Foundation (Applying Cognitive Psychology to Enhance Educational Practice) awarded to Kathleen McDermott and by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1745038.

Ruth A. Shaffer

Washington University in St. Louis

January 2021

Abstract of Thesis

Exploring the Mechanisms that Underlie the Benefits of Retrieval Practice

in Younger and Older Adults

by

Ruth A. Shaffer

Master of Arts in Psychological & Brain Sciences

Washington University in St. Louis, 2021

Professor Kathleen B. McDermott, Chair

The testing effect—or the benefit of retrieval practice to later memory—is often considered to be a recollection-related phenomenon. However, recent work (Shaffer & McDermott, 2020) has observed a benefit of testing to *both* recollection *and* familiarity processing on both immediate and delayed final tests. Further, although aging populations show marked declines in recollection, older and younger adults often benefit from testing to a similar degree (Meyer & Logan, 2013). This finding suggests that the testing effect in older adults may function via relatively preserved familiarity and lends further support to the notion that the testing effect does not function solely via recollection-related processes. The current project builds on this prior work with the aim of better understanding the mechanisms from the dual process perspective that underlie the testing effect in both younger and older adults. To this end, younger (18-22 yo) and older (65-82 yo) adults studied words, took cued-recall tests on half of the words, and took a final recognition test on all words immediately or 1 day later in which parameter estimates of recollection and familiarity were calculated. At both delays, older and younger adults exhibited a testing effect in both recollection and familiarity, although the magnitude of the testing effect in recollection was smaller for older than for younger adults. These findings add to the growing

literature suggesting that the testing effect can be supported by changes in *both* recollection *and familiarity* processing. Further, whereas the benefit to familiarity appears to persist across age, the benefit to recollection may decline. Implications for theories of the testing effect, as well as for its application in older adults, are explored.

Chapter 1: Introduction

1.1 The Testing Effect

An extensive literature has revealed that taking tests—or retrieval practice—benefits long term retention of material, a finding referred to as the "testing effect" (Roediger & Butler, 2011; Roediger & Karpicke, 2006a, 2006b; Rowland, 2014; Tulving, 1967). Indeed, retrieval practice has proven to be an effective tool for enhancing memory for a wide variety of materials, such as individual words, associations, and text passages (e.g., Carpenter, Pashler, Wixted, & Vul, 2008; Carrier & Pashler, 1992; Karpicke & Roediger, 2007; Roediger & Karpicke, 2006a), and using a wide variety of forms of initial testing, such as cued-recall, free recall, and even, in many cases, recognition testing (e.g., Carpenter, Pashler, & Vul, 2006; McDermott, Agarwal, D'Antonio, Roediger, & McDaniel, 2014; Zaromb & Roediger, 2010). The testing effect also has been validated outside of the laboratory. The benefits of retrieval practice have been found in middle schools, colleges, and medical schools (e.g., Larsen, Butler, & Roediger, 2008, 2013; McDaniel, Anderson, Derbish, & Morrisette, 2007; McDermott et al., 2014) and in both children and older adults (Karpicke, Blunt, & Smith, 2016; Coane, 2013; Meyer & Logan, 2013).

Although the robustness of the testing effect is well established, the mechanisms that underlie this benefit of retrieval practice to later memory remain unclear. This project seeks to explore the mechanisms that underlie the testing effect from the dual process perspective in both younger and older adults. In this Introduction I open with an overview of the literature on the mechanisms from the dual-process perspective that support the testing effect broadly and then move on to a discussion of this topic in older adults, specifically.

1.2 The Dual-Process Perspective and the Testing Effect

From the dual-process perspective, two memorial processes—recollection and familiarity—contribute to performance during tests of memory (Jacoby, 1991; Tulving, 1985; Yonelinas, 2001b, 2002; Yonelinas, Aly, Wang, & Koen, 2010). Recollection is conceptualized as experiential in nature, consisting of effortful or conscious memory for material accompanied by the contextual components of an episode. By contrast, familiarity is conceptualized as an automatic, unconscious form of memory for (or "familiarity" with) material, devoid of the contextual or experiential aspects of the episode in which the item was experienced (Jacoby, 1991; Yonelinas, 2001a, 2002).

1.2.1 A Role for Recollection

With respect to the testing effect, the dominant view has been that retrieval practice benefits later memory by enhancing recollection-related processing and that it leaves familiarity relatively unchanged. Indeed, one prominent theory of the testing effect—the Episodic Context Account—maintains that the memorial benefits following retrieval practice derive from a temporal context updating mechanism (Karpicke, Lehman, & Aue, 2014; Lehman, Smith, & Karpicke, 2014; Whiffen & Karpicke, 2017). The idea is that during retrieval practice elements of the original learning context and testing context combine to provide more probative retrieval cues during later tests of memory. Another prominent theory of the testing effect—the Elaborative Retrieval Hypothesis—suggests that testing induces greater elaboration of items in semantic memory, producing a greater number of retrieval cues to draw upon during final testing (Carpenter, 2009; Carpenter & DeLosh, 2006; Rawson, Vaughn, & Carpenter, 2015). While these accounts differ in many ways (for a recent review, see Karpicke, 2017), when considered

from the dual process perspective, they share an emphasis on recollection-related processing in producing the benefits of retrieval practice to later memory.

Several studies that have directly observed the effects of prior retrieval practice on later estimates of recollection and familiarity have reached the same conclusion. For example, Chan and McDermott (2007) had subjects study lists of words and then take cued-recall tests on some of the words. The authors measured recollection and familiarity on a final test taken a few minutes later using three different methods for obtaining parameter estimates. With all three methods, the authors found that items that had been studied and then tested yielded greater estimates of recollection on the final test than items that had been studied only. The authors found no differences between the two conditions, however, in estimates of familiarity (see also Jones & Roediger, 1995). Verkoeijen, Tabbers, and Verhage (2011) further validated this finding with a restudy control condition (rather than a no test control), suggesting that the benefit of testing to recollective processing was not simply due to re-exposure during testing, but rather to the specific effects of retrieval practice (see also Pu & Tse, 2014; Rowland, 2011).

1.2.2 A Role for Familiarity

Despite this emphasis on recollection, a number of studies more recently have begun to provide evidence that familiarity processes likewise may be involved in the testing effect. Indeed, Shaffer and McDermott (2020) examined the extent to which the findings from prior work, which primarily used retention intervals of 15 minutes or less between initial and final testing, would extend to longer delays (also see Bies-Hernandez, 2013; Guran, Lehmann-Grube, & Bunzeck, 2020, discussed below). The authors reasoned that estimates of recollection and familiarity may decline at differential rates over short intervals (Gardiner & Java, 1991; Yonelinas, 2002) and may change across successive tests (Conway, Gardiner, Perfect, Anderson,

& Cohen, 1997; Dewhurst, Conway, & Brandt, 2009). Further, several studies reveal a benefit of initial testing to estimates of familiarity-related processes on delayed final tests, either statistically or numerically based on calculations from means reported in the articles (e.g., Bies-Hernandez, 2013; Dudukovic, DuBrow, & Wagner, 2009, Exp. 1; Guran et al., 2020; Kessler et al., 2014). However, in all but one case (Bies-Hernandez, 2013), direct parameter estimates of familiarity were not calculated.

In two experiments, Shaffer and McDermott (2020) sought to address these issues by examining the effects of prior testing on immediate, 1-day delayed, and 4-day delayed final tests. The authors used two different commonly used methods for obtaining estimates of recollection and familiarity—confidence ratings fit to the Dual-Process Signal Detection model (Yonelinas, 1994; Yonelinas & Parks, 2007) and Remember-Know responses (Tulving, 1985) using the Independence Remember-Know procedure (Yonelinas, 2002; Yonelinas & Jacoby, 1995). Contrary to the majority of prior work that calculated parameter estimates after short delays, results revealed that testing was accompanied by increased estimates of *both* recollection *and* familiarity on immediate and delayed final tests.

Notably, results from a handful of additional studies have also provided either direct or indirect evidence of a testing effect in familiarity-processing on relatively immediate final tests (Gao et al., 2016; Guran et al., 2020; Jia, Gao, Cui, & Guo, 2019; Roediger & McDermott, 1995). Although in several of these cases the central aim of the study was not to address this question, and direct parameter estimates of familiarity were not always calculated.

Thus, despite the prevailing emphasis on recollection in much of the prior literature, there is emerging evidence that familiarity-processes may likewise be involved in producing the benefits of retrieval practice. However, it remains unclear which factors conspire to determine

whether or not a testing effect in familiarity *will* be revealed, as well as the robustness of this finding. In an examination of the discrepancies that exist in the literature, Shaffer and McDermott (2020) suggest that the testing effect in familiarity may be more readily revealed in conditions in which recollection is broadly reduced or de-emphasized at any stage in the task and must therefore be relied on to a lesser extent for success during final testing. The idea is that retrieval practice may generally enhance both recollection and familiarity processes, though a benefit to familiarity will not be revealed if subjects are able to respond primarily on the basis of recollection (see also Bies-Hernandez, 2013, for an exploration of this issue with respect to the impact of final test format, specifically, and reliance on familiarity or recollection for success; also see Chan & McDermott, 2007, and Bies-Hernandez, 2013, for similar lines of reasoning regarding the importance of reliance on recollection vs. familiarity during final testing).

One aim of the present study is to test the possibility that with reduced recollection a testing effect may be more readily revealed in estimates of familiarity. However, in doing so, it is necessary first to consider another, related puzzle in the literature: the testing effect in older adults.

1.3 The Testing Effect and Aging

Further complicating the picture is research on the testing effect in aging populations. Prior work has revealed marked declines in recollective processing in healthy older adults, with often much smaller declines in or relatively preserved familiarity (e.g., Anderson et al., 2008; Koen & Yonelinas, 2014, 2016; McCabe, Roediger, McDaniel, & Balota, 2009; Pitarque et al., 2016; Prull, Dawes, Martin, Rosenberg, & Light, 2006; Yonelinas, 2002, although this is not always the case, e.g. see Duarte, Ranganath, Trujillo, & Knight, 2006). For example, age-related declines have been observed in memory for associations (Naveh-Benjamin, 2000) as well as for

source information (Johnson, Hashtroudi, & Lindsay, 1993). Studies that have directly obtained parameter estimates of recollection yield consistent conclusions, revealing age-related reductions in recollection across a wide range of estimation procedures: the process-dissociation procedure (e.g., Jennings & Jacoby, 1997), Remember-Know analyses (e.g., Bastin & Van der Linden, 2003), and Receiver-Operating Characteristic (ROC) analyses (e.g., Koen & Yonelinas, 2016). By contrast, although there is some variability in terms of age effects in familiarity, results often reveal much smaller (or entirely absent) age-related reductions in parameter estimates of familiarity (for review and meta-analysis see Koen & Yonelinas, 2014; Prull et al., 2006; also see McCabe et al., 2009).

Critically, however, despite these marked reductions in recollection with healthy aging, prior work often reveals a testing effect in older adults that is similar in magnitude to that observed in younger adults. This result has been observed in prototypical studies of the testing effect, in which items are studied and then tested in a subsequent experimental block (Coane, 2013; Meyer & Logan, 2013; Rogers & Gilbert, 1997; although see Henkel, 2014, who obtained mixed results; and Guran, Herweg, & Bunzeck, 2019; Guran et al., 2020). This result has also been observed in numerous studies in the spaced retrieval literature, in which initial studying and testing occur either in the same experimental block, separated by lags typically on the order of 1 to 10 items, or in short interleaved blocks (Bishara & Jacoby, 2008; Kausler & Phillips, 1988; Kausler & Wiley, 1991; Logan & Balota, 2008; Rabinowitz & Craik, 1986; Maddox & Balota, 2015, with lags > 0; although see Tse, Balota, & Roediger, 2010).

However, to the extent that the testing effect is due to enhanced recollection only, as is suggested by much of the prior literature, how can one explain the existence of such consistent testing effects in a population shown to have marked deficits in recollection? The observance of

a testing effect in older adults similar in magnitude to that found in younger adults suggests one of two (non-mutually exclusive) possibilities:

1.3.1 Benefits via Preserved Familiarity Processing

The first possibility is that the benefits of retrieval practice in older adults accrue primarily via enhanced familiarity, given relatively preserved familiarity processing. This possibility would provide further evidence for the role of familiarity in supporting the benefits of retrieval practice, as well as for the idea that the testing effect in familiarity may be more readily revealed when recollection is reduced and must be relied on to a lesser degree for success during final testing. Some support for this possibility comes from the spaced retrieval training literature (Camp, Foss, Stevens, & O'Hanlon, 1996; Camp & Schaller, 1989; Cherry, Simmons, & Camp, 1999), in which older adults with severe memory deficits (e.g., Alzheimer's Disease) retrieve material at increasingly large intervals, with lags beginning at a few seconds and eventually approaching weeks. The idea is that this form of retrieval practice capitalizes on automatic processing preserved in these older adult populations for success.

More direct evidence for this possibility comes from a study that examined the impact of spaced retrieval on associative memory intrusions in healthy older and younger adults (Bishara & Jacoby, 2008). Bishara and Jacoby (2008) had subjects study and take cued-recall tests on word pairs after lags of 1 to 6 intervening items. In their second experiment, the authors used an opposition procedure such that increased familiarity (or automatic processing) as a result of prior testing would serve to increase the number of intrusions on a final test, whereas increased recollection would serve to reduce the number of these intrusions. Results revealed that initial testing increased the occurrence of intrusions in older adults but not in younger adults. This finding suggested that spaced retrieval had enhanced familiarity in older adults without

commensurate increases in recollection. Put another way, results supported the notion that the benefit of retrieval practice in older adults may differ from that of younger adults and derive specifically from improved familiarity.

Critically, however, in this study and in much of the spaced retrieval literature, initial testing occurs within the same experimental block as initial study, with lags of only a few items between studying and initial testing. This differs from other testing effects literature in that success on initial tests may not require much remembering from long term memory. As a result, the processes required for success in initial testing—and those subsequently enhanced as a result of initial testing—may differ from those observed when initial testing is performed in a separate block following studying. Thus, further work is required to connect this finding to the broader testing effect literature.

1.3.2 Ameliorating Existing Deficits in Recollection Processing

A second possibility is that the testing effect in older adults functions by ameliorating existing deficits in recollection. This possibility would comport with prior work suggesting that the benefits of retrieval practice derive primarily from enhanced recollection only. Some support for this possibility comes from work that has shown improved recollection-related processing in older adults via an expanding retrieval opposition procedure training (e.g., Jennings & Jacoby, 2003). However, the goal of this training was to improve recollective *processing* in general, rather than to improve recollection for any specific items tested.

The most direct evidence to suggest that the testing effect in older adults functions primarily via enhanced recollective processing comes from a recent study in which the authors used the Remember-Know procedure (Tulving, 1985; Gardiner, 1988) to explore the effect of retrieval practice in both younger and older adults (Guran et al., 2020). In the Remember-Know

procedure, a Remember response is thought to index recollection, and a Know response is thought to index familiarity in the absence of recollection (see Yonelinas, 2001b, for discussion). Across both age groups and on both immediate and delayed final tests, the authors found a large testing effect in Remember responses and a much smaller, although still significant, testing effect in Know responses. The authors interpret the finding as evidence of a primary role for recollection in supporting the testing effect in both younger and older adults.

Critically, however, the authors never calculate parameter estimates of familiarity, and instead use Know responding only as a proxy for familiarity. To the extent that prior testing enhances recollection, however, more Remember responses are made to items previously tested. As a result, fewer Know responses can be made to these items, even if familiarity is experienced. This situation would serve to artificially reduce the observed testing effect in Know responses, even when a large testing effect in familiarity may have existed. Thus, had the authors calculated parameter estimates of familiarity (via the Independence Remember-Know Procedure, Yonelinas, 2002, Yonelinas & Jacoby, 1995; see Supplementary Materials for standard calculations), results may have suggested a strong testing effect in both recollection and familiarity. Several methodological decisions further limit the generalizability of the study's findings (e.g., frequent task switching during initial testing and restudying, the use of pictorial stimuli, and initial recognition testing). Thus, although the study provides preliminary evidence to suggest that the testing effect in older adults may function via enhanced recollection, further work is required before strong conclusions can be drawn.

Of course, the possibilities that retrieval practice benefits older adults via enhanced familiarity or by alleviating a deficit in recollection are not mutually exclusive. However, observing benefits of retrieval practice in older adults *primarily* in one process or the other or

observing a change across age in the mechanisms that support the testing effect would have important implications for aging populations. Specifically, a greater understanding of the mechanisms that underlie the testing effect in older adults would aid in predicting the conditions under which retrieval practice is more or less likely to produce tangible memorial benefits in this population. If healthy older adults reliably show improvements in recollection at both short and long delays as a result of prior testing, this would argue strongly for the use of retrieval practice in aging populations as a tool for remediating declines in recollective-related processing—such as for improving memory for associative information or for improving the outcomes of source monitoring. However, to the extent that the effects of prior testing in older adults are driven more so by enhanced familiarity, as suggested by findings from the spaced retrieval literature (Bishara & Jacoby, 2008), greater constraints would be placed on the expected benefits of retrieval practice in aging populations.

Further, understanding the mechanisms that support the testing effect in older adults has implications for our understanding of the mechanisms that underlie the testing effect more broadly. In an effort to evaluate the discrepancies in the literature, Shaffer and McDermott (2020) suggest several factors that may impact whether or not a testing effect in familiarity is revealed on a final test. One suggestion, noted above, is that a testing effect in familiarity may be more readily revealed in conditions in which recollection is reduced. The idea is that testing may generally enhance both recollection and familiarity processes, but that a benefit to familiarity will not be revealed if subjects are responding primarily on the basis of recollection (see Bies-Hernandez, 2013, and Chan & McDermott, 2007, for related arguments). One test of this hypothesis would be to examine the mechanisms that underlie the testing effect in a population shown to have reduced recollective processing, with relatively preserved

familiarity—in this case, older adults. The key question would be whether older adults, relative to younger adults, exhibit a testing effect of greater magnitude in estimates of familiarity, given reduced ability to rely on recollection for success on the final test.

1.4 The Present Study

To explore these issues, the current study examines the impact of prior retrieval practice on estimates of recollection and familiarity in both younger and older adults. In doing so, this project seeks to address three aims:

1. The *first* and most basic aim is to examine the replicability of the findings from Shaffer and McDermott (2020) revealing a testing effect in *both* recollection *and* familiarity on both immediate and delayed final tests. The experiments in Shaffer and McDermott (2020) were conducted online via Amazon Mechanical Turk (MTurk), whereas much of the prior work on the topic has been conducted in more controlled laboratory settings and with undergraduate student populations. Thus, for comparison with prior work, as well as in order to ensure that the prior findings were not the result of the idiosyncrasies of online samples, the first aim of this master's thesis is to replicate our prior findings in an in-lab, undergraduate sample.

2. The *second aim* is to extend these questions to an older adult sample, in order to address the puzzle of why older adults often reveal a testing effect similar in magnitude to that of younger adults. Does retrieval practice for older adults ameliorate existing deficits in recollection or function primarily via preserved familiarity processing? As described above, the evidence in the literature is mixed, and to our knowledge no work yet has examined this question directly with parameter estimates of recollection and familiarity.

3. Finally, the *third aim* is to address the suggestion made in Shaffer and McDermott (2020): that the testing effect in familiarity may be more readily revealed when recollection is reduced and

must therefore be relied on to a lesser extent for success on the final test. This prediction is tested 1) by examining the mechanisms that underlie the testing effect in a population that exhibits reduced recollection with relatively preserved familiarity (the older adult sample here), and 2) by directly exploring the relation between estimates of overall recollection and the magnitude of the observed testing effect in estimates of familiarity.

In order to address these three aims, undergraduate students at Washington University in St. Louis and community-dwelling older adults (65+ years old) studied words and took cued-recall tests on half of the words. Subjects then took a final recognition test that included all of the old words and an equal number of new words. To examine the extent to which the results were robust to the retention interval between initial and final testing, half of the subjects completed the final test immediately and half completed the final test after a 1-day delay. Estimates of recollection and familiarity on the final recognition test were obtained via the Remember-Know-New procedure (described in detail below). In order to screen for cognitive impairment, older adults additionally completed the Mini-Mental State Examination (Folstein, Folstein, & McHugh, 1975).

Chapter 2: Method

2.1 Participants

2.1.1 Younger Adults

Sixty-two undergraduate subjects from the Washington University Psychological and Brain Sciences Undergraduate Research Participation Pool took part in the study (no delay condition: 31; 1-day delay condition: 31). Inclusion criteria indicated that the subject must be at least 18 years old, a native English speaker, and have normal or corrected-to-normal vision. One subject in the 1-day delay condition was excluded for failing to meet this inclusion criteria. In addition, one subject in the no delay condition was excluded for having response times of under 250ms on many final test trials (74 of 240 possible responses were made in under 250ms, indicating a lack of engagement in the task). Thus, the final sample included 60 participants (see Table 2.1): no delay condition (N = 30, mean age = 19.4 y, standard deviation (SD) age = 1.2 y, age range = 18 – 22 y, female (F) = 22, education = 13.4 y) and 1-day delay condition (N = 30, mean age = 19.7 y, SD age = 1.3 y, age range = 18 – 22 y, F = 22, education = 13.7 y). Three subjects in the included sample made 1 final test recognition response each in under 250ms; calculations excluded the specific trial for the 3 subjects. One subject pressed start before the experimenter had provided the instructions and indicated observing the first stimulus. After restarting the experiment, this stimulus was removed from all analyses for this subject. In a small number of cases, a test trial failed to log, in which case relevant calculations were made with one fewer trial for the subject (for 1 subject, 1 of the 60 initial test trials failed to log; for 4 subjects, 1 of the 240 final test trials failed to log). Subjects received 1 hour of course credit or $10.00 for their participation.

2.1.2 Older Adults

Sixty-five older adult subjects from the St. Louis area community took part in the study (no delay condition: 32; 1-day delay condition: 33). Subjects were recruited via the Volunteer for Health registry (https://sites.wustl.edu/wuvfh/). All subjects were at least 65 years old, had normal or corrected-to-normal vision, and were native English speakers. Subjects also completed the Mini-Mental State Examination (MMSE; Folstein et al., 1975) in order to assess potential cognitive impairment. All subjects scored above the standard single cutoff score of 24. One subject in the 1-day delay condition was unable to attend the second session of the study and was excluded from analyses. Thus, the final sample included 64 participants (see Table 2.1): no delay condition (N = 32, mean age = 70.7 y, standard deviation (SD) age = 4.7 y, age range = 65 – 82 y, female (F) = 23, education = 17.0 y) and 1-day delay condition (N = 32, mean age = 71.8 y, SD age = 4.2 y, age range = 65 – 81 y, F = 22, education = 16.3 y). One subject pressed start before the experimenter had provided the instructions and observed 2 stimuli. After restarting the experiment, these 2 stimuli were removed from all analyses for this subject. In a small number of cases, a test trial failed to log, in which case relevant calculations were made with one fewer trial for the subject (for 2 subjects, 1 of the 60 initial test trials failed to log; for 2 subjects, 1 of the 240 final test trials failed to log).

Subjects received $20.00 in the no delay condition and $25 in the 1-day delay condition for their participation in the study ($15 for completion of the study and $5 per visit to the Psychology Building to defray the costs of transportation to the study).

For both younger and older adult groups, subjects consented to participation before beginning the study and were debriefed upon completion of the study. The experiment was conducted in accordance with the Washington University in St. Louis Institutional Review Board.

Table 2.1 Demographic information broken down by age group (older, younger) and delay condition (no delay, 1-day delay). Age, education, and MMSE score are calculated as means (standard deviations). MMSE = Mini-Mental State Examination.

	Younger		Older	
	No Delay	**1-Day Delay**	**No Delay**	**1-Day Delay**
N	30	30	32	32
Age (Years)	19.4 (1.2)	19.7 (1.3)	70.7 (4.7)	71.8 (4.2)
Age Range (Years)	18 - 22	18 - 22	65 - 82	65 - 81
Sex (Female/Male)	22/8	22/8	23/9	22/10
Education (Years)	13.4 (1.2)	13.7 (1.4)	17.0 (3.1)	16.3 (3.3)
MMSE score	NA	NA	29 (1.2)	28.9 (1.2)
Hispanic/Latino (Yes/No/NA)	5/25/0	4/24/2	0/28/4	1/28/3
Race (N)				
Asian	4	7	0	0
Black / African American	3	2	2	5
Caucasian	18	19	28	26
Native Hawaiian / Pacific Islander	1	0	0	0
More than one race	3	2	1	1
Other	1	0	1	0

2.2 Materials

Stimuli were the same as those used in Shaffer and McDermott (2020). Specifically, stimuli consisted of 240 words obtained from the English Lexicon Project database (Balota et al., 2007). Stimulus selection parameters emulated those of Chan and McDermott (2007; Experiment 3): Kučera-Francis Frequency = 5 – 200; parts of speech = noun, adjective, and/or verb; word length = 4 – 9 letters. No words had the same first 3 letters and average frequency constraints matched those of Chan and McDermott (2007). Specifically, 8 lists of 30 words were constructed so that each list's average Kučera-Francis Frequency fell within 34.33 – 35. List

placement within the experiment was counterbalanced in an effort to place each list in the Test, No Test, and New (on the final test) conditions approximately 25%, 25%, and 50% of the time, respectively. The stimulus order within a given experimental block was newly randomized for each subject and session.

2.3 Procedure

The experiment consisted of two sessions, either completed back-to-back in a single day (no delay condition) or completed separately on two consecutive days (1-day delay condition). Subjects were tested individually and completed the experimental tasks in a laboratory office suite setting on an iMac computer (with the exception of the Mini-Mental State Examination, which was administered by the experimenter). In the first session, subjects studied words and took 3-letter stem cued-recall tests. In the second session, memory for previously studied, tested, and new items was probed on a recognition test. Older adults additionally completed the Mini-Mental State Examination (MMSE; Folstein et al., 1975) at the end of the second session. The experimental procedure follows that of Shaffer and McDermott (2020) and is based on Chan and McDermott (2007) Experiment 3. The procedure is described in detail below (see Fig. 2.1 for a depiction of the experimental design).

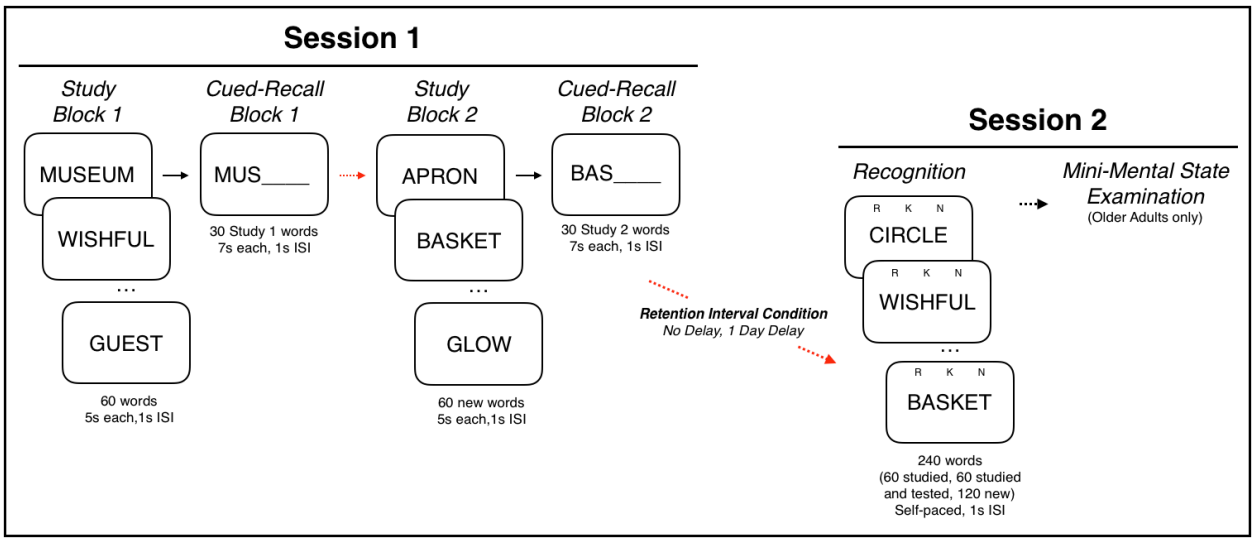

Figure 2.1 Experimental Design. In Session 1, subjects studied 60 words and took a 3-letter stem cued-recall test on half of the studied words. Subjects then completed an identical study-test cycle with a new set of words. Immediately (no delay condition) or 1-day later (1-day delay condition), subjects took a final Remember-Know-New recognition test that included all of the old words and an equal number of new words. Older adults additionally completed the Mini-Mental State Examination (Folstein, Folstein & McHugh, 1975).

Session 1

Subjects were instructed that during the experiment they would be shown a set of words and asked to remember the words for a later memory test. Before seeing the set of words, specific instructions regarding the nature of the memory test were provided (see Supplementary Materials for the complete set of Session 1 instructions):

> *"Specifically, we will later provide you with the first 3 letters of some of the words you see. We will ask that you type the ENTIRE word from the list of words you saw in the blank provided. If you cannot remember the word you saw earlier, just leave it blank. For example, on the test you might see 'TRA_____' if you had earlier studied the word 'TRACTOR', and so you would type 'TRACTOR' (the entire word) in the blank provided."*
>
> After listening to the instructions and asking the experimenter any questions, subjects began the task. Initial learning condition (test vs. no test) was manipulated within subjects and was divided into two blocks of study-test cycles in order to improve initial test performance. In the first initial learning block, 60 words were visually presented one at a time for 5s each (1s

ISI). Subjects then took a cued-recall test on half of the words from the preceding list (30 words). During the cued-recall test, subjects were shown the first 3 letters of a word and were given 7s to type the complete word in the space provided (1s ISI). Subjects were instructed to leave the space empty if they could not remember the previously studied word. In the second block, subjects completed an identical study session with a new set of 60 words and an identical cued-recall test on half of the new set of 60 words. No feedback was provided during the cued-recall tests. Once initial learning was complete, subjects gave ratings of both difficulty and effort for the task.

Retention Interval

Subjects began the Session 2 recognition test either 1) immediately after completing Session 1 (younger adults: average delay = 7.81 minutes, SD = 0.78 min; range = 6.63 min – 10.80 min; older adults: average delay = 9.32 min, SD = 1.94 min; range = 7.00 min – 16.20 min); or 2) 1 day later (younger adults: average delay = 0.99 days, SD = 0.003 days, range = 0.98 days – 0.99 days; older adults: average delay = 0.99 days, SD = 0.01 days, range = 0.97 days – 1.00 days).

Session 2

Instructions. Prior to completing the final recognition test, subjects were provided with extensive instructions regarding the Remember-Know-New procedure (see the Supplementary Materials for the complete set of Remember-Know instructions). Specifically, subjects were instructed to indicate for each word whether or not they had seen it in the previous section of the experiment, with two response options (Remember, Know) for words believed to have been previously presented in the experiment. Excerpts from the instructions are provided below:

> ***Remember:*** *"If you respond with a "remember" judgment, this means that you remember something specific about having seen this word before in the experiment. For*

example, you might remember a specific thought you had or connection you made when you originally saw this word in the experiment. You may visually remember seeing the word. In essence, you're indicating that you can consciously recall specific parts of the experience you had when you saw the word in the previous section."

Know: *"If you respond with a "know" judgment, this means that you know that you've seen the word before in the experiment, but you cannot remember the specific details regarding seeing the word before in the experiment. In other words, this response is telling us that you have a gut feeling that you saw the word, but you don't have a conscious recollection of "seeing" or "experiencing" it in the last section."*

New: *"If you respond with a "new" judgment, this means that you did not see the word in the previous part of the experiment."*

In accordance with recommendations in the literature (Koen & Yonelinas, 2016), additional instructions indicated that the subject should only provide a Remember response if the subject would be able to indicate the detail leading to the Remember response, if asked:

"In essence, a REMEMBER response indicates that you believe the word was previously presented in the experiment AND you can provide some specific detail about your experience of seeing that word earlier in the experiment. You should only provide a REMEMBER response if, when asked, you would be able to indicate the detail that is leading you to make a REMEMBER response."

"If you believe a word was presented earlier in the experiment but you DO NOT recall a specific detail from seeing the word before in the experiment and COULD NOT provide the detail if asked, then you should make a KNOW response. This means that you KNOW the word was presented earlier in the experiment, but you don't have any conscious recollection of specific details of seeing that word earlier in the study."

Because misunderstandings are common and can muddy interpretation of the results, after going over the Remember-Know instructions verbally, the subject was asked to explain each response type (Remember, Know, New) to the experimenter. In the event that the subject's response indicated a lack of understanding or clear distinction between the Remember and Know response options, discussion continued until the misunderstanding was resolved.

Recognition Test. During the final recognition test, subjects were shown all 120 words they had previously studied (half of which had also been tested), along with 120 new words.

Stimuli were presented sequentially, and the presentation order was newly randomized for each subject. For each word presented, subjects were asked to indicate their memory for the stimulus by making a Remember, Know, or New response via the 7-8-9 keys on the keyboard.

Once the recognition test was complete, subjects provided difficulty and effort ratings for the task, indicated whether or not they would like to be contacted for future participation in studies, and, for subjects in the 1-day delay condition, indicated whether or not they had noted down any words from the experiment after the previous day's session. No subjects indicated noting down words after the prior session.

Cognitive Test. Following the recognition test, older adults additionally completed the Mini-Mental State Examination (MMSE; Folstein et al., 1975) in order to assess potential cognitive impairment.

Chapter 3: Results

For all main text ANOVAs below, as well as for the multiple regression analyses under Aim 3, if outliers were detected in the groups of interest, a duplicate analysis excluding outliers was conducted. Specifically, outliers were defined as subject-level means that were over 3 SDs above or below the group means for the original analysis in each section. Any differences in the patterns of results after excluding outliers is reported in the results below.

In addition, for clarity and completeness, for all main text analyses pertaining to the primary questions of interest (i.e. Session 1 cued-recall performance, Session 2 accuracy, and Session 2 parameter estimates), two sets of supplementary analyses were conducted. Specifically, supplementary analyses were conducted that excluded subjects who were noted to have struggled initially with the Session 2 Remember-Know instructions (7 subjects total; 1 younger adult in the 1-day delay condition; 3 older adults from each of the no delay and the 1-day delay conditions). Of note, before beginning the final recognition test all subjects—even those who had struggled initially—provided verbal explanations to the experimenter that revealed an accurate understanding of the Remember-Know task instructions. In addition, analyses were conducted that excluded subjects who displayed potential guessing behavior on the initial cued-recall tests (35 subjects total; 6 younger adults in the no delay condition, 5 younger adults in the 1-day delay condition, 11 older adults in the no delay condition, and 13 older adults in the 1-day delay condition). Specifically, subjects were instructed not to guess on the initial cued-recall tests and to simply leave the item blank if they could not remember the studied word that corresponded to the 3-letter cue. In this additional supplementary analysis, subjects were excluded for potential guessing behavior if the proportion of cued-recall items given incorrect responses was both 1) greater than or equal to their proportion of items left blank

and 2) greater than or equal to .2. The idea behind this criterion for guessing was: 1) only to include subjects who had fewer incorrect responses than items left blank, or 2) if this was not the case, to include subjects if they had a very low overall proportion of incorrect, potential guessing responses (< .2).

For both sets of supplementary analyses, any changes to the patterns of results are described in detail the Supplementary Materials. Additional notes are included in the main text for any major change relevant to the central analysis of interest. To preview the results of both supplementary analyses, the majority of patterns of results remained the same.

3.1 Younger Adults

The first aim was to examine the extent to which prior work in an online MTurk sample would replicate in an in-lab, undergraduate sample. Specifically, will the undergraduate sample exhibit a testing effect in both recollection and familiarity at both immediate and delayed final tests? To this end, performance in the undergraduate, younger adult sample was examined first.

3.1.1 Session 1

Performance on the initial cued-recall test was first examined 1) to confirm above floor performance such that a testing effect should be expected on the final recognition test and 2) to confirm that participants in the no delay and 1-day delay conditions did not significantly differ in performance prior to the delay manipulation. Initial test responses were scored manually by the experimenter. A lenient criterion was used, such that clear misspellings, other forms of the correct word, and incomplete words in which the intended word was clear were scored as correct. The following are examples of items scored as correct: 1) misspellings: “LEMMON” instead of “LEMON”; 2) other forms: “ROUNDING” instead of “ROUNDED”; and 3) incomplete words**:** “ROADWA” instead of “ROADWAY.”

Figure 3.1 displays average proportion correct on the first and second initial test blocks for younger adults in the no delay and 1-day delay conditions, respectively (see Supplementary Material Figures S1-S3 for all subject-level data). As is apparent in the figure, initial test performance was above floor, was comparable for subjects in the two delay conditions, and improved from the first to the second initial test block, which was likely a result of practice effects.

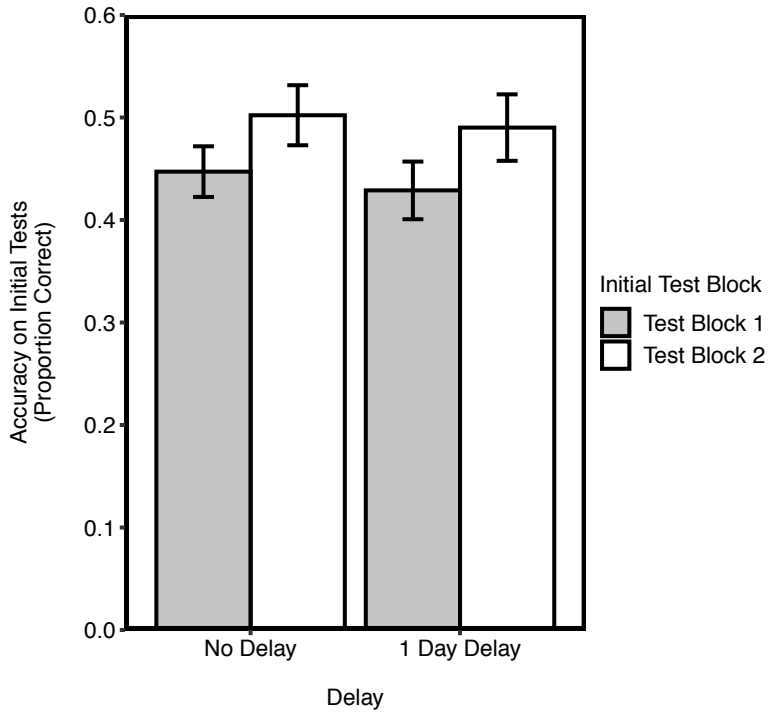

Figure 3.1 Initial cued-recall test accuracy: Younger adults. Proportion correct on the initial cued-recall tests (mean ± SE) is displayed for younger adults by delay condition (no delay, 1-day delay) and test block (block 1, block 2). Initial test performance was comparable for subjects in the two delay conditions.

A Two-Way Mixed ANOVA formally examined the effect of delay (between-subjects: no delay, 1-day delay) and initial test block (within-subjects: block 1, block 2) on initial test performance (measured as proportion correct cued-recall responses). The main effect of delay was not significant ($F(1, 58) = 0.17$, $p = .679$, $\eta_p^2 = .003$), such that initial test performance did not significantly differ for subjects the no delay and 1-day delay conditions. The main effect of

initial test block, however, was significant ($F(1, 58) = 10.40$, $p = .002$, $\eta_p^2 = .15$), such that performance improved from the first ($M = .44$) to the second ($M = .50$) test block (likely a result of practice effects). Finally, the interaction between delay and initial test block was not significant ($F(1, 58) = 0.03$, $p = .863$, $\eta_p^2 = .001$).

3.1.2 Session 2

Accuracy. To examine the magnitude of the testing effect, final recognition test accuracy was calculated for items previously tested and items previously studied only (no test condition). The primary calculation of accuracy was hit rate minus false alarm rate (i.e. where hits and false alarms were defined as the proportion of Remember and Know responses to old and new items, respectively). An alternative measure of accuracy, *d'*, was calculated for comparison and to adjust for potential differences in response bias across subjects. Using *d'*, all patterns of results remained the same. Thus, only accuracy in terms of hits minus false alarms is reported below.

Figure 3.2 displays average accuracy on the final recognition test for younger adults by initial learning condition and delay. As is evident in the figure, a testing effect was observed for younger adults in-lab on both the immediate and 1-day delayed final tests. Forgetting occurred across the delay, such that performance was worse on the 1-day delayed final test than on the immediate final test. However, the magnitude of the testing effect was stable over time.

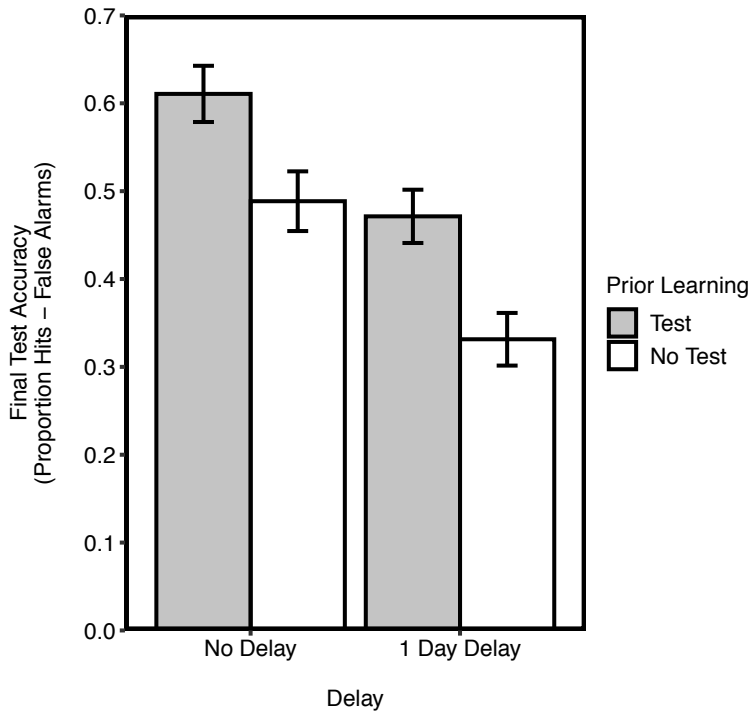

Figure 3.2 Recognition test accuracy: Younger adults. Proportion hits minus false alarms (mean ± SE) on the final recognition test is displayed for younger adults by initial learning condition (test, no test) and delay (no delay, 1-day delay). A testing effect occurred at both delays.

A Two-Way Mixed ANOVA formally examined the effect of initial learning condition (within-subjects: test, no test) and delay (between-subjects: no delay, 1-day delay) on final test accuracy. The main effect of initial learning condition was significant ($F(1, 58) = 96.29$, $p < .001$, $\eta_p^2 = .62$), revealing a testing effect in which performance was greater for items previously tested ($M = .54$) than for items previously studied only (no test; $M = .41$). The main effect of delay was also significant ($F(1, 58) = 12.08$, $p < .001$, $\eta_p^2 = .17$), such that performance decreased from the immediate ($M = .55$) to the 1-day delayed ($M = .40$) final test. However, the interaction between initial learning condition and delay was not significant ($F(1, 58) = 0.44$, $p = .509$, $\eta_p^2 = .01$), such that the magnitude of the testing effect did not significantly differ as a function of delay.

Parameter estimates. The preceding analysis verified that the undergraduate, younger adult sample exhibited a testing effect in overall accuracy on the final recognition test. With respect to Aim 1, the central question is the extent to which this observed testing effect in the younger adult, in-lab sample can be attributed to changes in recollection and/or familiarity—and, specifically, whether the results in this undergraduate sample would replicate the finding that familiarity contributes to the testing effect, as observed in Shaffer and McDermott (2020). To address this question, the following analysis examines the magnitude of the testing effect in recollection and familiarity on both the no delay and 1-day delayed final test. The Independence Remember-Know Procedure (Yonelinas, 2002; Yonelinas & Jacoby, 1995) was used to obtain estimates of recollection and familiarity from Remember, Know, and New recognition responses (see the Supplementary Materials for the formulas used).

Figure 3.3 displays average process estimates on the final recognition test for younger adults by parameter, initial learning condition, and delay. As can be seen in the figure, at both delays a testing effect was observed in estimates of both recollection and familiarity. However, the magnitude of the testing effect was larger for estimates of recollection than for familiarity.

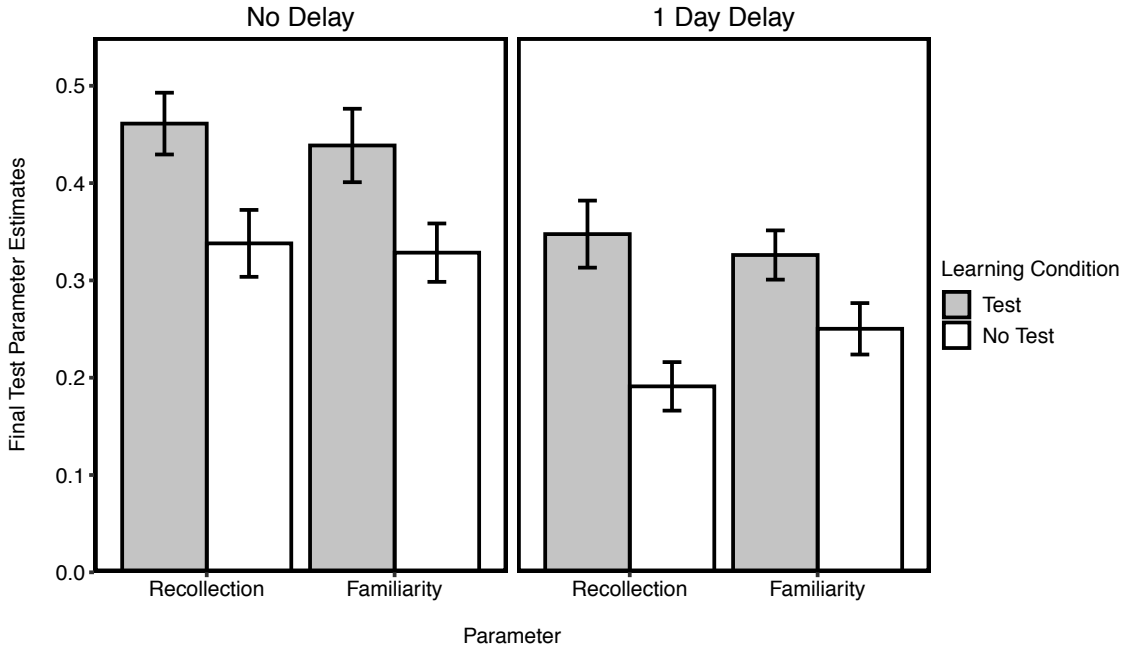

Figure 3.3 Recognition test parameter estimates: Younger adults. Process estimates (mean ± SE) on the final recognition test are displayed for younger adults by parameter (recollection, familiarity), initial learning condition (test, no test), and delay (no delay, 1-day delay). A testing effect occurred in estimates of both recollection and familiarity, with a larger effect in recollection.

A Three-Way Mixed ANOVA formally examined the effect of parameter (within-subjects: recollection, familiarity), initial learning condition (within-subjects: test, no test), and delay (between-subjects: no delay, 1-day delay) on process estimates during the final recognition test. The three-way interaction was not significant ($F(1, 58) = 2.63$, $p = .110$, $\eta_p^2 = .04$), nor were the two-way interactions between initial learning condition and delay ($p = .982$) or between parameter and delay ($p = .445$).

However, the two-way interaction between parameter and initial learning condition was significant ($F(1, 58) = 5.03$, $p = .029$, $\eta_p^2 = .08$). To explore the two-way interaction further, the effect of initial learning condition was examined separately for estimates of recollection and familiarity, collapsed across delay. There was a significant effect of initial learning condition for estimates of recollection ($F(1, 59) = 102.99$, $p < .001$, $\eta_p^2 = .64$; $M_{test} = .40$; $M_{notest} = .26$) and for

estimates of familiarity ($F(1, 59) = 32.89$, $p < .001$, $\eta_p^2 = .36$; $M_{test} = .38$; $M_{notest} = .29$), revealing that both recollection and familiarity estimates were enhanced by prior testing. However, the magnitude of the testing effect was larger for estimates of recollection than of familiarity.

Finally, in the full Three-Way Mixed ANOVA above, a main effect of delay was revealed such that estimates of recollection and familiarity decreased from the immediate ($M = .39$) to the 1-day delayed ($M = .28$) final test ($F(1, 58) = 10.83$, $p = .002$, $\eta_p^2 = .16$).

In sum, results from the in-lab undergraduate sample replicated prior work on MTurk, revealing a testing effect in estimates of both recollection and familiarity on both the immediate and 1-day delayed final tests (although the magnitude of the testing effect was greater in recollection). Moreover, the magnitude of the testing effect did not change significantly across the delay.

3.2 Older Adults

The second aim was to address the question of why older adults often reveal a testing effect similar in magnitude to that of younger adults. Does retrieval practice ameliorate existing deficits in recollection or function primarily via enhanced familiarity in older adults? To this end, performance in the older adult sample was examined next.

3.2.1 Session 1

As in the younger adult analysis, performance on the initial cued-recall test was first examined 1) to confirm above floor performance such that a testing effect should be expected on the final recognition test and 2) to confirm that participants in the no delay and 1-day delay conditions did not significantly differ in performance prior to the delay manipulation. Initial test responses were scored manually by the experimenter using the same lenient criterion scoring method as detailed above.

Figure 3.4 displays average proportion correct on the first and second initial test blocks for older adults in the no delay and 1-day delay conditions, respectively. As can be seen in the figure, initial test performance was above floor, was comparable for subjects in the two delay conditions, and improved from the first to the second initial test block (particularly in the 1-day delay condition), which was likely a result of practice effects.

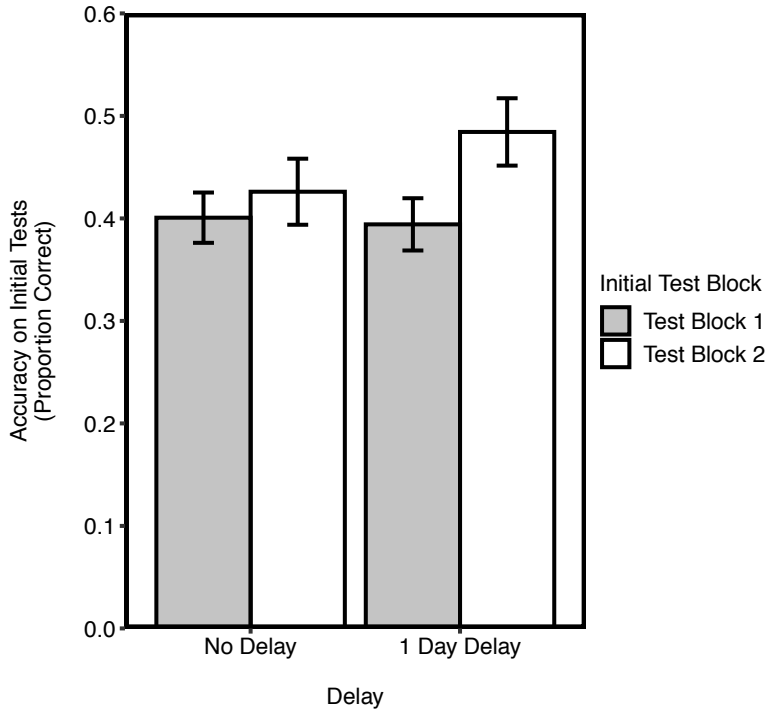

Figure 3.4 Initial cued-recall test accuracy: Older adults. Proportion correct on the initial cued-recall tests (mean ± SE) is displayed for older adults by delay condition (no delay, 1-day delay) and test block (block 1, block 2). Initial test performance was comparable for older adults in the two delay conditions.

A Two-Way Mixed ANOVA formally examined the effect of delay (no delay, 1-day delay) and initial test block (block 1, block 2) on initial test performance. The main effect of delay was not significant ($F(1, 62) = 0.49$, $p = .485$, $\eta_p^2 = .01$), such that initial test performance did not significantly differ for subjects in the no delay and 1-day delay conditions. The main effect of initial test block, however, was significant ($F(1, 62) = 10.23$, $p = .002$, $\eta_p^2 = .14$), such that performance improved from the first ($M = .40$) to the second ($M = .46$) test block, likely a

result of practice effects. Finally, the interaction between delay and initial test block failed to reach significance ($F(1, 62) = 3.22$, $p = .077$, $\eta_p^2 = .05$). However, a marginally significant effect ($p = .077$) suggested that subjects in the 1-day delay condition, relative to the no delay condition, may have improved more from the first to the second initial test block. Follow-up t-tests revealed that in the no delay condition, subjects did not improve significantly from the first to the second initial test block ($p = .268$), whereas there was significant improvement from the first to the second initial test in the 1-day delay condition ($p = .003$).

3.2.2 Session 2

Accuracy. The next analysis examined the extent to which the present study would replicate prior work suggesting that older adults often exhibit a testing effect. To examine the magnitude of the testing effect, final recognition test accuracy was calculated for items previously tested and items previously studied only. As in the younger adult analysis, accuracy was calculated both as hit rate minus false alarm rate and as *d'*. However, as before, all patterns of results remained the same when using *d'*. Thus, only accuracy in terms of hits minus false alarms is reported below.

Figure 3.5 displays average accuracy on the final recognition test for older adults by initial learning condition and delay. As is apparent in the figure, a testing effect was observed for older adults on both the immediate and 1-day delayed final tests. Forgetting occurred across the delay, such that performance was worse on the 1-day delayed final test than on the immediate final test. However, the magnitude of the testing effect was stable over time.

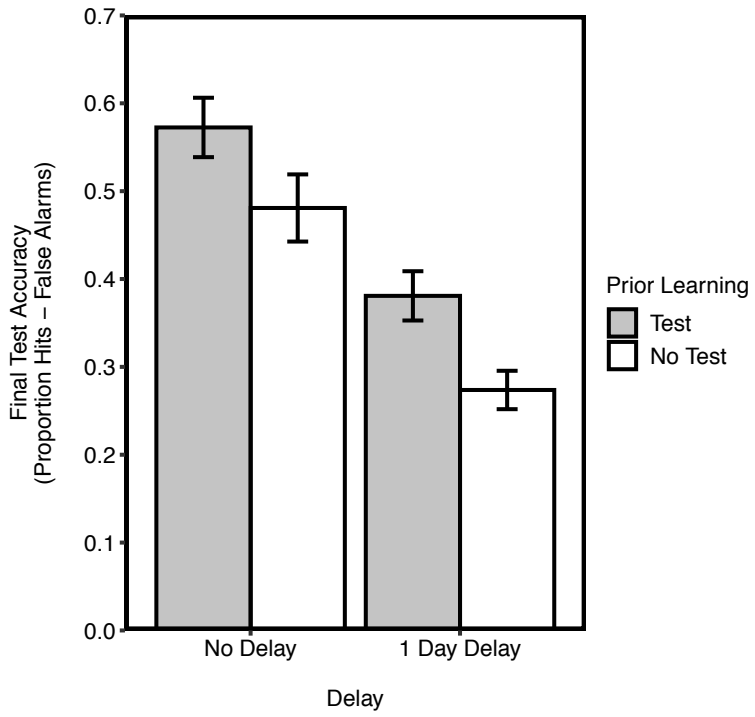

Figure 3.5 Recognition test accuracy: Older adults. Proportion hits minus false alarms (mean ± SE) on the final recognition test is displayed for older adults by initial learning condition (test, no test) and delay (no delay, 1-day delay). A testing effect occurred at both delays.

A Two-Way Mixed ANOVA formally examined the effect of initial learning condition (test, no test) and delay (no delay, 1-day delay) on final test accuracy. The main effect of initial learning condition was significant ($F(1, 62) = 73.23$, $p < .001$, $\eta_p^2 = .54$), revealing a testing effect in which performance was greater for items previously tested ($M = .48$) than for items previously studied only (no test; $M = .38$). The main effect of delay was also significant ($F(1, 62) = 22.11$, $p < .001$, $\eta_p^2 = .26$), such that performance decreased from the immediate ($M = .53$) to the 1-day delayed ($M = .33$) final test. However, the interaction between initial learning condition and delay was not significant ($F(1, 62) = 0.44$, $p = .508$, $\eta_p^2 = .01$), such that the magnitude of the testing effect did not significantly differ as a function of delay. Thus, the above analysis replicated prior work, revealing a robust testing effect in recognition accuracy in an older adult sample.

Parameter Estimates. With respect to Aim 2, the central question concerns the extent to which the observed testing effect in the older adult sample can be conceptualized as alleviating a recollection deficit and/or as functioning through relatively preserved familiarity. The first step in addressing this question is to examine whether, in the older adult sample, a testing effect is observed in estimates of recollection, familiarity, or both processes. Thus, the following analysis examines the magnitude of the testing effect in recollection and familiarity on both the no delay and 1-day delayed final test for the older adult sample. As before, the Independence Remember-Know Procedure (Yonelinas, 2002; Yonelinas & Jacoby, 1995) was used to obtain estimates of recollection and familiarity from Remember, Know, and New recognition responses.

Figure 3.6 displays average process estimates on the final recognition test for older adults by parameter, initial learning condition, and delay. As is evident in the figure, a testing effect (of similar magnitude) was observed in estimates of both recollection and familiarity at both delays.

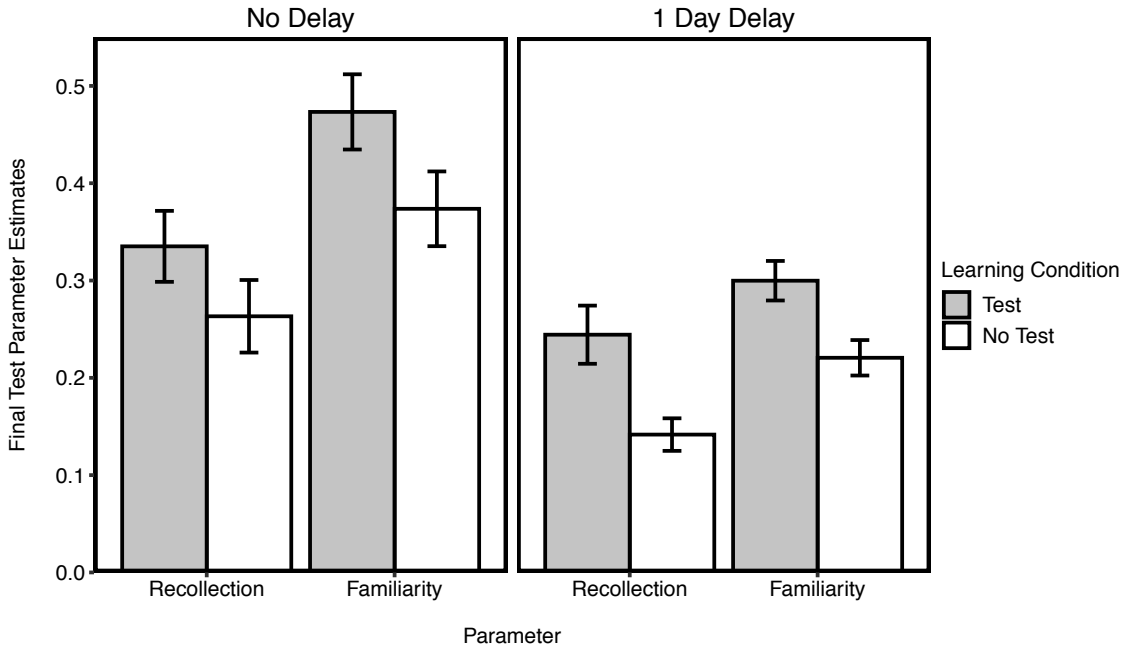

Figure 3.6 Recognition test parameter estimates: Older adults. Process estimates (mean ± SE) on the final recognition test are displayed for older adults by parameter (recollection, familiarity), initial learning

condition (test, no test), and delay (no delay, 1-day delay). A testing effect of similar magnitude occurred in estimates of recollection and familiarity.

A Three-Way Mixed ANOVA formally examined the effect of parameter (recollection, familiarity), initial learning condition (test, no test), and delay (no delay, 1-day delay) on process estimates during the final recognition test. The three-way interaction was not significant ($F(1, 62) = 1.71$, $p = .195$, $\eta_p^2 = .03$), nor were any of the two-way interactions (initial learning condition and parameter: $p = .913$; initial learning condition and delay: $p = .786$; parameter and delay: $p = .254$).

There was, however, a significant main effect of initial learning condition ($F(1, 62) = 85.47$, $p < .001$, $\eta_p^2 = .58$), revealing a testing effect in which estimates of recollection and familiarity were greater for items previously tested ($M = .34$) than for items previously studied only (untested; $M = .25$). The main effect of delay was also significant ($F(1, 62) = 16.52$, $p < .001$, $\eta_p^2 = .21$), such that estimates of recollection and familiarity decreased from the immediate ($M = .36$) to the 1-day delayed ($M = .23$) final test. Finally, the main effect of parameter was significant ($F(1, 62) = 14.88$, $p < .001$, $\eta_p^2 = .19$), such that, overall, estimates of familiarity ($M = .34$) were greater than estimates of recollection ($M = .25$).

In sum, a testing effect of similar magnitude was observed in estimates of both recollection and familiarity on both the immediate and 1-day delayed final tests.

3.3 Age Group Comparisons

The preceding analyses confirm a testing effect in older adults that is accompanied by both enhanced recollection and familiarity. In order to address the second aim of this master's thesis, the next step is to examine the extent to which age differences exist in the effects of testing on estimates of both recollection and familiarity.

To account for the potential effects of differences in initial test performance and years of education across age groups, additional analyses were conducted. Specifically, separate ANCOVAs controlling for overall proportion correct on the initial tests and reported years of education were conducted for the primary analyses in the Age Group Comparisons section (i.e. session 2 accuracy, parameter estimates in the no test condition, and the testing effect in parameter estimates). The specifics of any changes to patterns of results are described in the Supplementary Materials. Any major change related to a central analysis of interest is also noted in the main text. To preview these results, for most effects, the analyses produced the same patterns of results as in the original analysis.

3.3.1 Session 1

Prior analyses have already established that for both younger and older adults, performance on the initial cued-recall test was above floor and that within each age group performance in the no delay and 1-day delay conditions did not significantly differ prior to the delay manipulation. The following analysis tests for differences in initial test performance across the older and younger adult groups.

Figure 3.7 displays average proportion correct on the first and second initial test blocks for younger and older adults in the no delay and 1-day delay conditions, respectively (the data plotted in Figure 3.7 is the same as that in Figures 3.1 and 3.4 and is reproduced side-by-side below for easy comparison). As is apparent in the figure, initial cued-recall test performance was comparable for older and younger adult subjects across both delays.

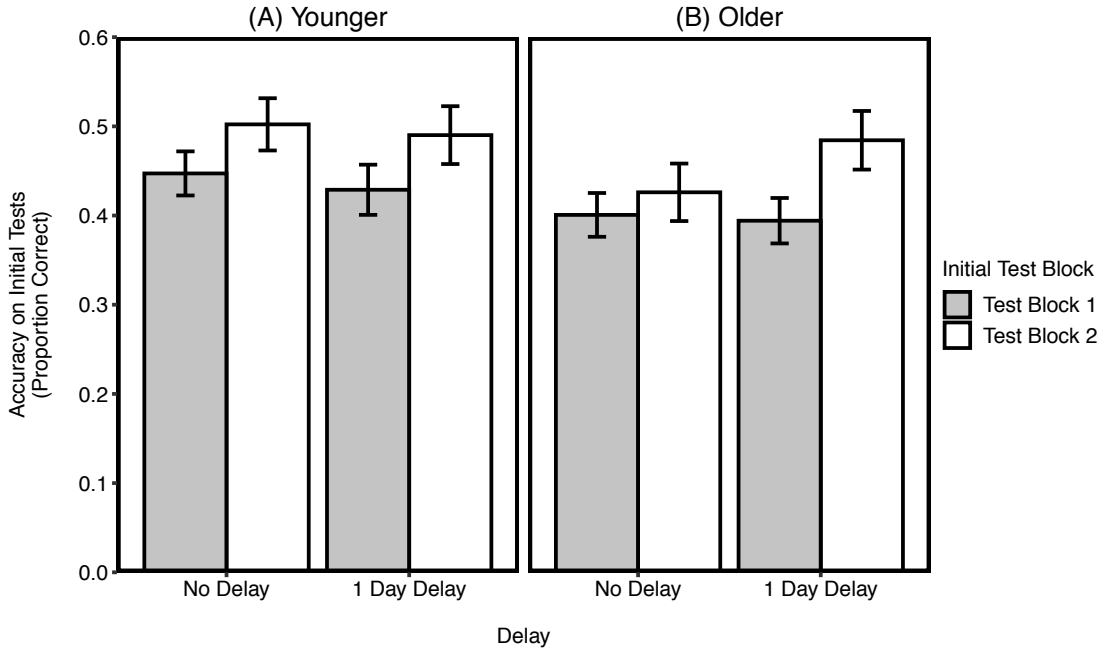

Figure 3.7 Initial cued-recall test accuracy: Younger and Older adults. Proportion correct on the initial cued-recall tests (mean ± SE) is displayed for younger (A) and older (B) adults by delay condition (no delay, 1-day delay) and test block (block 1, block 2). Initial test performance was comparable for older and younger adults across both delay conditions.

A Three-Way Mixed ANOVA formally examined the effect of age group (younger adults, older adults), delay (no delay, 1-day delay), and initial test block (block 1, block 2) on initial test performance. The main effect of age group was not significant ($F(1, 120) = 2.47$, $p = .119$, $\eta_p^2 = .02$), such that initial test performance did not significantly differ for younger ($M = .47$) and older ($M = .43$) adults. In addition, the main effect of delay was not significant ($F(1, 120) = 0.04$, $p = .837$, $\eta_p^2 = .000$), such that initial test performance did not significantly differ for the no delay ($M = .44$) and 1-day delay ($M = .45$) conditions. As expected, the main effect of initial test block, however, was significant ($F(1, 120) = 20.58$, $p < .001$, $\eta_p^2 = .15$), such that performance improved from the first ($M = .42$) to the second ($M = .48$) test block. Finally, none of the two- or three-way interactions were significant (age group X delay: $p = .430$; age group X

initial test block: p = .988; delay X initial test block: p = .167; age group X delay X initial test block: p = .254).

Given similar initial test performance across younger and older adults, any differences observed across age groups in the testing effect (in overall accuracy, recollection, or familiarity) can more readily be attributed to differences across age groups in the effects, themselves, of successful testing, rather than to differences in initial test performance.

3.3.2 Session 2

Accuracy. The magnitude of the testing effect in overall accuracy was next examined. Prior analyses have already established that both younger and older adults exhibit a testing effect in accuracy at both delays. Thus, the focus of the following analysis is to examine potential differences between older and younger adults in the magnitude of the testing effect. As before, accuracy was calculated both as hit rate minus false alarm rate and as *d'*. However, with one exception (described below), all patterns of results remained the same when using *d'*.

Figure 3.8 displays average accuracy on the final recognition test for younger and older adults by initial learning condition and delay (the data plotted in Figure 3.8 is reconfigured from prior figures below for easy comparison). As is evident in the figure, the magnitude of the testing effect in overall accuracy was roughly comparable for older and younger adult subjects across both delays (although there is a suggestion that the magnitude of the testing effect may be slightly larger for younger adults).

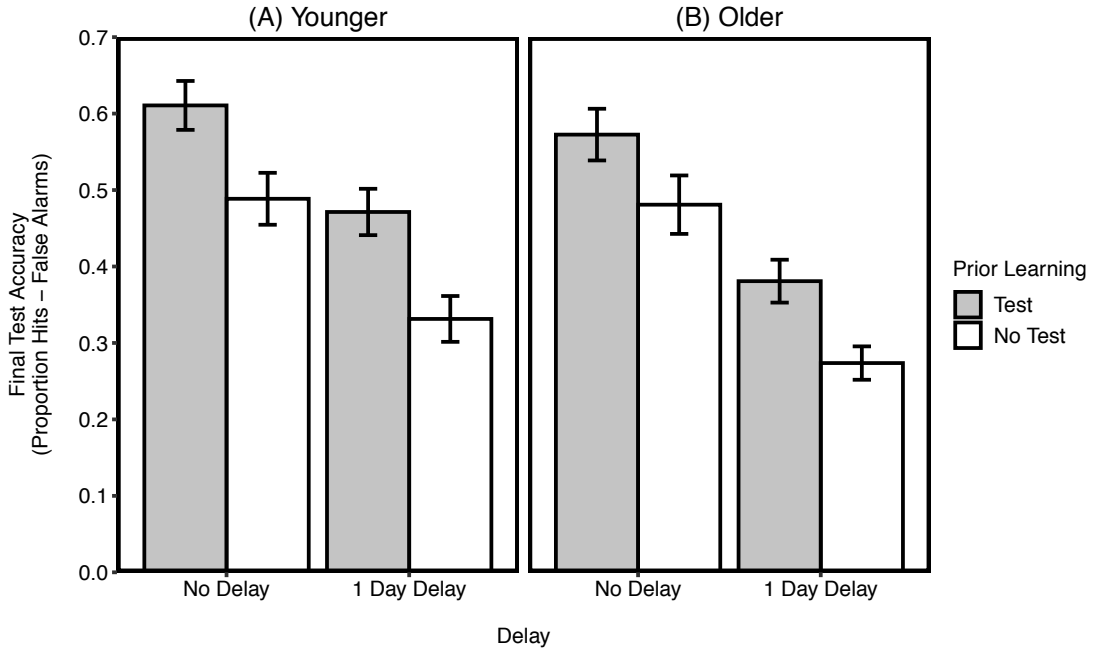

Figure 3.8 Recognition test accuracy: Younger and Older adults. Proportion hits minus false alarms (mean ± SE) on the final recognition test is displayed for younger (A) and older (B) adults by initial learning condition (test, no test) and delay (no delay, 1-day delay). The magnitude of the testing effect was roughly comparable for older and younger adults across both delays.

A Three-Way Mixed ANOVA formally examined the effect of age group (younger adults, older adults), initial learning condition (test, no test), and delay (no delay, 1-day delay) on final test accuracy. The primary effects of interest concerned the main effect of age group, as well as the interaction between age group and initial learning condition. Accordingly, these effects are reported first.

Specifically, the main effect of age group was not significant ($F(1, 120) = 2.60$, $p = .110$, $\eta_p^2 = .02$), such that overall final test accuracy did not significantly differ for older ($M = .43$) and younger ($M = .48$) adult age groups. The interaction between age group and initial learning condition was also not significant at $p < .05$ ($F(1, 120) = 3.21$, $p = .076$, $\eta_p^2 = .03$); however, a marginally significant effect ($p = .076$) suggested that younger adults may have improved more from prior testing than older adults. Follow-up analyses examined the effect of initial learning

condition collapsed across delay for younger and older adults, separately. Results revealed that although both younger and older adults exhibited a testing effect (younger: $F(1, 59) = 97.21$, $p < .001$, $\eta_p^2 = .62$, $M_{test} = .54$, $M_{notest} = .41$; older: $F(1, 63) = 73.89$, $p < .001$, $\eta_p^2 = .54$, $M_{test} = .48$, $M_{notest} = .38$), the effect was numerically greater for younger adults. The corresponding *d'* analysis produced the same patterns of results. However, when a single outlier was removed from the *d'* analysis (younger adult, no delay group), the above marginally significant interaction between age group and initial learning condition no longer approached significance ($p = .139$).

Although not of primary interest, the remaining results of the Three-Way Mixed ANOVA are reported below for the sake of completeness. As expected, the main effect of initial learning condition was significant ($F(1, 120) = 170.71$, $p < .001$, $\eta_p^2 = .59$), revealing a testing effect in which performance was greater for items previously tested ($M = .51$) than for items previously studied only (no test; $M = .39$). The main effect of delay was also significant ($F(1, 120) = 33.35$, $p < .001$, $\eta_p^2 = .22$), such that performance decreased from the immediate ($M = .54$) to the 1-day delayed ($M = .36$) final test. Finally, neither of the remaining two-way interactions were significant (age group X delay: $p = .397$; initial learning condition X delay: $p = .348$), nor was the three-way interaction between age group, initial learning condition, and delay ($F(1, 120) = 0.00$, $p = .948$, $\eta_p^2 = .00$).

Parameters.

No Test (Baseline) Condition. The preceding analysis established that older and younger adults exhibited similar overall accuracy on the final recognition test. In addition, as expected given prior work, the analysis confirmed that the magnitude of the testing effect was similar across age groups (although perhaps with the suggestion that the effect was larger in the younger adult group).

Before examining age effects in the magnitude of the testing effect in estimates of recollection and familiarity, it is important first to verify the existence of the expected patterns of results in terms of age effects in overall recollection and familiarity. Thus, the below analysis serves as a manipulation check to determine whether the older adult sample, relative to the younger adult sample, showed reduced recollection and relatively preserved familiarity, as prior literature would suggest.

To the extent that the benefits of testing operate by alleviating a recollection deficit or via relatively preserved familiarity, age effects in estimates of recollection and familiarity in the test condition may not be representative of prior work on the topic. Thus, to examine age effects in process estimates in which the influence of testing would not be a factor, parameter estimates in the no test condition were examined, specifically, rather than overall estimates of recollection and familiarity in the test and no test conditions combined.

Figure 3.9 displays average process estimates on the final recognition test in the no test (baseline) condition for younger and older adults by parameter and delay (Figure 3.9 is reconfigured below from prior figures for easy comparison). As can be seen in the figure, baseline estimates of familiarity were comparable for younger and older adults, as expected given prior literature on the topic. Also in line with prior work, estimates of recollection were reduced for older relative to younger adults (although the age effect in recollection was only marginally significant, $p = .052$).

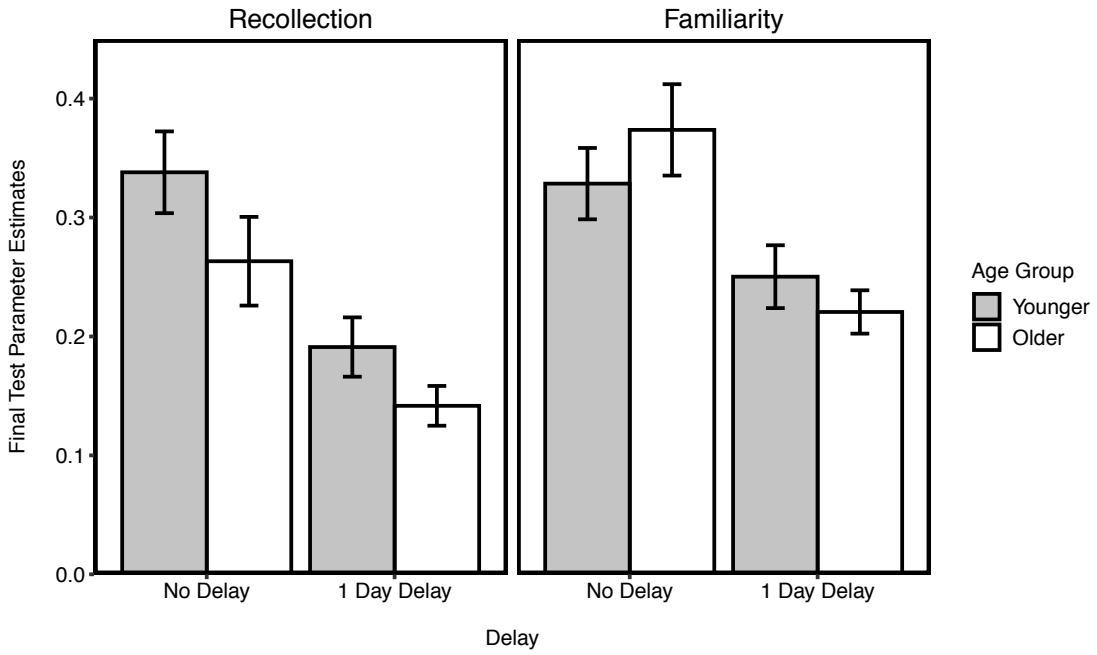

Figure 3.9 Parameter estimates in the no test condition: Younger and Older adults. Process estimates (mean ± SE) on the final recognition test in the no test condition are displayed for younger and older adults by parameter (recollection, familiarity) and delay (no delay, 1-day delay). Baseline estimates of familiarity were comparable for older and younger adults, whereas baseline estimates of recollection were numerically reduced.

A Three-Way Mixed ANOVA formally examined the effect of age group (younger adults, older adults), parameter (recollection, familiarity), and delay (no delay, 1-day delay) on process estimates in the no-test, baseline condition. The three-way interaction was not significant ($F(1, 120) = 2.10$, $p = .149$, $\eta_p^2 = .02$), nor were the two-way interactions between age group and delay ($p = .603$) or between parameter and delay ($p = .591$). The primary analysis of interest was the two-way interaction between age group and parameter, in order to explore the extent to which recollection and familiarity show the expected patterns of results across age groups. As expected, the two-way interaction between age group and parameter was significant ($F(1, 120) = 4.09$, $p = .045$, $\eta_p^2 = .03$), revealing that the effect of age group differed for estimates of recollection and familiarity at baseline. To explore this interaction further, the effect of age group was examined separately in estimates of recollection and familiarity, collapsed across delay. For

estimates of familiarity in the no test condition, there was no significant or marginally significant effect of age group ($F(1, 122) = 0.06$, $p = .803$, $\eta_p^2 = .00$), such that estimates of familiarity were similar for younger ($M = .29$) and older ($M = .30$) adults. By contrast, for estimates of recollection in the no test condition, the effect of age group was marginally significant ($F(1, 122) = 3.85$, $p = .052$, $\eta_p^2 = .03$), suggesting that older adults displayed reduced recollection ($M = .20$) relative to younger adults ($M = .26$).

Notably, in the supplementary exclusion analyses (i.e., excluding subjects who struggled with the Remember-Know instructions, and excluding subjects based on initial test guessing behavior), the interaction between age group and parameter was no longer significant at $p < .05$, but only approached significance in both cases (Remember-Know exclusion: $p = .060$; Initial Test Guessing exclusion: $p = .080$). An analysis of familiarity in the no test condition revealed the same pattern of results as above, such that estimates of familiarity were similar for younger and older adults. In contrast to the above analysis, however, the effect of age group in estimates of recollection in the no test condition was no longer marginally significant in either case (Remember-Know exclusion: $p = .113$; Initial Test Guessing exclusion: $p = .138$). However, as in the main text analysis, in both supplementary analyses older adults displayed numerically reduced recollection ($M = .21$) relative to younger adults ($M = .27$).

Results from the two additional supplementary covariate analyses (i.e., covarying years of education, and covarying session 1 performance) are also worth mentioning here. Specifically, when years of education was included as a covariate in the above analyses, the effect of age group in estimates of no test recollection was significant ($p = .003$), such that older adults exhibited reduced estimates of recollection relative to younger adults, as expected. However,

when session 1 performance was instead included as a covariate, the effect of age group in recollection was no longer significant ($p = .148$).

Thus, estimates of familiarity followed the expected pattern of results in all cases, revealing comparable familiarity estimates in older and younger adults. By contrast, analyses pertaining to estimates of recollection revealed a more diverse set of outcomes. The expected reduction in recollection in older adults was not always observed statistically; however, the effect was observed numerically in the group means. Thus, although certainly not a robust effect, analyses suggest that older adults showed some expected reductions in recollection in the no test condition.

To further explore the significant age group by parameter interaction from the main analysis above, the effect of parameter was examined separately in the younger and older adult age groups, collapsed across delay. Whereas in the younger adult age group, estimates of recollection and familiarity in the no test condition did not differ significantly in magnitude ($F(1, 59) = 1.18$, $p = .281$, $\eta_p^2 = .02$), for older adults, estimates of recollection were significantly lower than estimates of familiarity ($F(1, 63) = 13.48$, $p < .001$, $\eta_p^2 = .18$).

Finally, in the full Three-Way Mixed ANOVA above, the main effect of delay was significant ($F(1, 120) = 27.68$, $p < .001$, $\eta_p^2 = .19$), indicating that estimates of recollection and familiarity in the no test condition decreased over the delay.

Thus, the significant interaction between age group and parameter indicated that the effect of age group depends on the parameter of interest (recollection, familiarity). As expected, relative to younger adults, older adults exhibited similar levels of familiarity and somewhat reduced levels of recollection on the final recognition test. Given the observance broadly of the

expected patterns of results at baseline, analysis could then proceed to examining age effects in the magnitude of the testing effect in recollection and familiarity.

The Testing Effect. First, age effects in the magnitude of the testing effect in recollection were examined. Figure 3.10 displays average estimates of recollection on the final recognition test for younger and older adults by initial learning condition and delay (Figure 3.10 is reconfigured below from prior figures for easy comparison). As can be seen in the figure, although both older and younger adults exhibited a testing effect in recollection, the magnitude of the testing effect in recollection was reduced for older, relative to younger, adults at both delays.

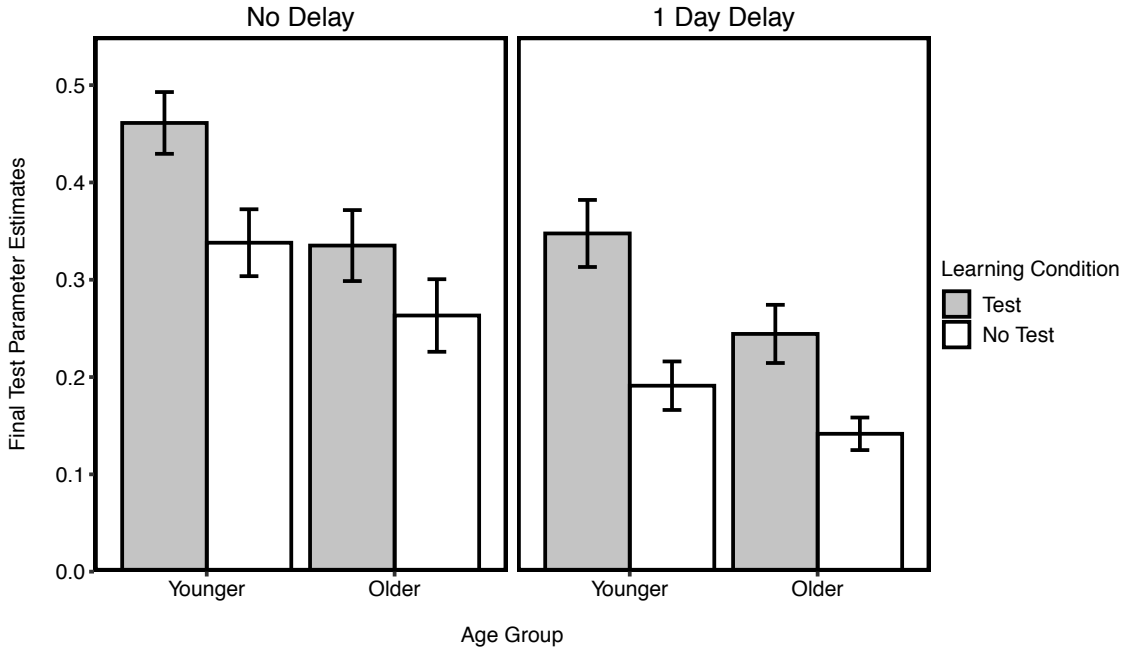

Figure 3.10 Estimates of recollection on the final test: Younger and Older adults. Estimates of recollection (mean ± SE) on the final recognition test are displayed for younger and older adults by initial learning condition (test, no test) and delay (no delay, 1-day delay). The magnitude of the testing effect in recollection was reduced for older adults relative to younger adults.

A Three-Way Mixed ANOVA formally examined the effect of age group (younger adults, older adults), initial learning condition (test, no test), and delay (no delay, 1-day delay) on

estimates of recollection on the final test. The three-way interaction was not significant ($F(1, 120) = 0.00$, $p = .945$, $\eta_p^2 = .00$). The primary analysis of interest was the two-way interaction between age group and initial learning condition, which was significant ($F(1, 120) = 8.07$, $p = .005$, $\eta_p^2 = .06$), such that the magnitude of the testing effect in recollection was greater for younger ($M_{test - no\ test} = .14$) than for older ($M_{test - no\ test} = .09$) adults.

Follow-up analyses examined the effect of initial learning condition in younger and older adults separately, collapsed across delay, finding a significant effect in both age groups, but a larger effect in younger than in older adults (Younger: $F(1, 59) = 102.99$, $p < .001$, $\eta_p^2 = .64$; Older: $F(1, 63) = 48.66$, $p < .001$, $\eta_p^2 = .44$). Follow-up analyses further examined the effect of age group for items in the test and no test conditions, separately, collapsed across delay. For the no test condition, as explored in the previous section, the effect of age group on estimates of recollection was marginally significant ($F(1, 122) = 3.85$, $p = .052$, $\eta_p^2 = .03$), such that older adults ($M = .20$) had numerically lower estimates of recollection than younger adults ($M = .26$). For the test condition, however, the effect of age group was significant and larger in magnitude ($F(1, 122) = 11.18$, $p = .001$, $\eta_p^2 = .08$). Thus, the gap between estimates of recollection for older ($M = .29$) and younger ($M = .40$) adults for items previously tested was larger than that for items not previously tested. In other words, retrieval practice increased the differences between older and younger adults in estimates of recollection.

For the sake of completeness, the remaining results of the Three-Way Mixed ANOVA are reported below. Specifically, neither of the remaining two-way interactions were significant (age group X delay, $p = .689$; initial learning condition X delay, $p = .085$). However, the latter effect (initial learning condition X delay) was marginally significant, indicating a potential

increase in the testing effect in estimates of recollection from the immediate to the 1-day delayed final test. Finally, the main effect of delay was significant ($p < .001$).

Age effects were next examined in the magnitude of the testing effect in familiarity. Figure 3.11 displays average estimates of familiarity on the final recognition test for younger and older adults by initial learning condition and delay (Figure 3.11 is reconfigured below from prior figures for easy comparison). As is apparent in the figure, the magnitude of the testing effect in familiarity was similar for older and younger adults at both delays.

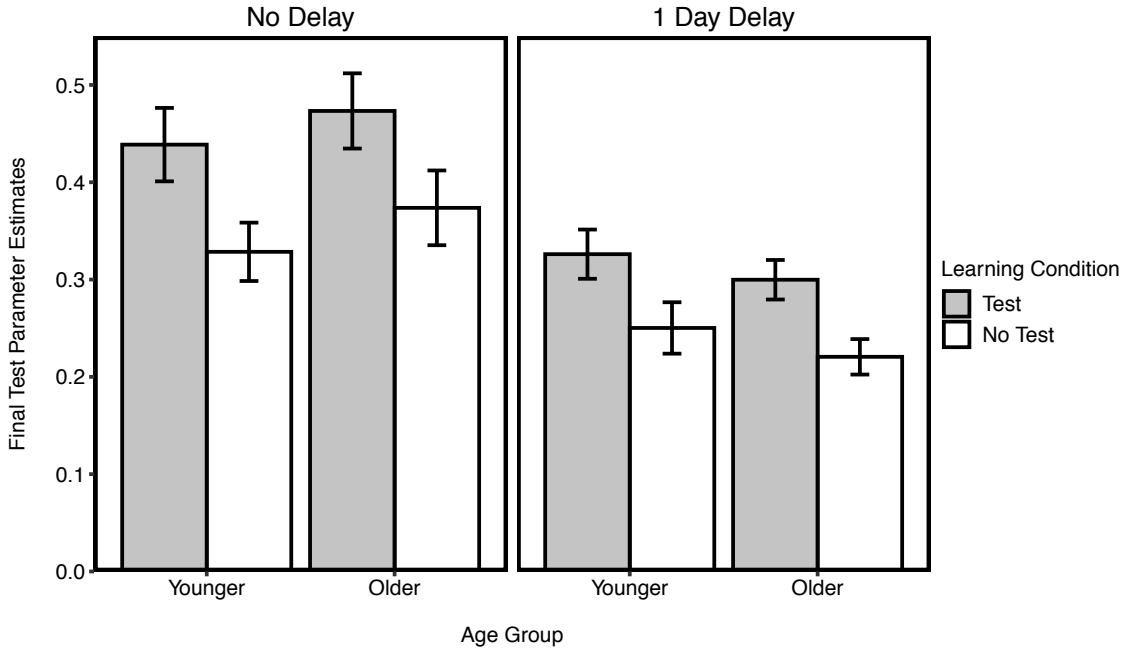

Figure 3.11 Estimates of familiarity on the final test: Younger and Older adults. Estimates of familiarity (mean ± SE) on the final recognition test are displayed for younger and older adults by initial learning condition (test, no test) and delay (no delay, 1-day delay). The magnitude of the testing effect in familiarity was comparable for older and younger adults.

A Three-Way Mixed ANOVA formally examined the effect of age group (younger adults, older adults), initial learning condition (test, no test), and delay (no delay, 1-day delay) on estimates of familiarity on the final test. The three-way interaction was not significant ($F(1, 120) = 0.10$, $p = .751$, $\eta_p^2 = .00$). Of primary interest is the two-way interaction between age group and

initial learning condition, which was not significant ($F(1, 120) = 0.03$, $p = .870$, $\eta_p^2 = .00$), indicating that the magnitude of the testing effect in familiarity was similar for younger ($M_{test - no\ test} = .09$) and older ($M_{test - no\ test} = .09$) adults.

As before, for the sake of completeness, the remaining results of the Three-Way Mixed ANOVA are reported below. Specifically, neither of the remaining two-way interactions were significant (age group X delay, $p = .234$; initial learning condition X delay, $p = .214$). In addition, the main effects of delay and of initial learning condition were both significant (both $ps < .001$). However, the main effect of age group was not significant ($p = .834$).

For an analysis and discussion of the effects of initial learning condition, age group, and delay on raw Know accuracy, rather than parameter estimates of familiarity, see the Supplementary Materials (along with Supplementary Figure S4).

Thus, whereas in estimates of recollection older adults showed a reduced testing effect relative to younger adults, for estimates of familiarity older and younger adults showed a testing effect of similar magnitude. This result may seem counterintuitive when considered together with the finding that older and younger adults showed a testing effect of similar magnitude in overall accuracy. However, estimates of recollection and familiarity are made to be independent of one another using the Independence Remember-Know procedure (Yonelinas, 2002; Yonelinas & Jacoby, 1995) and are not assumed to be mutually exclusive (i.e. the assumption made if raw Know responding is used as an estimate of familiarity). Thus, the magnitude of the testing effect in overall accuracy is not the direct sum of the magnitude of the testing effect in recollection and familiarity estimates (as it would be if using raw Remember and Know responses only).

Finally, when directly compared in a Four-Way Mixed ANOVA (age group X initial learning condition X parameter X delay), the three-way interaction between age group, initial

learning condition, and parameter was marginally significant, but was not significant at conventional thresholds ($F(1, 120) = 2.94$, $p = .089$, $\eta_p^2 = .02$). Although not the primary analysis of interest, for the sake of completeness, results of the entire Four-Way Mixed ANOVA are provided in the Supplementary Materials.

3.3.3 Manipulation Checks

Prior work (e.g., Kornell, Bjork, & Garcia, 2011) would suggest that the overall benefit of testing derives from a boost in performance for items answered correctly on the initial tests, but not for items missed on the initial tests (in the absence of feedback). Thus, there should be little to no difference in final test performance for items previously tested and missed and for items previously studied only (no test condition). As a manipulation check, performance on the final recognition test in terms of overall hit rates and estimates of recollection and familiarity was compared for items previously missed on the initial test and items previously studied only.

The complete set of analyses can be found in the Supplementary Materials (along with Supplementary Figures S5-S7). Briefly, as expected for both older and younger adults, final test hit rates were comparable for items missed on the initial test and items not tested. Further, estimates of recollection for items missed on the initial test were not greater than for items not tested. In fact, lower estimates of recollection were observed for items missed on the initial test. Finally, estimates of familiarity on the 1-day delayed final test were comparable for items missed on the initial test and items not tested. By contrast, on the immediate final test, estimates of familiarity were greater for items previously missed than for items not tested. Critically, however, for both older and younger adults, estimates of familiarity on the immediate final test were significantly higher for items answered *correctly* on the initial test than for both items missed and items that were not tested. Thus, findings conformed to expected patterns of results.

In addition, response times (RTs) on the final recognition test were examined to ensure that the expected patterns of results were exhibited regarding the effect of age group (younger adults < older adults), response type (Remember < Know), and initial learning condition (test < no test). In short, all expected patterns were observed. The Supplementary Materials provides a complete mixed effects RT analysis and discussion (along with Supplementary Figures S8-S10).

3.4 The Magnitude of the Testing Effect in Familiarity

The third aim of this master's thesis is to examine the extent to which the testing effect in familiarity is more readily revealed when recollection is reduced and must be relied on to a lesser extent for success on the final test. In order to test this prediction, prior analyses examined the mechanisms that underlie the testing effect in a population shown to have reduced recollection with relatively preserved familiarity (the older adult sample in this case). This preceding analysis revealed that the magnitude of the testing effect in familiarity did not differ for older and younger adults. Older adults showed a reduced testing effect in recollection relative to younger adults, but a comparable (and no greater) testing effect in familiarity. Thus, prior analyses suggest that the testing effect in familiarity is not more readily revealed when recollection is reduced and must be relied on to a lesser extent for success on the final test.

In order to examine more directly the question posed in the third aim, however, rather than use age group as a proxy for the presence or absence of a recollection deficit, the following multiple regression analyses directly examine the relation between estimates of overall recollection and the magnitude of the testing effect in familiarity. In the analyses below, continuous predictor variables were grand mean centered to improve interpretation of model coefficients.

The scatterplot in Figure 3.12 depicts the relation between estimates of overall recollection and the magnitude of the testing effect in familiarity. As is evident in the figure, the relation was not significant.

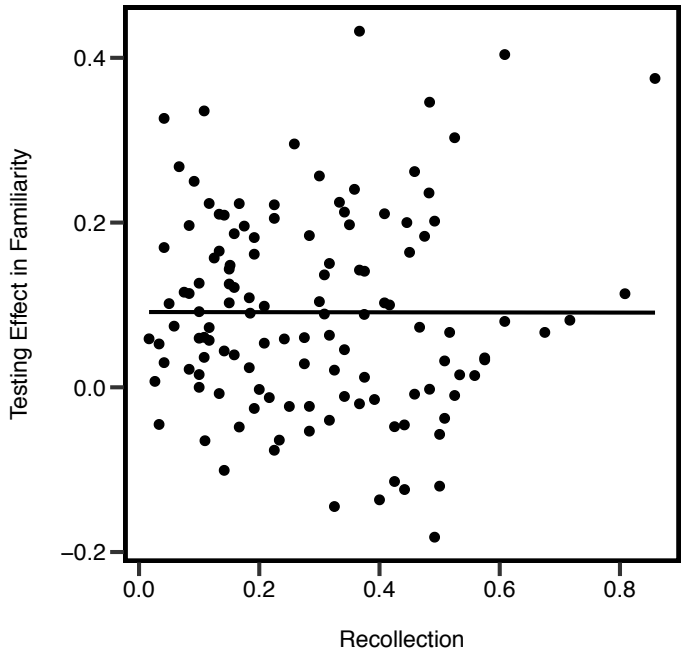

Figure 3.12 Scatterplot: Recollection and the testing effect in familiarity. A scatterplot depicts the relation between estimates of overall recollection and the magnitude of the testing effect in familiarity. No significant relation was observed. See Supplementary Materials Figure S11 for scatterplots of this relation broken down by age group and delay.

Multiple regression analysis formally examined the relation between estimates of recollection and the magnitude of the testing effect in estimates of familiarity. Specifically, recollection estimates did not explain significant variance in the magnitude of the testing effect in familiarity across subjects ($F(1,122) = 0.00$, $p = .989$, $R^2 = .00$; all model outputs can be found in Supplementary Tables S1-S6). Including delay condition (the main effect and interaction with recollection) in the model did not significantly improve the fit above and beyond the recollection-only model (model comparison: $F(2,120) = 1.67$, $p = .193$). Similarly, including age as a continuous variable in the model did not significantly improve fit above the recollection-

only model (model comparison: $F(2,120) = 0.17$, $p = .843$). Further, the same patterns of results occurred when predicting the magnitude of the testing effect in familiarity from the magnitude of the testing effect in recollection (rather than from overall recollection).

For comparison, multiple regression analysis examined the relation between raw Remember hit rates and the magnitude of the testing effect in raw Know hit rates. As is apparent in the scatterplot in Figure 3.13, as estimates of overall Remember hit rates increased, the magnitude of the testing effect in Know hit rates decreased.

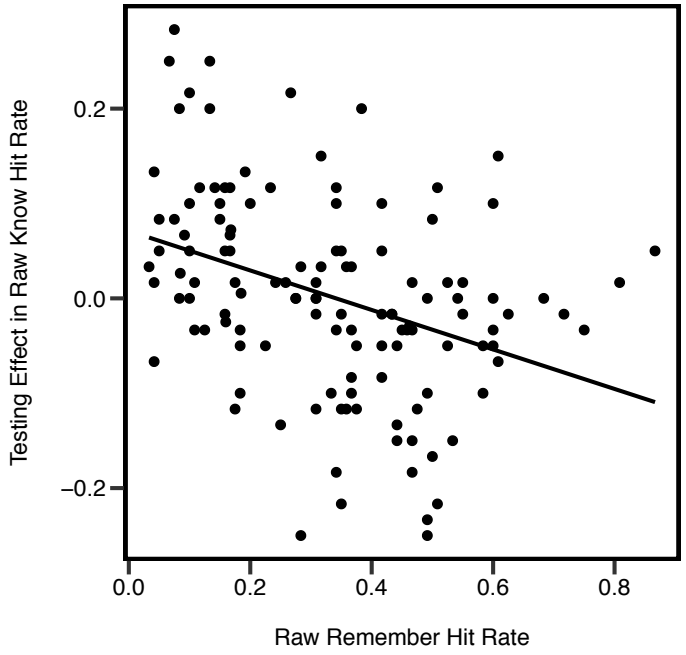

Figure 3.13 Scatterplot: Remember hit rates and the testing effect in Know hit rates. A scatterplot depicts the relation between estimates of raw Remember hit rates and the magnitude of the testing effect in raw Know hit rates. A negative relation was observed. See Supplementary Materials Figure S12 for scatterplots of this relation broken down by age group and delay.

Multiple regression analysis formally examined the relation between raw Remember hit rates and the magnitude of the testing effect in raw Know hit rates. In contrast to the above analysis, Remember hit rates explained significant variance in the magnitude of the testing effect in Know responses across subjects ($F(1,122) = 19.09$, $p < .001$, $R^2 = .14$). Higher Remember hit

rates predicted reductions in the magnitude of the testing effect in Know responses ($\beta_{Prop.\ Remember\ hit}$ = -0.37). Including age as a continuous variable in the model (the main effect and interaction with Remember hit rate) did not significantly improve model fit (model comparison: $F(2,120) = 0.82$, $p = .444$). Including delay in the model, however, significantly improved the fit of the model relative to the Remember hit rate only model (model comparison: $F(2,120) = 4.97$, $p = .008$; delay X Remember hit rate interaction: $t(120) = -2.49$, $p = .014$, $\beta = -.42$). On both the immediate and delayed final tests, the relation between Remember hit rates and the testing effect in Know hits was significantly negative; however, beta coefficients of simple slopes revealed a larger negative relation at the 1-day delay ($t(120) = -4.99$, $p < .001$, $\beta = -.66$) than on the immediate final test ($t(120) = -2.13$, $p = .035$, $\beta = -.23$). The same patterns of results occurred when predicting the magnitude of the testing effect in Know hit rates from the magnitude of the testing effect in Remember hit rates (rather than overall Remember hit rates), except that adding delay to the model did not significantly improve the fit of the model ($F(2,120) = 0.12$, $p = .887$).

Finally, a potential quadratic trend in the relation between recollection (or Remember hit rates) and the testing effect in familiarity (or the testing effect in Know hit rates) was explored. First, multiple regression analysis examined whether the inclusion of a quadratic term for overall recollection improved the fit of the model predicting the testing effect in familiarity. The quadratic term was marginally significant ($t(121) = 1.87$, $p = .064$, $\beta = .14$); however, when a single outlier (> 3 SDs from the mean in recollection) was removed, the term was no longer marginally significant ($t(120) = 0.65$, $p = .518$, $\beta = .05$).

By contrast, for raw Remember hit rates predicting the magnitude of the testing effect in raw Know hit rates, the quadratic term was significant ($t(121) = 3.58$, $p = .001$, $\beta = .25$; see Supplementary Figure S13). However, at grand mean levels of Remember hit rates, the relation

between raw Remember hit rates and the testing effect in raw Know hit rates remained significantly negative ($t(121) = -5.40$, $p < .001$, $\beta = -.45$). Visual examination of the quadratic curve in Supplementary Figure S13 reveals that subjects with relatively higher overall Remember hit rates had an increased testing effect in raw Know hit rates. When predicting the magnitude of the testing effect in raw Know hit rates from the magnitude of the *testing effect* in raw Remember hit rates (rather than from *overall* raw Remember hit rates), however, this quadratic relation no longer appeared ($t(121) = -0.89$, $p = .376$, $\beta = -.05$).

Thus, with respect to the third aim of this study, the present analyses aligned with the prior analyses, which failed to reveal significant differences between older and younger adults in the magnitude of the testing effect in familiarity. In the present analysis, overall recollection (and the magnitude of the testing effect in recollection) did not significantly predict the magnitude of the testing effect in familiarity.

By contrast, in the present analysis overall raw Remember hit rates (and the magnitude of the testing effect in raw Remember hit rates) predicted the magnitude of the testing effect in raw Know hit rates. The relation was significantly negative, such that with a smaller testing effect in Remember hit rates, a larger testing effect in raw Know hit rates was observed. This relation is to be expected, given that if a subject has recollection for an old item, they must supply a Remember (and not a Know) response, even if they feel familiarity with the item, as well. Thus, with a greater testing effect in Remember responses, there is less potential to observe a testing effect in raw Know responses. This underscores the need for the Independence Remember-Know procedure (Yonelinas, 2002; Yonelinas & Jacoby, 1995) used in prior work and in the present study to obtain parameter estimates of recollection and familiarity. A significant quadratic relation was also observed between overall Remember hit rates and the testing effect in Know hit

rates; however, at average levels of Remember hit rates, the relation between overall Remember hit rates and the magnitude of the testing effect in Know hit rates remained negative.

3.4.1 Reliability Analysis

Finally, qualifying the above results is an examination of the reliability of estimates of the testing effect in recollection and familiarity. Split-half reliabilities using the Spearman-Brown Prophecy formula (Brown, 1910; Spearman, 1910) were obtained for the magnitude of the testing effect in familiarity, recollection, raw Know hit rates, and raw Remember hit rates. Reliability estimates of the magnitude of the testing effect in recollection (.52), raw Know hit rates (.48), and raw Remember hit rates (.52) were moderate (reliability of the testing effect in recollection and in raw Remember hit rates were the same, given the recollection calculation). However, the reliability estimate for the magnitude of the testing effect in familiarity was extremely low (.07).

By contrast, and as expected, reliability estimates for recollection, familiarity, raw Remember hit rates, and raw Know hit rates in the test and no test conditions separately (i.e., not as a testing effect difference score), were a great deal higher in all cases ($recollection_{test}$ = .92; $recollection_{notest}$ = .91; $familiarity_{test}$ = .79; $familiarity_{notest}$ = .80; raw $Remember_{test}$ = .93; raw $Remember_{notest}$ = .92; raw $Know_{test}$ = .87; raw $Know_{notest}$ = .86).

Together, these results suggest that although the reliability of difference scores are often low, the reliabilities of the difference scores representing the testing effect in recollection, raw Remember hit rates, and raw Know hit rates were moderate. Critically, however, the reliability of the difference score representing the testing effect in familiarity was extremely low. This suggests that although testing may increase familiarity broadly, the magnitude of this increase in familiarity does not exhibit reliable variability between subjects. Given these reliabilities, with

respect to Aim 3, it is not surprising that the magnitude of the testing effect in familiarity was not significantly predicted by overall recollection or the testing effect in recollection.

Chapter 4: Discussion

This project examined the mechanisms from the dual-process perspective that underlie the benefits of retrieval practice in both younger and older adults. In addressing the three aims of the current study, I will begin with a discussion of the present findings within the context of the mechanisms that underlie the testing effect broadly. I will then move on to discuss the findings as they concern the mechanisms that underlie the testing effect in older adults specifically. Finally, implications for theories of the testing effect and for aging populations, as well as limitations and future directions, will be discussed.

4.1 The Testing Effect from the Dual-Process Perspective

Some prior studies that directly examine recollection and familiarity following initial testing find a testing effect in recollection only (Chan & McDermott, 2007; Pu & Tse, 2014; Verkoeijen et al., 2011, also see Jones & Roediger, 1995). However, more recent work has suggested that changes in familiarity may be involved in the testing effect as well (Bies-Hernandez, 2013; Gao et al., 2016; Jia et al., 2019; Shaffer & McDermott, 2020; even if such changes are small in magnitude, e.g., Guran et al., 2020).

With respect to the first aim of this project, results from the in-lab undergraduate sample reveal a testing effect in estimates of both recollection and familiarity, replicating prior work from several online samples via Amazon Mechanical Turk (Shaffer & McDermott, 2020). This suggests that the prior findings were not a result of the idiosyncrasies of an online sample and underscores the replicability of the results in a more controlled testing environment and within a high-performing, undergraduate sample. Thus, this project adds to the growing literature suggesting that the testing effect can be supported by changes in both recollection and familiarity.

What might account for the discrepancies observed across the literature? The present study tested the prediction that the magnitude of the observed testing effect in familiarity would increase as subjects are able to rely less on high levels of recollection for success on the final recognition test. This prediction was explored by testing a sample of subjects—older adults—known to have reductions in recollection with relatively preserved familiarity. Although estimates of recollection on the final test were somewhat reduced in the older adult group, older adults did not exhibit a testing effect in familiarity larger than that of younger adults. Further, direct examination across subjects of the relation between overall recollection and the testing effect in familiarity revealed no effect. However, an examination of reliabilities suggested that, although familiarity estimates in the test and no test conditions separately show moderate-to-high reliability and prior testing led to group-level increases in familiarity, the magnitude of this increase was not reliable.

Together, with respect to the third aim of this project, these results suggest two things. First, they suggest that the magnitude of the testing effect in familiarity does not relate to overall individual levels of recollection. Second, they suggest that although familiarity estimates may be reliable and relate to other individual difference factors, the magnitude of the *testing effect* in familiarity may not be a reliable measure. Thus, it is likely not a fruitful avenue for future work to seek out individual difference factors that predict the magnitude of the testing effect in familiarity specifically.

4.1.1 Final Test Type

If not overall levels of recollection inherent to the subject, what else might explain the discrepancies observed in the literature with respect to the contribution of familiarity to the testing effect? One possibility, first discussed in Bies-Hernandez (2013), is that the discrepancies

in the literature may result from the use of different final testing formats. The suggestion is that different final test formats induce different levels of reliance on the use of recollection or familiarity for success during final testing, which, in turn, may influence whether or not a testing effect appears in either process.

Indeed, Bies-Hernandez (2013) noted that two prior studies (Chan & McDermott, 2007, and Verkoeijen et al., 2011) use primarily source or exclusion final testing and observe testing effects in recollection only. In source or exclusion final tests, the subject must accurately recall a particular feature of the initial learning or testing context (e.g., a temporal feature, such as the list in which an item was studied) in order to supply a correct recognition response to an old item. These forms of final tests emphasize the use of recollection for success. By contrast, Bies-Hernandez (2013) used confidence-based final testing and primarily observed testing effects in familiarity. In a confidence-based final test, subjects may provide an accurate high-confidence recognition response to an old item (thought to index recollection) or a range of lower-confidence responses (thought to index a signal detection familiarity process; see Parks & Yonelinas, 2007; Yonelinas, 2001b). Thus, there is no emphasis on the use of recollection for success on the final test.

An examination of the current state of the literature provides continued support for this possibility. Indeed, the current study, along with other studies that observe a testing effect in familiarity (Bies-Hernandez, 2013; Gao et al., 2016; Guran et al., 2020; Jia et al., 2019; Shaffer & McDermott, 2020), used either Remember-Know or confidence-based responding on the final recognition test. In the Remember-Know procedure, as in confidence-based final testing, subjects may provide an accurate recognition response to an old item with or without accompanying recollection (via a Remember or Know response, respectively). Thus, there is no emphasis on the

use of recollection for success on the final test. By contrast, studies in which a testing effect is not observed in familiarity have primarily—although not always—used source or exclusion final tests (e.g., Pu & Tse, 2014; Verkoeijen et al., 2011; Chan & McDermott, 2007, Exp. 1a,1b, 3; although see Chan & McDermott, 2007, Exp. 2, and Jones & Roediger, 1995).

Thus, the possibility remains that one factor leading to the discrepancies observed in the literature is an increased or decreased emphasis on or attention to recollective details when responding on different final tests. However, as noted in prior work (Bies-Hernandez, 2013), this suggestion is offered tentatively, as varied estimation procedures (Remember-Know, confidence, process-dissociation) often show convergence (e.g., Koen & Yonelinas, 2016; Yonelinas, 2001b; Chan & McDermott, 2007). Further work will be necessary in order to test this possibility.

4.2 The Testing Effect and Aging

The second aim of this project is to examine the mechanisms that underlie the testing effect in older adults, specifically. Does retrieval practice for older adults ameliorate existing deficits in recollection or does it function primarily via enhanced familiarity?

The present study replicates prior work in observing a testing effect in overall recognition accuracy in the older adult group similar in magnitude to that observed in the younger adult group (e.g., Coane, 2013; Meyer & Logan, 2013; Bishara & Jacoby, 2008; Kausler & Phillips, 1988; Kausler & Wiley, 1991; Logan & Balota, 2008; Rabinowitz & Craik, 1986; although see Henkel, 2014 and Tse et al., 2010 for mixed results; Guran et al., 2019, 2020). When parameter estimates of recollection and familiarity were examined on the final test, the older adult group, like the younger adult group, exhibited a testing effect in estimates of both recollection and familiarity on both immediate and delayed final tests. However, the testing effect in recollection was smaller in magnitude in older relative to younger adults, such that prior testing actually

increased age differences in estimates of recollection. By contrast, the testing effect in familiarity was similar across the two age groups. Critically, these differences in final testing were likely not due entirely to differences in overall initial test performance. Indeed, older and younger adults performed comparably during initial testing; and, after controlling for individual level initial test performance, the relevant patterns of results on the final test remained the same. Finally, the above patterns of results were consistent across both immediate and 1-day delayed final tests, suggesting that the findings are robust to the retention interval between initial and final testing. Thus, the answer to the question of whether retrieval practice in older adults functions by improving recollection or via preserved familiarity appears to be: some of both. However, given that age-related differences in recollection actually increased as a result of prior testing, the notion that retrieval practice can be considered to *alleviate* deficits in recollection in older adults is dubious.

These results align in part with prior work from the spaced retrieval literature. Following spaced retrieval, Bishara and Jacoby (2008) observed commensurate increases in recollection and familiarity in younger adults. By contrast, the authors observed in older adults greater increases in familiarity than in recollection, reflected by higher levels of intrusions on an exclusion final test for items previously tested than for items untested or restudied. However, unlike in a standard testing effect study, in a spaced retrieval paradigm items are studied and then tested after short lags of, for example, 1-6 items. By contrast, in the present study items were studied and then tested in separate blocks (i.e., after much longer lags), likely requiring more effortful or controlled retrieval of items during initial testing (Pyc & Rawson, 2007, 2009; Roediger & Karpicke, 2010). Such a situation may serve to increase broadly the resulting testing effect in recollection in both younger and older adults, which could produce the pattern of results

observed in the present study. Together, these findings underscore the importance of the choice of initial test type in determining the processes observed to be enhanced in final testing.

To our knowledge, only one prior study (Guran et al., 2020) has directly compared the testing effect in older and younger adults in recollection and familiarity using a testing effect paradigm in which initial studying and testing occur in separate blocks. Although Guran et al. (2020) conclude that the testing effect in both older and younger adults accrues primarily via increases in recollection (indexed via Remember responses in the Remember-Know paradigm), the authors also observe a small testing effect in Know responses. On the surface, these results seem in part to conflict with the present findings. However, had Guran et al. (2020) calculated estimates of familiarity (rather than analyzing raw Know responses only), the same conclusions as in the present study would likely have been reached with respect to the testing effect in familiarity. This partial convergence of results occurred despite several methodological differences between the two studies.

Thus, the present findings, in the context of the prior literature, suggest that testing can improve both recollection and familiarity in older adults, depending perhaps in part on the form of the initial test. However, the improvements in recollection may decline to some extent with age, while improvements in familiarity appear to persist. Implications of the present findings for theories of the testing effect and for aging populations will be discussed next.

4.3 Implications for Theories of the Testing Effect

What do the present results mean for theories of the testing effect? Although there is no agreed upon or completely satisfying theory of the mechanisms that underlie the testing effect, two prominent theories of the testing effect—the Episodic Context Account (Karpicke et al., 2014; Lehman et al., 2014; Whiffen & Karpicke, 2017) and the Elaborative Retrieval Hypothesis

(Carpenter, 2009; Carpenter & DeLosh, 2006; Rawson et al., 2015)—suggest that the mechanisms that underlie the benefits of retrieval relate to temporal context updating or associative memory mechanisms. In either case, theoretical discussion of the testing effect often refers to the effect as a recollection-specific phenomenon. The current results align with prior work, as well as with the preeminent theories of the testing effect, in supporting the idea that testing improves recollection-related processing—whether via a temporal context updating or elaboration that leads to a greater number of associative cues during retrieval or via another recollection-related process. However, the present results also replicate and extend prior work to reveal a testing effect in familiarity with in-lab undergraduate samples, as well as in multiple age groups and across multiple delays. In light of these results, the notion that the testing effect functions via recollection to the exclusion of familiarity is not supported. This suggests that one must assume either that familiarity processes can underlie or benefit from the mechanisms proposed by current theories to account for the testing effect or that in some cases current theoretical explanations of the testing effect do not fully account for the effect.

Further, many studies, in reporting findings, focus on the benefit of testing to recollection-related processes (e.g., source memory, memory for context; e.g., Akan, Stanley, & Benjamin, 2018; Brewer, Marsh, Meeks, Clark-Foos, & Hicks, 2010; Hong, Polyn, & Fazio, 2019), and dismiss, fail to report, or do not provide a means of assessing effects in familiarity (e.g., Guran et al., 2020). Here we provide further evidence that observing a testing effect in overall accuracy does not imply the existence solely of benefits to recollective processes. Thus, future work should endeavor to report effects of testing on familiarity-related processes when relevant.

4.4 Implications for Aging Populations

Bishara and Jacoby (2008) observe a greater benefit of spaced retrieval in familiarity than in recollection in older adults and warn: "the present results suggest that older adults and perhaps other populations with controlled memory impairments might suffer from inflexible behavior following spaced retrieval practice" (pg. 56). The suggestion is that with larger increases in familiarity than in recollection as a result of testing, the benefits of testing in older adults may be restricted to situations in which automatic processing (familiarity) will lead to accurate performance. Using a standard testing effect paradigm, however, the present study observed a testing effect in older adults in both familiarity and in recollection, albeit a smaller testing effect in recollection than that observed in younger adults. These results suggest that given particular initial testing conditions, retrieval practice can be useful for improving later memory performance for older adults even when successful performance will require controlled processing or recollection (e.g., recalling the source of information, memory for new associations). This suggestion aligns with prior work that has revealed a testing effect in older adults in final free recall (Rabinowitz & Craik, 1986; Rogers & Gilbert, 1997, although see Henkel, 2014), which is thought to rely heavily on recollective processing for success (Yonelinas, 2002, although automatic influences can contribute as well, see McCabe, Roediger, & Karpicke, 2011).

Notably, however, the magnitude of the testing effect in recollection—but not familiarity—was reduced in the older relative to the younger adults. This suggests that, although the mechanisms that support the testing effect in younger and older adults are perhaps similar, with increased age the benefits of retrieval practice to recollection-related processing may be more limited. Of course, the present results may hinge heavily on particular attributes of the

current sample of older adults—who achieved high levels of performance on both initial and final testing. This issue will be considered further in the limitations section below.

4.5 Limitations and Future Directions

There are several limitations of the present study, as well as avenues for further research that present themselves. First, the present study observes the effects of prior testing relative to a no test control condition. However, re-exposure to material can serve to increase estimates of familiarity (Jacoby, 1991). To the extent that familiarity is found to support the testing effect, it is important to ascertain whether retrieval practice per se, and not simple re-exposure to material during testing, accounts for the change in familiarity. While we cannot address this issue in the current study, prior literature that uses a restudy control condition (Bies-Hernandez, 2013) and that observes changes from recollective to familiarity-based processing across repeated tests (Conway et al., 1997; Dewhurst et al., 2009) suggests that the effect in familiarity estimates is unlikely to be due merely to re-exposure effects.

In addition, the present study conducted initial cued-recall testing to enable more direct comparison with prior work (e.g., Shaffer & McDermott, 2020; Chan & McDermott, 2007, Exp. 3). While not necessarily a limitation of the present study, the generalizability of the present results to other forms of initial testing has yet to be determined. Indeed, the benefit of retrieval practice has been shown to be influenced by the form of initial test—such as the use of recognition- or recall-based initial testing (Kang, McDermott, & Roediger, 2007; Roediger & Karpicke, 2006b)—as well as by other design choices—such as the lag between initial studying and the first test of an item (Carpenter & DeLosh, 2005; Pyc & Rawson, 2007, 2009; Roediger & Karpicke, 2006b, 2010). Critically, when initial learning task demands are less constrained (and success may derive from various processes or strategies during initial testing), older adults may

exhibit less effective processing strategies (Craik & Rabinowitz, 1985; Rabinowitz & Craik, 1986). This suggests that particularly in older adults the benefit of retrieval practice observed in recollection and familiarity may depend on the form of initial testing employed (e.g., recognition, free recall, continuous vs. blocked initial testing). Future work should explore the generalizability of the present findings in older adults by varying the type and difficulty of the initial testing task.

The present study assesses recollection and familiarity at final testing via the Remember-Know procedure (Tulving, 1985; Gardiner, 1988), which necessarily involves introspection on the part of the subject and incorporates distinctions in memory that can be difficult to understand (McCabe & Geraci, 2009; Migo, Mayes, & Montaldi, 2012). This situation leaves open the possibility that older and younger adults interpret or comply with the final test instructions differently, which could lead to systematic differences in responding and muddy the interpretation of results (Koen & Yonelinas, 2014, 2016). Although it is possible, it is for several reasons unlikely that age-related differences in the effects of testing observed in the present study are due simply to differences in interpretation or adherence to the Remember-Know instructions.

Specifically, all participants were given extensive Remember-Know instructions that incorporated recommendations from the literature (Koen & Yonelinas, 2016; Migo et al., 2012; Yonelinas, 2001b). Instructions asked subjects to use a Remember response only when subjects would be able to report the detail leading them to make the Remember response (Koen & Yonelinas, 2016; Yonelinas, 2001b). When this best practice instruction is given, estimates of recollection and familiarity in older and younger adults derived from the Remember-Know procedure generally align with those of other procedures for estimating recollection and familiarity (Koen & Yonelinas, 2016). Further, in the present study, after instructions were

provided, subjects were asked to describe in detail each response type. The experimenter noted subjects who struggled with the instructions, although all subjects eventually provided correct explanations before completing the final recognition test. Supplementary analyses that excluded subjects who had initially struggled to understand the instructions were conducted, and all findings relating to age effects (or lack thereof) in the magnitude of the testing effect remained the same. Despite the above considerations, future work should seek to replicate these findings using less subjective or introspective techniques for obtaining estimates of recollection and familiarity (e.g., a confidence-based or process-dissociation final test procedures). The use of other techniques would ensure that the present findings are not the result of interpretational differences of the Remember-Know instructions in older and younger adults.

Another limitation of the present study concerns the representativeness of the older adult sample and the resulting generalizability of the present results to the healthy older adult population. Efforts were made to include a representative older adult sample by recruiting older subjects via Volunteer for Health (https://sites.wustl.edu/wuvfh/), which draws subjects from throughout the greater St. Louis area. In addition, response time analyses revealed the expected patterns of age-related effects: older adults responded significantly more slowly than younger adults across all response types on the final test. Finally, the expected reductions in baseline (no test) estimates of recollection for older relative to younger adults were observed numerically. Problematically, however, this observed reduction in baseline recollection was only marginally significant and no longer reached significance in several covariate and exclusion analyses (although significant age-related reductions in the magnitude of the *testing effect* in recollection remained in all cases). In addition, older and younger adults exhibited similar levels of overall accuracy during initial cued-recall and final recognition testing. Although prior work reveals a

benefit of prior testing of similar magnitude in older and younger adults, overall final test performance typically exhibits reductions with age. Together these results suggest that the present older adult sample fell on the higher end of cognitive function and was perhaps a selective sample within this age group.

Critically, improvements in later memory performance following initial testing depend upon a certain level of initial test performance (Kornell et al., 2011; Rowland, 2014). One would not expect to see a testing effect in later memory performance in older adults with extremely poor performance during initial testing, at least in the absence of feedback (e.g., see Tse et al., 2010). Further, with more marked reductions in recollection, initial test performance may rely on different processes or strategies for success. This situation could, in turn, influence later estimates of the testing effect in recollection and familiarity. Thus, in a more representative older adult sample in which significant reductions are observed in baseline recollection and overall accuracy, different conclusions regarding the mechanisms that underlie the testing effect in older adults may be reached. Further, the present results likely do not apply to older adult groups with more severely restricted recollection, such as is observed in populations with amnesic mild cognitive impairment or Alzheimer's Disease (Anderson et al., 2008; Koen & Yonelinas, 2014; Pitarque et al., 2016). These groups may need a great deal more support during initial testing to benefit from retrieval practice, such as via the use of spaced retrieval practice with relatively short lags between initial study and testing, along with the provision of feedback (Balota, Duchek, Sergent-Marshall, & Roediger, 2006) or via a spaced retrieval shaping procedure (e.g., Camp et al., 1996; Camp & Schaller, 1989; Cherry et al., 1999).

Finally, the present older adult sample had an age range of 65-82 years (M = 71.27 yo; SD = 4.45 yo). Only 13 participants (of 64 older adults) were age 75 and above. Critically,

within the older adult sample in the present study, increased age related to decreases in the magnitude of the testing effect in recollection (r_s = -.38, p = .002), even after controlling for differences in initial test performance (*partial* r_s = -.41, p = .001). This suggests that in a generally older sample than that obtained in the present study, one might not expect to see a testing effect in recollection. Indeed, for the oldest one-third of the present older adult sample (M = 76.2 yo; range = 73-82 yo), the testing effect in recollection failed to reach significance at p < .05 (t = 1.99, p = .061), whereas the testing effect in familiarity (t = 3.08, p = .006) and in overall hit rates (t = 4.08, p = .001) were significant at conventional thresholds. By contrast, in the youngest one-third of the older adult sample (M = 66.6 yo; range = 65-69 yo), the testing effect was significant in all cases (i.e., the testing effect in recollection, familiarity, and overall hit rates: all *ts* >= 3.50, all *ps* <= .002). This suggests that the present results regarding the existence of a robust testing effect in recollection in older adults may not extend to more elderly populations, where increases in familiarity may better account for the testing effect.

Future work that extends the present findings to a more diverse population of older adults will enhance our understanding of the mechanisms that underlie the testing effect in aging populations, as well as the situations in which retrieval practice can be expected to benefit later memory performance.

4.6 Conclusion

Further work is necessary in order to resolve the discrepancies that exist in the literature as to when a testing effect in familiarity will appear. However, the present study observed a benefit of retrieval practice to both recollection and familiarity processing in both younger and older adults, with consistent findings from an immediate to a 1-day delayed final test. Thus, although not always observed, the present findings bolster the notion that, rather than being a

recollection-only phenomenon, the testing effect can be supported by changes in both recollection and familiarity. Further, whereas the benefit to familiarity appears to persist across age, the benefit to recollection may decline. These findings at once suggest that retrieval practice can lead to memorial benefits in older adults that are qualitatively similar to those observed in younger adults and warn of potential constraints on the benefits with increasing age. Together, the present results pose challenges for theories of the testing effect that propose a primarily recollection-related phenomenon. Our understanding of the mechanisms that underlie the testing effect will benefit from future work that explores the generalizability of the present results to more elderly populations, as well as to the use of other testing procedures.

References

Akan, M., Stanley, S. E., & Benjamin, A. S. (2018). Testing enhances memory for context. *Journal of Memory and Language*, *103*, 19–27. DOI:10.1016/j.jml.2018.07.003

Anderson, N. D., Ebert, P. L., Jennings, J. M., Grady, C. L., Cabeza, R., & Graham, S. J. (2008). Recollection- and familiarity-based memory in healthy aging and amnestic mild cognitive impairment. *Neuropsychology*, *22*(2), 177–187. DOI:10.1037/0894-4105.22.2.177

Balota, D. A., Duchek, J. M., Sergent-Marshall, S. D., & Roediger, H. L. (2006). Does expanded retrieval produce benefits over equal-interval spacing? Explorations of spacing effects in healthy aging and early stage Alzheimer's Disease. *Psychology and Aging*, *21*(1), 19–31. DOI:10.1037/0882-7974.21.1.19

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., … Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*(3), 445–459. DOI:10.3758/BF03193014

Bastin, C., & Van der Linden, M. (2003). The contribution of recollection and familiarity to recognition memory: A study of the effects of test format and aging. *Neuropsychology*, *17*(1), 14–24. DOI:10.1037/0894-4105.17.1.14

Bies-Hernandez, N. J. (2013). *Examining the testing effect using the dual-process signal detection model* (Doctoral Dissertation, University of Nevada, Las Vegas). Retrieved from https://digitalscholarship.unlv.edu/thesesdissertations/1804

Bishara, A. J., & Jacoby, L. L. (2008). Aging, spaced retrieval, and inflexible memory performnace. *Psychonomic Bulletin & Review*, *15*(1), 52–57. DOI:10.3758/PBR.15.1.52

Brewer, G. A., Marsh, R. L., Meeks, J. T., Clark-Foos, A., & Hicks, J. L. (2010). The effects of free recall testing on subsequent source memory. *Memory*, *18*(4), 385–393. DOI:10.1080/09658211003702163

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, *3*(3), 296–322. DOI:10.1111/j.2044-8295.1910.tb00207.x

Camp, C. J., Foss, J. W., Stevens, A. B., & O'Hanlon, A. M. (1996). Improving prospective memory task performance in persons with Alzheimer's disease. In M. Brandimonte, G. O. Einstein, & M. A. McDaniel (Eds.), *Prospective memory: Theory and applications* (pp. 351–367). Retrieved from https://psycnet.apa.org/record/2002-02930-018

Camp, C. J., & Schaller, J. R. (1989). Epilogue: Spaced-retrieval memory training in an adult day-care center. *Educational Gerontology*, *15*(6), 641–648. DOI:10.1080/0380127890150608

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(6), 1563–1569. DOI:10.1037/a0017021

Carpenter, S. K., & DeLosh, E. L. (2005). Application of the testing and spacing effects to name learning. *Applied Cognitive Psychology*, *19*(5), 619–636. DOI:10.1002/acp.1101

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition*, *34*(2), 268–276. DOI:10.3758/BF03193405

Carpenter, S. K., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued recall test? *Psychonomic Bulletin & Review*, *13*(5), 826–830. DOI:10.3758/BF03194004

Carpenter, S. K., Pashler, H., Wixted, J. T., & Vul, E. (2008). The effects of tests on learning and forgetting. *Memory & Cognition*, *36*(2), 438–448. DOI:10.3758/MC.36.2.438

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition*, *20*(6), 633–642. DOI:10.3758/BF03202713

Chan, J. C. K., & McDermott, K. B. (2007). The testing effect in recognition memory: A dual process account. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(2), 431–437. DOI:10.1037/0278-7393.33.2.431

Cherry, K. E., Simmons, S. S., & Camp, C. J. (1999). Spaced retrieval enhances memory in older adults with probable Alzheimer's disease. *Journal of Clinical Geropsychology*, *5*(3), 159–175. DOI:10.1023/A:1022983131186

Coane, J. H. (2013). Retrieval practice and elaborative encoding benefit memory in younger and older adults. *Journal of Applied Research in Memory and Cognition*, *2*(2), 95–100. DOI:10.1016/j.jarmac.2013.04.001

Conway, M. A., Gardiner, J. M., Perfect, T. J., Anderson, S. J., & Cohen, G. M. (1997). Changes in memory awareness during learning: The acquisition of knowledge by psychology undergraduates. *Journal of Experimental Psychology: General*, *126*(4), 393–413. DOI:10.1037/0096-3445.126.4.393

Craik, F. I. M., & Rabinowitz, J. C. (1985). The Effects of Presentation Rate and Encoding Task on Age-Related Memory Deficits. In *Journal of Gerontology* (Vol. 40). Retrieved from https://academic.oup.com/geronj/article-abstract/40/3/309/658903

Dewhurst, S. A., Conway, M. A., & Brandt, K. R. (2009). Tracking the R-to-K shift: Changes in memory awareness across repeated tests. *Applied Cognitive Psychology*, *23*(6), 849–858. DOI:10.1002/acp.1517

Duarte, A., Ranganath, C., Trujillo, C., & Knight, R. T. (2006). Intact recollection memory in high-performing older adults: ERP and behavioral evidence. *Journal of Cognitive Neuroscience*, *18*(1), 33–47. DOI:10.1162/089892906775249988

Dudukovic, N. M., DuBrow, S., & Wagner, A. D. (2009). Attention during memory retrieval enhances future remembering. *Memory & Cognition*, *37*(7), 953–961. DOI:10.3758/MC.37.7.953

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state": A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, *12*(3), 189–198. DOI:10.1016/0022-3956(75)90026-6

Gao, C., Rosburg, T., Hou, M., Li, B., Xiao, X., & Guo, C. (2016). The role of retrieval mode and retrieval orientation in retrieval practice: Insights from comparing recognition memory testing formats and restudying. *Cognitive, Affective, & Behavioral Neuroscience*, *16*(6), 977–990. DOI:10.3758/s13415-016-0446-z

Gardiner, J. M. (1988). Functional aspects of recollective experience. *Memory & Cognition*, *16*(4), 309–313. DOI:10.3758/BF03197041

Gardiner, J. M., & Java, R. I. (1991). Forgetting in recognition memory with and without recollective experience. *Memory & Cognition*, *19*(6), 617–623. DOI:10.3758/BF03197157

Guran, C.-N. A., Herweg, N. A., & Bunzeck, N. (2019). Age-related decreases in the retrieval practice effect directly relate to changes in alpha-beta oscillations. *The Journal of Neuroscience*, *39*(22), 4344–4352. DOI:10.1523/JNEUROSCI.2791-18.2019

Guran, C.-N. A., Lehmann-Grube, J., & Bunzeck, N. (2020). Retrieval practice improves recollection-based memory over a seven-day period in younger and older adults. *Frontiers in Psychology*, *10*, Article 2997. DOI:10.3389/fpsyg.2019.02997

Henkel, L. A. (2014). The retrieval context of intervening tasks influences subsequent memory in younger and older adults. *Experimental Aging Research*, *40*(5), 555–577.

DOI:10.1080/0361073X.2014.956622
Hong, M. K., Polyn, S. M., & Fazio, L. K. (2019). Examining the episodic context account: Does retrieval practice enhance memory for context? *Cognitive Research: Principles and Implications*, *4*(46), 1–9. DOI:10.1186/s41235-019-0202-3
Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, *30*(5), 513–541. DOI:10.1016/0749-596X(91)90025-F
Jennings, J. M., & Jacoby, L. L. (1997). An opposition procedure for detecting age-related deficits in recollection: Telling effects of repetition. *Psychology and Aging*, *12*(2), 352–361. DOI:10.1037/0882-7974.12.2.352
Jennings, J. M., & Jacoby, L. L. (2003). Improving memory in older adults: Training recollection. *Neuropsychological Rehabilitation*, *13*(4), 417–440. DOI:10.1080/09602010244000390
Jia, X., Gao, C., Cui, L., & Guo, C. (2019). The role of emotion arousal in the retrieval practice effect. *Experimental Brain Research*, *237*, 3241–3252. DOI:10.1007/s00221-019-05658-0
Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source Monitoring. *Psychological Bulletin*, *114*(1), 3–28. Retrieved from http://memlab.yale.edu/sites/default/files/files/1993_Johnson_Hashtroudi_Lindsay_PsychBull.pdf
Jones, T. C., & Roediger, H. L. (1995). The experiential basis of serial position effects. *European Journal of Cognitive Psychology*, *7*(1), 65–80. DOI:10.1080/09541449508520158
Kang, S. H. K., McDermott, K. B., & Roediger, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, *19*(4–5), 528–558. DOI:10.1080/09541440601056620
Karpicke, J. D. (2017). Retrieval-Based Learning: A Decade of Progress. In J. T. Wixted & J. H. Byrne (Eds.), *Learning and Memory: A Comprehensive Reference (Second Edition)* (pp. 487–514). DOI:10.1016/B978-0-12-809324-5.21055-9
Karpicke, J. D., Blunt, J. R., & Smith, M. A. (2016). Retrieval-based learning: Positive effects of retrieval practice in elementary school children. *Frontiers in Psychology*, *7*, Article 350. DOI:10.3389/fpsyg.2016.00350
Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning: An episodic context account. In B. H. Ross (Ed.), *Psychology of Learning and Motivation* (Vol. 61, pp. 237–284). DOI:10.1016/B978-0-12-800283-4.00007-1
Karpicke, J. D., & Roediger, H. L. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language*, *57*(2), 151–162. DOI:10.1016/j.jml.2006.09.004
Kausler, D. H., & Phillips, P. L. (1988). Instructional variation and adult age differences in activity memory. *Experimental Aging Research*, *14*(4), 195–199. DOI:10.1080/03610738808259747
Kausler, D. H., & Wiley, J. G. (1991). Effects of short-term retrieval on adult age differences in long-term recall of actions. *Psychology and Aging*, *6*(4), 661–665. DOI:10.1037/0882-7974.6.4.661
Kessler, Y., Vandermorris, S., Gopie, N., Daros, A., Winocur, G., & Moscovitch, M. (2014). Divided attention improves delayed, but not immediate retrieval of a consolidated memory. *PLoS ONE*, *9*(3), e91309. DOI:10.1371/journal.pone.0091309

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. DOI:10.1007/s11065-014-9266-5

Koen, J. D., & Yonelinas, A. P. (2016). Recollection, not familiarity, decreases in healthy aging: Converging evidence from four estimation methods. *Memory*, *24*(1), 75–88. DOI:10.1038/nmeth.2839.A

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language*, *65*(2), 85–97. DOI:10.1016/j.jml.2011.04.002

Larsen, D. P., Butler, A. C., & Roediger, H. L. (2008). Test-enhanced learning in medical education. *Medical Education*, *42*(10), 959–966. DOI:10.1111/j.1365-2923.2008.03124.x

Larsen, D. P., Butler, A. C., & Roediger, H. L. (2013). Comparative effects of test-enhanced learning and self-explanation on long-term retention. *Medical Education*, *47*(7), 674–682. DOI:10.1111/medu.12141

Lehman, M., Smith, M. A., & Karpicke, J. D. (2014). Toward an episodic context account of retrieval-based learning: Dissociating retrieval practice and elaboration. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(6), 1787–1794. DOI:10.1037/xlm0000012

Logan, J. M., & Balota, D. A. (2008). Expanded vs. equal interval spaced retrieval practice: Exploring different schedules of spacing and retention interval in younger and older adults. *Aging, Neuropsychology, and Cognition*, *15*(3), 257–280. DOI:10.1080/13825580701322171

Maddox, G. B., & Balota, D. A. (2015). Retrieval practice and spacing effects in young and older adults: An examination of the benefits of desirable difficulty. *Memory & Cognition*, *43*, 760–774. DOI:10.3758/s13421-014-0499-6

McCabe, D. P., & Geraci, L. D. (2009). The influence of instructions and terminology on the accuracy of remember-know judgments. *Consciousness and Cognition*, *18*(2), 401–413. DOI:10.1016/j.concog.2009.02.010

McCabe, D. P., Roediger, H. L., & Karpicke, J. D. (2011). Automatic processing influences free recall: Converging evidence from the process dissociation procedure and remember-know judgments. *Memory & Cognition*, *39*, 389–402. DOI:10.3758/s13421-010-0040-5

McCabe, D. P., Roediger, H. L., McDaniel, M. A., & Balota, D. A. (2009). Aging reduces veridical remembering but increases false remembering: Neuropsychological test correlates of remember–know judgments. *Neuropsychologia*, *47*(11), 2164–2173. DOI:10.1016/j.neuropsychologia.2008.11.025

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, *19*(4/5), 494–513. DOI:10.1080/09541440701326154

McDermott, K. B., Agarwal, P. K., D'Antonio, L., Roediger, H. L., & McDaniel, M. A. (2014). Both multiple-choice and short-answer quizzes enhance later exam performance in middle and high school classes. *Journal of Experimental Psychology: Applied*, *20*(1), 3–21. DOI:10.1037/xap0000004

Meyer, A. N. D., & Logan, J. M. (2013). Taking the Testing Effect Beyond the College Freshman: Benefits for Lifelong Learning. *Psychology and Aging*, *28*(1), 142–147. DOI:10.1037/a0030890.supp

Migo, E. M., Mayes, A. R., & Montaldi, D. (2012). Measuring recollection and familiarity:

Improving the remember/know procedure. *Consciousness and Cognition*, *21*(3), 1435–1455. DOI:10.1016/j.concog.2012.04.014

Naveh-Benjamin, M. (2000). Adult age differences in memory performance: Tests of an associative deficit hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(5), 1170–1187. DOI:10.1037/0278-7393.26.5.1170

Parks, C. M., & Yonelinas, A. P. (2007). Moving beyond pure signal-detection models: Comment on Wixted (2007). *Psychological Review*, *114*(1), 188–202. DOI:10.1037/0033-295X.114.1.188

Pitarque, A., Meléndez, J. C., Sales, A., Mayordomo, T., Satorres, E., Escudero, J., & Algarabel, S. (2016). The effects of healthy aging, amnestic mild cognitive impairment, and Alzheimer's disease on recollection, familiarity and false recognition, estimated by an associative process-dissociation recognition procedure. *Neuropsychologia*, *91*, 29–35. DOI:10.1016/j.neuropsychologia.2016.07.010

Prull, M. W., Dawes, L. L. C., Martin, A. M., Rosenberg, H. F., & Light, L. L. (2006). Recollection and familiarity in recognition memory: Adult age differences and neuropsychological test correlates. *Psychology and Aging*, *21*(1), 107–118. DOI:10.1037/0882-7974.21.1.107

Pu, X., & Tse, C.-S. (2014). The influence of intentional versus incidental retrieval practices on the role of recollection in test-enhanced learning. *Cognitive Processing*, *15*(1), 55–64. DOI:10.1007/s10339-013-0580-2

Pyc, M. A., & Rawson, K. A. (2007). Examining the efficiency of schedules of distributed retrieval practice. *Memory & Cognition*, *35*(8), 1917–1927. DOI:10.3758/BF03192925

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, *60*(4), 437–447. DOI:10.1016/j.jml.2009.01.004

Rabinowitz, J. C., & Craik, F. I. M. (1986). Prior retrieval effects in young and old adults. *Journal of Gerontology*, *41*(3), 368–375. DOI:10.1093/geronj/41.3.368

Rawson, K. A., Vaughn, K. E., & Carpenter, S. K. (2015). Does the benefit of testing depend on lag, and if so, why? Evaluating the elaborative retrieval hypothesis. *Memory & Cognition*, *43*(4), 619–633. DOI:10.3758/s13421-014-0477-z

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20–27. DOI:10.1016/j.tics.2010.09.003

Roediger, H. L., & Karpicke, J. D. (2006a). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249–255. DOI:10.1111/j.1467-9280.2006.01693.x

Roediger, H. L., & Karpicke, J. D. (2006b). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*(3), 181–210. DOI:10.1111/j.1745-6916.2006.00012.x

Roediger, H. L., & Karpicke, J. D. (2010). Intricacies of spaced retrieval: A resolution. In A. S. Benjamin (Ed.), *Successful Remembering and Successful Forgetting: Essays in Honor of Robert A. Bjork* (pp. 23–47). Psychology Press.

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

Rogers, W. A., & Gilbert, D. K. (1997). Do performance strategies mediate age-related differences in associative learning? *Psychology and Aging*, *12*(4), 620–633.

DOI:10.1037/0882-7974.12.4.620
Rowland, C. A. (2011). *Testing effects in context memory* (Master's Thesis, Colorado State University). Retrieved from https://mountainscholar.org/bitstream/handle/10217/46908/Rowland_colostate_0053N_10630.pdf?sequence=1&isAllowed=y
Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432–1463. DOI:10.1037/a0037559
Shaffer, R. A., & McDermott, K. B. (2020). A role for familiarity in supporting the testing effect over time. *Neuropsychologia*, *138*, 1–14. DOI:10.1016/j.neuropsychologia.2019.107298
Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, *3*(3), 271–295. DOI:10.1111/j.2044-8295.1910.tb00206.x
Tse, C.-S., Balota, D. A., & Roediger, H. L. (2010). The benefits and costs of repeated testing on the learning of face-name pairs in healthy older adults. *Psychology and Aging*, *25*(4), 833–845. DOI:10.1037/a0019933
Tulving, E. (1967). The effects of presentation and recall of material in free-recall learning. *Journal of Verbal Learning and Verbal Behavior*, *6*(2), 175–184. DOI:10.1016/S0022-5371(67)80092-6
Tulving, E. (1985). Memory and consciousness. *Canadian Psychology/Psychologie Canadienne*, *26*(1), 1–12. DOI:10.1037/h0080017
Verkoeijen, P. P. J. L., Tabbers, H. K., & Verhage, M. L. (2011). Comparing the effects of testing and restudying on recollection in recognition memory. *Experimental Psychology*, *58*(6), 490–498. DOI:10.1027/1618-3169/a000117
Whiffen, J. W., & Karpicke, J. D. (2017). The role of episodic context in retrieval practice effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(7), 1036 –1046. DOI:10.1037/xlm0000379
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*(6), 1341–1354. DOI:10.1037/0278-7393.20.6.1341
Yonelinas, A. P. (2001a). Components of episodic memory: The contribution of recollection and familiarity. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *356*(1413), 1363–1374. DOI:10.1098/rstb.2001.0939
Yonelinas, A. P. (2001b). Consciousness, control, and confidence: The 3 Cs of recognition memory. *Journal of Experimental Psychology: General*, *130*(3), 361–379. DOI:10.1037/0096-3445.130.3.361
Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, *46*(3), 441–517. DOI:10.1006/JMLA.2002.2864
Yonelinas, A. P., Aly, M., Wang, W.-C., & Koen, J. D. (2010). Recollection and familiarity: Examining controversial assumptions and new directions. *Hippocampus*, *20*(11), 1178–1194. DOI:10.1002/hipo.20864
Yonelinas, A. P., & Jacoby, L. L. (1995). The relation between remembering and knowing as bases for recognition: Effects of size congruency. *Journal of Memory and Language*, *34*(5), 622–643. DOI:10.1006/jmla.1995.1028
Yonelinas, A. P., & Parks, C. M. (2007). Receiver operating characteristics (ROCs) in recognition memory: A review. *Psychological Bulltein*, *133*(5), 800–832. DOI:10.1037/0033-2909.133.5.800

Zaromb, F. M., & Roediger, H. L. (2010). The testing effect in free recall is associated with enhanced organizational processes. *Memory & Cognition*, *38*(8), 995–1008. DOI:10.3758/MC.38.8.995