Psychology and Aging
2010, Vol. 25, No. 4, 833–845


0882-7974/10/$12.00 DOI: 10.1037/a0019933

# The Benefits and Costs of Repeated Testing on the Learning of Face–Name Pairs in Healthy Older Adults

Chi-Shing Tse
The Chinese University of Hong Kong

David A. Balota and Henry L. Roediger, III
Washington University in St. Louis

We compared the benefits of repeated testing and repeated study on cued recall of unfamiliar face–name pairs in healthy middle-aged and older adults. We extended Karpicke and Roediger's (2008) paradigm to compare the effects of repeated study versus repeated testing after each face–name pair was correctly recalled once. The results from Experiment 1, which provided no feedback during the acquisition phase, yielded a crossover interaction: Middle-aged adults showed the expected benefit of repeated testing, whereas older adults produced a benefit of repeated study. When participants were given feedback in Experiment 2, both middle-aged and older adults benefited from repeated testing. We suggest that for face–name pairs, feedback may be particularly important for individuals who have relatively poor memory to produce benefits from repeated testing.

*Keywords:* face–name learning, repeated study, repeated testing, older adults

Tests have often been regarded as a way of measuring learning rather than as a method to facilitate learning. While preparing for an exam, most college students choose to repeatedly read their notes or textbooks rather than to self-test (e.g., Karpicke, Butler, & Roediger, 2009; Kornell & Bjork, 2007). However, many studies have shown a *testing effect*—the advantage in long-term retention for materials that are repeatedly tested over those that are represented for additional study during the retention interval (see Roediger & Karpicke, 2006, for a review). The robust benefit of testing on retention has been found with study materials that include paired associates (e.g., Carrier & Pashler, 1992), pictures (e.g., Wheeler & Roediger, 1992), semantic associates (e.g., McDermott, 2006), general knowledge facts (e.g., McDaniel & Fisher, 1991), and textbook passages (e.g., Agarwal, Karpicke, Kang, Roediger, & McDermott, 2008). In addition, repeated testing has been shown to produce benefits in free recall (e.g., Karpicke & Roediger, 2007; Tulving, 1967), cued recall (e.g., Carpenter, Pashler, & Vul, 2006), and episodic recognition (e.g., Glover, 1989).

Most of the testing-effect studies have focused on young-adult populations (i.e., high school or college students), perhaps due to the obvious relevance for classroom learning. In contrast, relatively few studies have directly examined the testing effect in healthy middle-aged and older adults. In the present study, we address this issue by using face–name pairs as study materials, because older adults often report considerable difficulty remembering such items (Cohen & Faulkner, 1986). Before reporting the experiments, we first review the literature on face–name learning in older adults and outline the repeated study/repeated testing paradigm introduced by Karpicke and Roediger (2008). According to the extant literature, which suggests breakdowns in associative binding (Naveh-Benjamin, 2000) and attentional control processes in older adults (Balota & Faust, 2001), one might expect that repeated testing (without feedback) might not always be beneficial for older adults, due to the difficulty of recalling the correct name paired with a given face across the tests. When there is no feedback during the acquisition phase, older adults may actually benefit more from repeated study than repeated testing.

## Face–Name Learning in Healthy Older Adults

There is considerable evidence indicating that names are more difficult to recall than other information about a person, such as occupation (e.g., Cohen, 1990; Cohen & Burke, 1993). This occurs even when the same word (e.g., *cook*) is introduced as an occupation for one group and as a last name for another group (e.g., James, 2004; McWeeny, Young, Hay, & Ellis, 1987; Rendell, Castel, & Craik, 2005). Compared with young adults, older adults show even larger impairments in learning face–name pairs compared with face–occupation pairs in cued recall (e.g., Barresi, Obler, & Goodglass, 1998), associative recognition (e.g., Naveh-Benjamin, Guez, Kilb, & Reedy, 2004; Naveh-Benjamin et al., 2009), and forced-choice recognition (e.g., James, Fogler, & Tauber, 2008). Because of the difficulty in acquiring face–name

This article was published Online First August 16, 2010.

Chi-Shing Tse, Department of Educational Psychology, The Chinese University of Hong Kong, Hong Kong, China; David A. Balota and Henry L. Roediger, III, Department of Psychology, Washington University in St. Louis.

This work was supported by National Institute on Aging Grants PO1AG003991 and PO1AG026276. We thank John Morris and the Clinical Core at the Washington University Alzheimer's Disease Research Center for their careful recruitment and description of the healthy older adults; Jeffrey Karpicke for his comments as well as providing the experimental programs; Jessica Logan and Pooja Agarwal for constructing the stimuli; Martha Storandt for providing the psychometric data; Rebecca Howard for coding participants' responses; and Christopher Grant and Betsy Hemphill for testing the participants.

Correspondence concerning this article should be addressed to Chi-Shing Tse, Department of Educational Psychology, The Chinese University of Hong Kong, Hong Kong, China; or David A. Balota, Department of Psychology, Campus Box 1125, Washington University in St. Louis, One Brookings Drive, St. Louis, MO 63130-4899. E-mail: cstse@cuhk.edu.hk or dbalota@artsci.wustl.edu

pairs, several mnemonic aids have been proposed to enhance acquisition for both young adults (e.g., Helder & Shaughnessy, 2008; Neuschatz, Preston, Toglia, & Neuschatz, 2005) and older adults (e.g., Cohen, 1990; Groninger & Murray, 2004; Troyer, Hafliger, Cadieux, & Craik, 2006). However, only a few studies have examined the effect of repeated testing on the acquisition of face–name pairs in older adults (Barresi et al., 1998; James, 2004; James et al., 2008). For example, Barresi et al. showed that when older adults were tested twice, they showed an improvement on memory for unfamiliar face–name associations, albeit not as much as on memory for unfamiliar face–occupation associations. When older adults were given immediate feedback for their incorrect responses, James (2004; see also James et al., 2008) reported similar findings in multiple-choice recognition memory tests. However, neither of these studies directly contrasted the effects of repeated testing versus repeated study on older adults' performance to examine the mnemonic benefit of repeated testing.

## Karpicke and Roediger's (2008) Testing-Effect Paradigm

In the Karpicke and Roediger (2008) study, college students learned 40 Swahili–English word pairs in four study–test cycles during an acquisition phase. At study, participants saw the word pairs presented one at a time on a screen. A 30-s distractor task followed list presentation; then, participants were given each Swahili word as a cue and asked to recall its corresponding English translation. The participants were assigned to one of four acquisition conditions: standard, adjusted learning, test dropout (i.e., repeated study), and study dropout (i.e., repeated testing). Here we focus on the test dropout and study dropout conditions because these conditions yielded powerful evidence for a testing effect.

The repeated testing and repeated study conditions differed in what occurred across the four cycles during the acquisition phase. Specifically, once a word pair was correctly recalled in one of the four cycles, it was either (a) removed from the test list but still retained on the study list in the subsequent cycles (*repeated study*) or (b) removed from the study list but still retained on the test list in the subsequent cycles (*repeated testing*). After 1 week, all participants were tested for all 40 word pairs in a final cued-recall test. It is worth noting that feedback was not given during acquisition or during the final cued-recall test. The results indicated that participants in the repeated testing condition showed dramatically better final recall performance (~80%) than those in the repeated study condition (~35%). All participants in the repeated study and repeated testing conditions went through four study–test cycles before the termination of the acquisition phase, and the number of study–test trials in the acquisition phase did not differ for the two conditions. Hence, the robust testing effect was not compromised by participants having seen the word pairs more often in the repeated testing condition. Finally, a self-report questionnaire given at the end of the acquisition phase indicated that participants in all conditions predicted that they would recall only about 50% of the word pairs, so they overestimated their performance in the repeated study condition and underestimated their performance in the repeated testing condition. Similar to the typical college students (Karpicke et al., 2009), these participants were unaware of the benefit of repeated testing on long-term retention.

## The Role of Testing Versus Study in Individuals With Compromised Episodic Memory

There is substantial evidence of a breakdown in episodic memory performance in healthy older adults compared with younger adults, and multiple theoretical mechanisms have been proposed to account for these deficits (see McDaniel, Einstein, & Jacoby, 2008, for a review). Interestingly, older adults often benefit from the same mnemonic techniques as young adults but overall produce lower performance. For example, older adults clearly benefit from spacing during study as much as younger adults but produce overall lower performance (see, for example, Balota, Duchek, & Paullin, 1989; Balota, Duchek, Sergent-Marshall, & Roediger, 2006). Similarly, older adults also benefit strongly from depth of processing (e.g., Troyer et al., 2006) and other mnemonic techniques such as the method of loci (e.g., Verhaeghen & Marcoen, 1996). Hence, one might expect older adults to also benefit more from repeated testing compared to repeated study, albeit producing overall lower performance. On the other hand, it is also possible that individuals with relatively poor episodic memory may actually benefit less than younger adults from repeated testing compared with repeated study. Specifically, in the repeated testing procedure, individuals with episodic memory deficits may sometimes either forget the correct face–name pair or possibly produce an incorrect face–name pair, thereby confusing the names with the wrong faces. Because participants do not receive feedback in the repeated testing session during acquisition or have the opportunity to restudy the items (unless they never retrieved an item correctly to begin with), it is possible that there may be confusion between the correct response and the earlier incorrect responses, which would compromise final recall performance. In contrast, in the repeated study condition, individuals repeatedly study the correct face–name pairs on every trial, even when they have made a prior mistake on an item, and so there is an opportunity on each trial to correct any previous incorrectly stored face–name pair. Thus, in contrast to other mnemonic techniques, it is possible that individuals with poor episodic memory may not benefit as much from repeated testing compared with repeated study and may actually produce a reversal of the testing effect.[1]

## Present Research

The goal of the present study was to examine the testing effect on the retention of face–name pairs, using Karpicke and Roediger's (2008) repeated study and repeated testing procedures with non-college-student samples (i.e., middle-aged and older adults). During the acquisition phase, they participated in two sessions: one with the repeated study procedure and one with the repeated testing procedure. For each face, an occupation label was presented as a contextual cue at both study and test to avoid floor performance. The participants were instructed to study the associ-

[1] In Karpicke and Roediger's (2008) repeated testing condition, their college student participants rarely produced intra-list intrusion errors during the acquisition phase (J. D. Karpicke, personal communication, June 2008). Given their proficiency in encoding Swahili–English pairs during the acquisition phase, these participants rarely needed to discriminate the correct English translation from their wrong answer to a Swahili cue in the delayed final recall test.

ation between the face and first name because they would later be asked to recall the first name corresponding to each face on the test.

There were four study–test cycles in each session during the acquisition phase (see Figures 1 and 2 for simplified versions in which only the first three cycles and three face–name pairs are illustrated). As shown in Figure 1, when participants correctly recalled a name in one cycle (e.g., Jacob), the face–name pair was either removed from further testing but still studied in subsequent cycles (i.e., the repeated study session) or removed from further study but still tested in subsequent cycles (i.e., the repeated testing session). In the repeated study session, when participants wrongly recalled a name in one cycle (e.g., Nicole and Aaron being recalled as Vivien and Dave, respectively), the face–name pair was studied and tested again in the immediately subsequent cycle. In the repeated testing session, only face–name pairs that were never correctly recalled (i.e., Nicole and Aaron in Figure 1) were studied and tested again. However, as depicted in Figure 2, once a pair was correctly recalled (e.g., Jacob), it would be dropped out from further study in the repeated testing condition, and hence participants might consistently yet incorrectly recall the name in subsequent cycles (e.g., Jacob being recalled as Patrick).

At the end of each acquisition phase, participants were asked to predict their performance on recalling the face–name pairs 1.5 hr later. This judgment allowed us to examine whether participants' actual final recall performance was correlated with their predicted final recall performance. On the basis of pilot data, we shortened the retention interval from 1 week in Karpicke and Roediger (2008) to 1.5 hr to avoid floor effects. After the acquisition phase, the participants were tested with other unrelated tasks (e.g., Stroop and lexical decision tasks) during the 1.5-hr retention interval and then took a final recall test for all of the 16 face–name pairs they learned in both sessions, which were randomly interleaved.

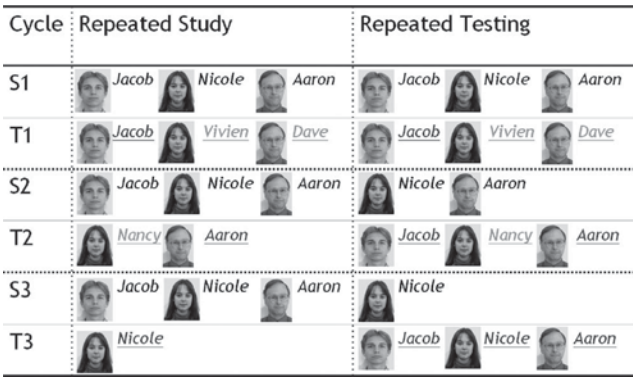


*Figure 1.* Examples illustrating the repeated study and repeated testing procedures. S = study trials; T = test trials. The numbers next to S and T indicate the cycle number. In the S rows, the correct name appears next to the face. In the T rows, underlined names in black indicate correct responses and underlined names in light gray indicate incorrect responses. Only three face–name pairs (Jacob, Nicole, and Aaron) are shown here as examples. This is a simplified version in which only the first three cycles and three face–name pairs are illustrated. In the actual experiment, four study–test cycles and eight face–name pairs were used.

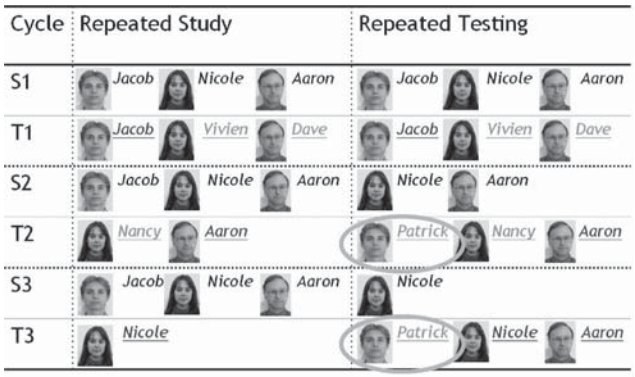


*Figure 2.* Examples illustrating the occurrence of intra-list intrusion errors. S = study trials; T = test trials. The numbers next to S and T indicate the cycle number. In the S rows, the correct name appears next to the face. In the T rows, underlined names in black indicate correct responses and underlined names in light grey indicate incorrect responses. Only three face–name pairs (Jacob, Nicole, and Aaron) are shown here as examples. This is a simplified version in which only the first three cycles and three face–name pairs are illustrated. In the actual experiment, four study–test cycles and eight face–name pairs were used.

## Experiment 1

### Method

**Participants.** Ninety-six healthy middle-aged and older adults (ages 46–95) were recruited from Washington University Alzheimer's Disease Research Center and paid for their participation. They did not meet criteria for probable Alzheimer's disease of the National Institute of Neurological and Communications Disorders and Stroke–Alzheimer's Disease and Related Disorders Association (McKhann et al., 1984). On the basis of the Clinical Dementia Rating (CDR; Morris, 1993), these individuals were all rated as nondemented (i.e., CDR 0). The CDR is based on a 90-min clinical interview that directly assesses the participant's cognitive status and also relies on information from a close collateral source. The reliability of the CDR (Burke et al., 1988) and the validity of the diagnosis based on autopsy by the clinicians and research scientists at Washington University in St. Louis have been excellent (93% diagnostic accuracy; Berg et al., 1998). The participants were also screened for depression, untreated hypertension, reversible dementias, and other disorders that can potentially produce cognitive impairment, so they are likely to be cognitive healthy individuals (see Table 1). This study was approved by the Institutional Review Board at Washington University School of Medicine, and all participants provided informed consent at the beginning of the study.

**Apparatus, materials, and design.** The experiment was programmed using E-Prime (Version 1.0) and ran on a Dell desktop computer with a standard 15-in. (38.1-cm) monitor. Two sets of eight color photographs of faces were chosen from the Psychological Image Collection at Stirling (n.d.). Each set consists of two young male, two young female, two old male, and two old female faces. Eight male and eight female names were randomly assigned to each of the eight male and eight female faces, respectively. Sixteen occupation labels were randomly assigned to the 16 face–

Table 1
*Mean and Standard Deviation in Psychometric Test Performance*

| | Experiment 1 (without feedback) | | | | | | | | Experiment 2 (with feedback) | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Middle-aged | | Young-old | | Old-old | | Overall[a] | | Middle-aged | | Young-old | | Old-old | | Overall[a] | |
| Variable | *M* | *SD* | *M* | *SD* | *M* | *SD* | *M* | *SD* | *M* | *SD* | *M* | *SD* | *M* | *SD* | *M* | *SD* |
| Age | 60.81 | 6.46 | 71.97 | 2.09 | 81.28 | 4.99 | 71.35* | 9.69 | 57.53 | 4.61 | 68.36 | 2.37 | 80.33 | 4.50 | 68.75* | 10.29 |
| Mini-Mental State Exam | 29.53 | 0.63 | 28.73 | 1.31 | 28.88 | 1.24 | 29.07* | 1.13 | 29.50 | 0.65 | 29.40 | 1.26 | 28.69 | 1.03 | 29.19 | 1.02 |
| Forward Digit Span | 6.33 | 1.19 | 6.28 | 1.17 | 6.92 | 1.02 | 6.54 | 1.15 | 7.56 | 0.73 | 6.80 | 1.79 | 6.69 | 1.03 | 7.00 | 1.14 |
| Backward Digit Span | 4.94 | 1.21 | 4.04 | 1.10 | 4.85 | 1.16 | 4.58* | 1.21 | 5.56 | 1.24 | 4.60 | 1.14 | 4.77 | 1.24 | 5.00 | 1.24 |
| Logical Memory | 13.00 | 4.11 | 11.92 | 3.17 | 11.92 | 4.23 | 12.20 | 3.82 | 12.39 | 2.52 | 13.60 | 2.41 | 13.08 | 4.19 | 12.94 | 3.34 |
| Associate Recall | 14.94 | 3.70 | 13.50 | 3.39 | 13.44 | 3.55 | 13.86 | 3.54 | 17.06 | 1.94 | 16.90 | 2.53 | 13.73 | 3.30 | 15.43* | 3.16 |
| Selective Reminding Free Recall | 32.77 | 5.93 | 29.65 | 5.99 | 27.27 | 5.12 | 30.04* | 6.08 | 34.71 | 3.56 | 34.00 | 4.47 | 30.62 | 7.49 | 33.08 | 5.63 |
| WAIS Information | 21.00 | 5.16 | 20.12 | 5.18 | 20.50 | 4.52 | 20.49 | 4.87 | 21.00 | 2.65 | 23.20 | 2.95 | 22.08 | 3.99 | 21.93 | 3.38 |
| WAIS Block Design | 36.06 | 7.75 | 27.88 | 8.39 | 30.08 | 8.11 | 30.88* | 8.64 | 37.33 | 7.55 | 30.60 | 5.27 | 31.85 | 7.12 | 33.44 | 7.30 |
| Animal Naming | 23.50 | 4.95 | 20.88 | 5.02 | 16.96 | 3.99 | 20.60* | 5.37 | 25.57 | 4.48 | 22.60 | 4.58 | 20.46 | 5.30 | 22.97* | 5.18 |
| Boston Naming | 57.56 | 2.96 | 53.80 | 7.23 | 53.88 | 4.76 | 54.81 | 5.63 | 57.56 | 2.65 | 56.80 | 1.10 | 54.92 | 3.95 | 56.15 | 3.32 |
| Reading Span | 7.83 | 1.85 | 6.92 | 1.62 | 7.12 | 1.48 | 7.31 | 1.69 | 8.79 | 3.02 | 7.91 | 1.38 | 6.85 | 1.34 | 7.87 | 2.23 |
| Rotation Span | 9.67 | 3.02 | 7.77 | 2.64 | 7.72 | 3.02 | 8.46* | 3.02 | 11.93 | 3.20 | 8.90 | 4.36 | 8.18 | 3.03 | 9.89* | 3.82 |
| Computation Span | 8.44 | 2.79 | 7.19 | 2.98 | 6.77 | 2.98 | 7.48 | 2.97 | 12.23 | 5.59 | 9.45 | 4.32 | 7.92 | 2.64 | 9.94 | 4.65 |
| Word Fluency S-P | 31.94 | 9.23 | 26.36 | 9.75 | 30.42 | 12.47 | 29.35 | 10.84 | 31.11 | 10.06 | 41.40 | 17.90 | 30.54 | 8.88 | 32.74 | 11.60 |
| WAIS Digit Symbol | 57.94 | 11.09 | 47.52 | 10.67 | 45.00 | 11.10 | 49.16* | 11.97 | 57.11 | 10.23 | 47.60 | 5.94 | 46.69 | 9.12 | 50.33* | 10.00 |
| Crossing Off | 177.83 | 32.27 | 159.40 | 25.56 | 150.96 | 39.83 | 161.03* | 34.48 | 173.33 | 26.36 | 204.20 | 44.32 | 147.33 | 21.57 | 167.27* | 34.90 |
| Trail Making A | 29.07 | 9.06 | 39.42 | 13.56 | 37.04 | 10.36 | 34.88* | 11.83 | 27.79 | 5.22 | 28.30 | 10.25 | 33.62 | 12.05 | 29.97 | 9.59 |
| Trail Making B | 63.17 | 23.03 | 98.42 | 35.67 | 94.12 | 32.38 | 84.16* | 34.16 | 65.57 | 20.91 | 72.60 | 25.43 | 84.77 | 36.26 | 74.22 | 28.77 |

*Note.* The references for these tasks are as follows: Mini-Mental State Exam (Folstein, Folstein, & McHugh, 1975); Forward and Backward Digit Span, Logical Memory, and Associate Recall subtests from the Wechsler Memory Scale (Wechsler, 1987); Selective Reminding Free Recall (Grober et al., 1988); Block Design, Information, and Digit Symbol subtests of Wechsler Adult Intelligence Scale (WAIS; Wechsler, 1997); Boston Naming and Animal Naming (Goodglass & Kaplan, 1983); three working memory span tasks (Reading, Rotation, and Computation Span; see Engle, Tuholski, Laughlin, & Conway, 1999); Word Fluency Test S-P (Thurstone & Thurstone, 1949); Crossing Off (Botwinick & Storandt, 1973); Trail Making A and B (Armitage, 1945). All of the psychometric and span tasks are scored such that higher scores indicate better performance, except the Trail Making A and B, where higher scores indicate poorer performance. The number of participants who received the 2-hr battery of psychometric tests (within a 1-year window of the current study) ranged from 95 to 119. The findings of this subset of participants were similar to those of the whole sample reported in the text.
[a] An asterisk indicates that differences between the three age groups were significant.
* $p < .05$.

name pairs. Assignment of stimuli to the repeated study and repeated testing conditions, as well as the presentation order of repeated study and repeated testing sessions, was counterbalanced across participants. All names and occupation labels are listed in the Appendix, and all face pictures are available upon request.

**Procedure.** Participants took the test individually while seated comfortably in front of a computer monitor. They completed two sessions (repeated study and repeated testing) during the acquisition phase and a final test 1.5 hr later, in which all face–name pairs were tested. During the acquisition phase, participants learned the list of face–name pairs across a total of four study–test cycles. The first study–test cycle consisted of eight study trials followed by eight test trials. On each study trial, participants saw a face (on the left) along with its corresponding first and last names and occupation label (on the right) on a black background for 8 s. The first–last name and occupation label, colored in yellow, were presented: "My name is [first name] [last name]. I am a [occupation label]." The participants were told to study the first name for each face so they could recall the first name in response to the face for a later memory test. The eight pairs were randomly presented on each cycle. To familiarize the participants with the task, they completed a sample study–test cycle with two study and test trials at the start of repeated study and repeated testing sessions. The experimenter made sure the participants understood the task before proceeding to the actual study–test cycles. A 500-ms blank screen was inserted between study trials.

After the eight trials on each study cycle, participants saw a series of arithmetic equations on the monitor to which they were to respond "correct" or "incorrect." They then started the test cycle, which consisted of eight or fewer test trials, depending on the condition. On each test trial, participants saw a face (on the left) and a description, "I am _____. I am a [occupational label]" (on the right). The participants were instructed to recall aloud the first name corresponding to the face or occupation, and the experimenter typed in the answer on the keyboard. The display stayed on the screen until the participants responded, after which the next test trial appeared. The participants were allowed to pass if they failed to recall any name and an "omission" response was coded. Thus, participants were not encouraged to guess (or not guess) if they could not recall the correct name corresponding to its face. Immediately after all test trials were presented, another study–test cycle began.

In the repeated study session, participants across study–test cycles were tested only on pairs that they had not yet recalled on the previous cycle until they reached the point where all names had been recalled. That is, all eight face–name pairs were studied repeatedly, but the acquisition phase ended immediately after

participants correctly recalled all eight face–name pairs. In the repeated testing session, across study–test cycles participants were consistently retested on all face–name pairs but only received an opportunity to restudy those pairs that were not recalled in the previous cycle. Hence, in the repeated testing condition, when a face–name pair was correctly recalled and then subsequently incorrectly recalled, participants would not receive the correct pair again in a subsequent study cycle. All eight face–name pairs were tested in all four cycles even after they had been correctly recalled. Hence, in the repeated study session, participants recalled each face–name pair only once during the acquisition phase, whereas in the repeated testing session, they repeatedly attempted to recall the face–name pairs during every study–test cycle (see Figure 1 for a simplified version in which only the first three cycles and three face–name pairs are illustrated). In the repeated study session, the face–name pair could be correctly recalled once at most, whereas in the repeated testing session, it could be correctly recalled more than once; therefore, participants correctly recalled the face–name pairs more frequently in the repeated testing session than in the repeated study session.

At the end of the acquisition phase of both the repeated study and the repeated testing conditions, we asked participants to predict how many of the eight just-learned face–name pairs they would recall on a final recall test after 1.5 hr. Participants then performed a set of unrelated tasks. Participants then took the final recall test, in which the procedure was identical to the test trial during the acquisition phase, except that they were now tested for the randomly ordered 16 face–name pairs learned in both repeated study and repeated testing sessions.

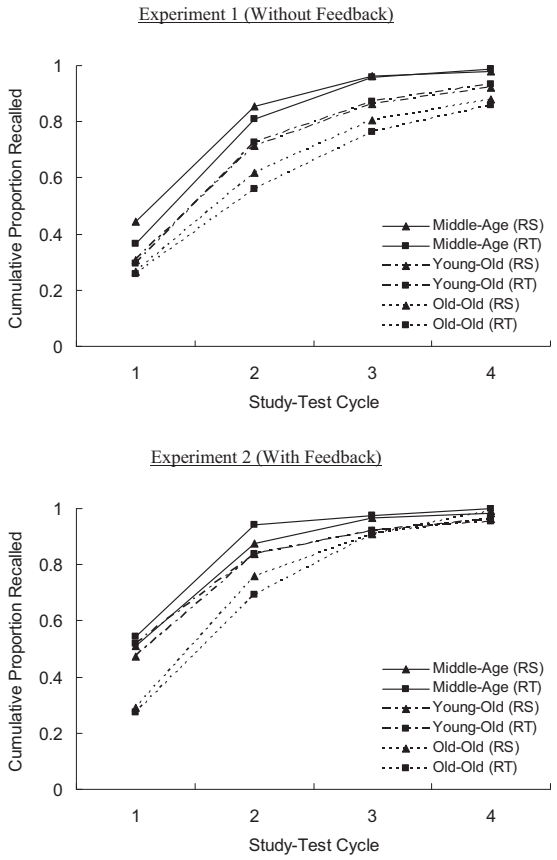


*Figure 3.* Mean proportion of face–name pairs recalled in repeated study and repeated testing sessions across the four study–test cycles in the acquisition phase. RS = repeated study; RT = repeated testing.

## Results

Unless otherwise specified, the significance level was set at .05. The effect sizes of $F$ and $t$ are represented by $\eta_p^2$ and Cohen's $d$, respectively. Because the order in which participants received the repeated study and repeated testing sessions did not interact with any factor in the following analyses (all $F$s $< 1$), we do not consider this factor any further.

Three sets of analyses were conducted to examine how participants' age modulates the acquisition of face–name pairs and the effectiveness of repeated testing on enhancing recall. First, we divided the participants into three equal groups based on their age (see Table 1) and conducted repeated-measures analyses of variance (ANOVAs) on their acquisition and final recall performance; age was a between-subjects variable and condition (repeated study vs. repeated testing) was a within-subjects variable. Second, we conducted correlational analyses to confirm the ANOVA findings with age as a continuous variable. Third, we conducted item analyses on the final recall performance, conditioned on whether a face–name pair had yielded intra-list intrusion errors, to address the influence of these errors on long-term retention.

**Acquisition phase.** Figure 3 (top panel) shows the cumulative proportion of face–name pairs recalled during the acquisition phase as a function of group and test cycle. As shown, recall generally declined with participants' age and increased over study–test cycles. A 3 (age) × 2 (condition) × 4 (cycle) mixed factor ANOVA on these data, using a Greenhouse-Geisser correction for the potential violation of sphericity (see Table 2 for the statistics), revealed main effects of age and cycle, with a marginally significant Age × Cycle interaction indicating that the age-related difference on proportion recall of face–name pairs became smaller as a function of increases in the number of study–test cycles. This interaction likely reflected the fact that old-old adults acquired the face–name pairs more slowly than young-old adults, but all participants approached a common ceiling. More importantly, there was no difference in cumulative recall proportion between repeated study and repeated testing sessions, nor did this variable interact with age (cf. Karpicke & Roediger, 2007, 2008).

In the repeated testing session, participants received the test trials with the correct face–name pairs dropped out from the study list across all four cycles, even after they correctly recalled each of the eight face–name pairs. Thus, the maximum number of study and test trials was 64: 8 (per cycle) × 4 (cycles) × 2 (study and test trials). In contrast, the acquisition phase in the repeated study session ended immediately after participants reached 100% accuracy in the face–name acquisition. Given that some of the participants did not go through all four study–test cycles in the repeated study session (i.e., the maximum number of study and test trials was less than 64), the mean number of study and test trials was lower for the repeated study condition (33.7 middle-aged, 42.3 young-old, and 45.5 old-old) than for the repeated testing condition (46.8 middle-

Table 2
*Analyses of Performance in the Acquisition Phase, Number of Study–Test Trials, and Proportion of Final Recall in Experiments 1 and 2*

| Effect | Experiment 1 (without feedback) | | | | | Experiment 2 (with feedback) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *df* | *F* | *MSE* | *p* | $\eta_p^2$ | *df* | *F* | *MSE* | *p* | $\eta_p^2$ |
| Acquisition phase | | | | | | | | | | |
| Cycle | 1.88, 174.91 | **570.11** | 0.04 | <.01 | .86 | 1.90, 78.07 | **201.24** | 0.05 | <.01 | .82 |
| Age | 1, 93 | **8.28** | 0.22 | <.01 | .15 | 1, 41 | **3.84** | 0.13 | <.01 | .16 |
| Condition | 1, 93 | 1.15 | 0.06 | .29 | .01 | 1, 41 | 0.05 | 0.03 | .82 | .001 |
| Age × Condition | 2, 93 | 0.47 | 0.06 | .63 | .01 | 2, 41 | 0.90 | 0.03 | .41 | .04 |
| Age × Cycle | 3.76, 174.91 | 2.25 | 0.04 | .07 | .05 | 3.81, 78.07 | **4.23** | 0.04 | <.01 | .17 |
| Cycle × Condition | 2.12, 196.95 | 0.89 | 0.02 | .42 | .01 | 1.85, 76.02 | 0.27 | 0.02 | .75 | .01 |
| Age × Condition × Cycle | 4.24, 196.95 | 0.88 | 0.02 | .48 | .02 | 3.71, 76.02 | 0.68 | 0.02 | .60 | .03 |
| Number of study–test trials | | | | | | | | | | |
| Age | 2, 93 | **9.91** | 111.07 | <.01 | .18 | 2, 41 | 2.88 | 93.41 | .07 | .12 |
| Condition | 1, 93 | **81.46** | 43.84 | <.01 | .47 | 1, 41 | **44.70** | 36.63 | <.01 | .52 |
| Age × Condition | 2, 93 | **5.69** | 43.84 | <.01 | .11 | 2, 41 | 0.37 | 36.63 | .69 | .02 |
| Proportion of final recall | | | | | | | | | | |
| Age | 2, 92 | **3.50** | 0.09 | .03 | .07 | 2, 40 | **4.57** | 0.06 | .02 | .19 |
| Condition | 1, 92 | 2.18 | 0.03 | .14 | .07 | 1, 40 | **13.08** | 0.03 | <.01 | .25 |
| Age × Condition | 2, 92 | **6.29** | 0.03 | <.01 | .12 | 2, 40 | 0.96 | 0.03 | .39 | .05 |

*Note.* The analyses for the proportion of final recall were done with the number of study–test trials being partialled out. Bold values are statistically significant.

aged, 49.3 young-old, and 51.3 old-old). The 3 (age) × 2 (condition) ANOVA yielded significant main effects of age and condition and Age × Condition interaction. The difference in the number of study–test trials between the repeated study and repeated testing sessions decreased as a function of age (13.1 middle-aged, 7.0 young-old, and 5.8 old-old).[2] However, because some participants might reach 100% accuracy in the repeated testing session prior to the end of four study–test cycles, it is important to examine whether they still took more trials to *first* acquire the face–name pairs (i.e., reaching 100% accuracy) in the repeated testing session than in the repeated study session. The mean number of study and test trials that participants took to first reach 100% accuracy in the repeated testing session was 37.3 (middle-aged), 41.8 (young-old), and 47.0 (old-old). Although middle-aged adults still took more trials to first reach 100% accuracy in the repeated testing session than in the repeated study session, $t(31) = 2.12$, young-old and old-old adults did not take differential numbers of trials to first reach 100% accuracy in repeated testing and repeated study sessions (both $t$s $< 1.00$), indicating that repeated study did not produce faster and more efficient learning than repeated testing. Nevertheless, as the participants did receive overall more trials in the repeated testing session than in the repeated study session, the difference in the number of study–test trials for repeated study versus repeated testing sessions was controlled (i.e., treated as a covariate) in all of the following analyses.[3]

**Final recall phase.** Figure 4 (upper left panel) displays the proportion of face–name pairs correctly recalled in the final recall phase. A 3 (age) × 2 (condition) ANOVA yielded a significant main effect of age and an Age × Condition interaction (see Table 2 for the statistics), indicating that middle-aged adults showed the predicted testing effect of better performance in repeated testing versus repeated study session (.76 vs. .68), $t(31) = 2.02$, $p = .052$, $d = .51$; young-old adults showed a null effect (.55 vs. .58), $t(31) = .57$, $d = .14$; and old-old adults showed a nearly significant reversal of the testing effect (i.e., a benefit from repeated study; .43 vs. .53), $t(31) = 1.88$, $p = .07$, $d = .48$. Because the acquisition was at or near ceiling by the end of four study–test cycles, the analyses on the proportion of face–name pairs forgotten in the final recall phase mirrored the final recall data. These analyses are available upon request.

**Correlational analyses.** Figure 5 displays final recall as a function of age and encoding condition. As shown, overall final recall performance decreases as a function of age in both the repeated study and repeated testing sessions, as supported by the

[2] One might be puzzled by the finding that the number of study–test trials interacted with condition (repeated study vs. repeated testing), whereas cumulative recall proportion did not. It is noteworthy that in the repeated study session when participants attained 100% accuracy in the third cycle, the cumulative recall proportion remained nominally 100% in the fourth cycle, even though they never studied items in that cycle because the acquisition phase ended immediately after the participants attained 100% accuracy. This fact explains why the overall number of study–test trials differed in repeated study and repeated testing sessions, even though the cumulative recall proportion did not (see Figure 3 and text for further discussion).

[3] One could argue that the total number of study–test trials might not be an accurate measure for how much participants learned in the acquisition phase because it includes both correctly and incorrectly recalled test trials. However, when we counted only the study trials and correct test trials, we found that all age groups received at least numerically more adjusted study–test trials in repeated testing sessions than in repeated study sessions. We conducted additional analyses in which we treated this adjusted number instead of the raw number of study–test trials as a covariate in all of our reported analyses and replicated the same overall pattern of results.

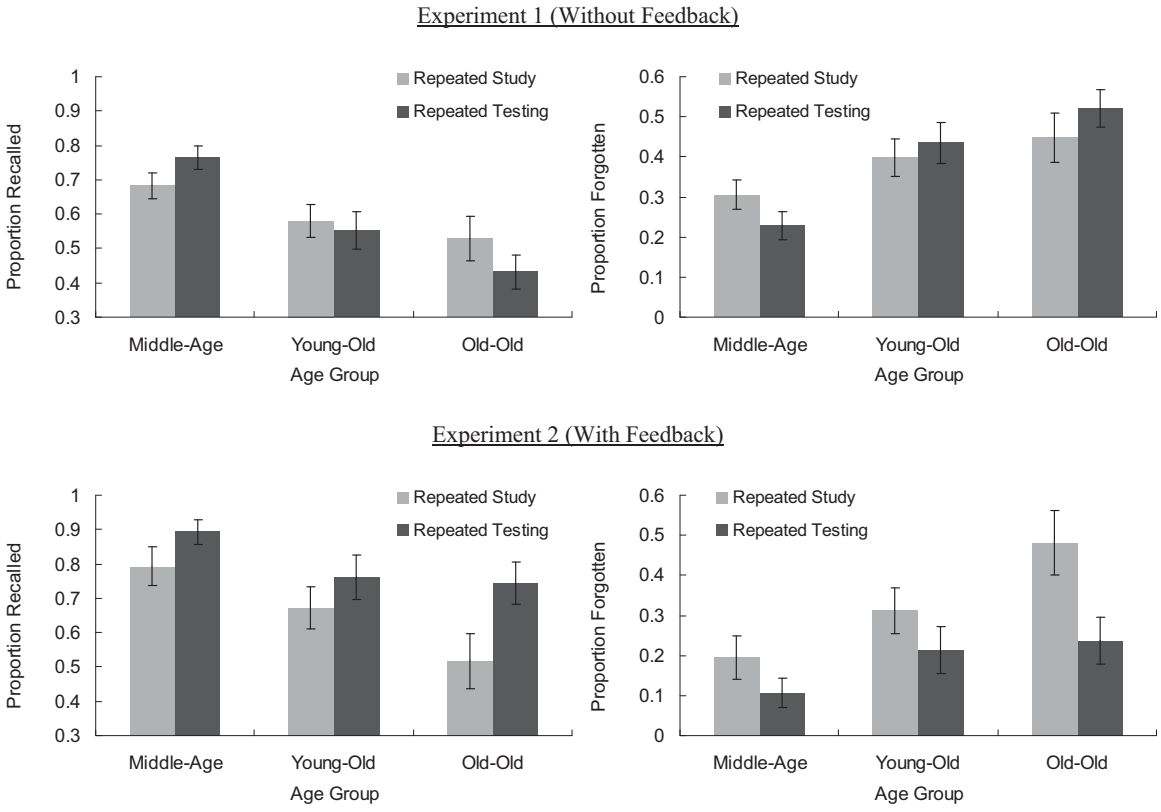


*Figure 4.* Mean proportion of face–name pairs recalled and proportion of face–name pairs forgotten (i.e., the difference between the proportion of pairs acquired at the end of acquisition phase and the proportion of pairs recalled in the final test) in repeated study and repeated testing sessions in the final-test phase.

significant Pearson correlation coefficients between final recall and age. More importantly, the age-related decrease appears to be steeper for the repeated testing session than for the repeated study session. Converging with the results from the ANOVAs, there was a reliable correlation ($r = -.25$, $p = .02$) between the difference of repeated study versus repeated testing sessions (i.e., the testing effect) and the participants' age.

**Intrusion error analyses.** Next, we investigated how the final recall performance could be influenced when the face–name pair had yielded intrusion errors in the acquisition phase. More specifically, in the repeated testing session, because a given face–name pair that had been correctly recalled in an earlier cycle was dropped from further study in the later cycle, older adults might confuse one face–name pair with another and produce an intrusion error in the subsequent test trial. Because participants in the repeated testing session do not restudy such items after they are correctly recalled once, they would not have the opportunity to correct the error. If older adults are deficient in recollection and/or attentional control, they might fail to discriminate their own intrusion errors from the correct answer, thereby reducing their final recall performance in the repeated testing session. This problem would be minimized in the repeated study session, because once a given pair is correctly recalled it is dropped from the test trials in the subsequent cycle.

To test the influence of intrusion errors in the repeated testing and repeated study sessions on final recall, we conducted item analyses by conditionalizing the proportion of final recall on each face–name pair by the number of intra-list and extra-list intrusion errors that participants made during the acquisition. Table 3 presents the cell means of the intrusion errors. Although age did not reliably modulate the intrusion errors, we did find evidence of such a relationship in the more powerful correlational analyses. The correlation between participants' age and final recall performance when they made one or more intra-list intrusion errors during acquisition was significant in the repeated testing session ($-.29$, $p < .01$), but not in the repeated study session ($-.09$). Hence, despite making intrusion errors equally often relative to the young-old adults, old-old adults were affected by their prior intra-list intrusion errors in the final recall test when they acquired the face–name pairs via a repeated testing procedure but not when they did so via a repeated study procedure. The correlation between participants' age and final recall performance when they made one or more extra-list intrusion errors in the acquisition phase was not significant in the repeated study session ($-.13$) or in the repeated testing session ($-.01$). Hence, the correlational analyses are consistent with the notion that part of the age-related reversal in the testing effect can be attributed to participants being confused by their prior intra-list intrusion errors.

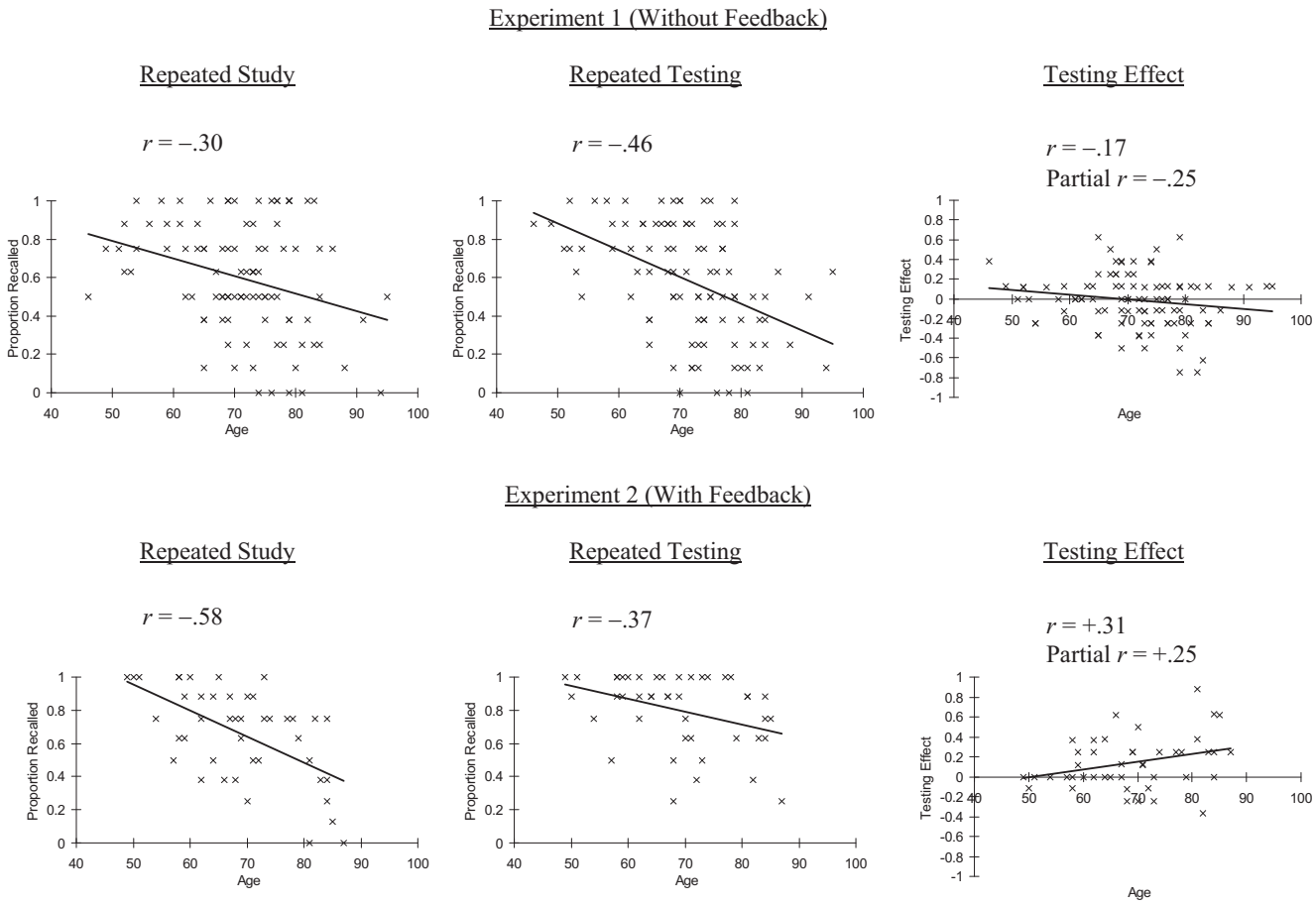


*Figure 5.* The scatter plots for the mean proportion of face–name pairs recalled in the repeated study session (left panels), repeated testing session (middle panels), and the mean difference in the proportion of face–name pairs between the repeated study and repeated testing sessions as a function of age in the final-test phase (right panels). The partial *r*s were based on analyses that controlled for the number of study–test trials in the acquisition phase.

In sum, the results from Experiment 1 indicate that middle-aged adults produced the expected testing effect, but the old-old adults showed no testing effect and even an advantage of repeated study. As participants' age increases, their episodic memory loss makes them increasing susceptible to intra-list intrusion errors during acquisition. Because they did not have opportunities to restudy the items in the repeated testing condition (after they correctly recalled a pair once), they were more likely to strengthen an erroneous face–name association during the acquisition phase. Hence, it may be critical for participants to have an opportunity to correct any errors.

Table 3
*Mean Proportions of Omission, Extra-List Intrusion Errors and Intra-List Intrusion Errors in the Acquisition Phase in the Repeated Study and Repeated Testing Sessions as a Function of Age*

| | Experiment 1 (without feedback) | | | Experiment 2 (with feedback) | | |
|---|---|---|---|---|---|---|
| Error type | Middle-aged | Young-old | Old-old | Middle-aged | Young-old | Old-old |
| Repeated study | | | | | | |
| Omission | 0.19 | 0.24 | 0.31 | 0.15 | 0.09 | 0.16 |
| Extra-list intrusion | 0.07 | 0.12 | 0.08 | 0.09 | 0.09 | 0.08 |
| Intra-list intrusion | 0.12 | 0.13 | 0.17 | 0.10 | 0.22 | 0.24 |
| Repeated testing | | | | | | |
| Omission | 0.10 | 0.21 | 0.28 | 0.07 | 0.03 | 0.14 |
| Extra-list intrusion | 0.10 | 0.08 | 0.09 | 0.03 | 0.07 | 0.07 |
| Intra-list intrusion | 0.10 | 0.13 | 0.15 | 0.05 | 0.16 | 0.13 |

If the reversal of the testing effect in the older adults was due to the persistence of errors from acquisition into the final recall phase, then one might expect the typical testing advantage to return when older adults are given correct feedback during the acquisition phase. Experiment 2 directly tested this possibility by providing feedback after all responses during the acquisition phase. This is the only procedural difference between Experiments 1 and 2. The feedback consisted of two parts. Immediately after the participant's response on each test trial, a *correct* or *incorrect* verbal signal was visually presented at the center of the screen. Regardless of whether participants' responses were correct or incorrect, the correct face–name pair was then displayed identically to a normal study trial (i.e., a face and a description, "I am _____. I am a [occupational label]").

## Experiment 2

### Method

Forty-four healthy middle-aged and older adults were recruited from the same participant pool as in Experiment 1 (see Table 1). All instructions, materials, design and procedures were identical to those in Experiment 1, except that participants received feedback immediately after their responses in the acquisition and final recall phases. The *correct* or *incorrect* signal was first presented for 2 s; then, regardless of whether participants' responses were correct or incorrect, the correct face–name pair was displayed for 5 s.

### Results

The analytic procedure was the same as in Experiment 1.

**Acquisition phase.** As shown in the lower portion of Figure 3 and consistent with the statistics summarized in Table 2, the overall pattern was similar to those obtained in Experiment 1, with main effects of both age (older adults learned more slowly) and study–test cycle. The reliable Cycle × Age interaction showed that the age-related difference on proportion of correct recall of face–name pairs again reliably decreased across the study–test cycles and is likely due to ceiling effects at the final study–test cycle. There was also no difference in cumulative recall proportion between repeated study and repeated testing sessions, nor was this affected by participants' ages.

As in Experiment 1, the mean number of study–test trials to reach criteria of 100% accuracy was lower for repeated study sessions, 35.0 (middle-aged), 36.4 (young-old), and 41.9 (old-old), than for repeated testing sessions, 44.4 (middle-aged), 45.9 (young-old), and 49.0 (old-old), as reflected by a significant main effect of condition. We again examined whether participants took more trials to *first* acquire the face–name pairs (i.e., reaching 100% accuracy) in the repeated testing session than in the repeated study session. The mean number of study and test trials when participants first reached 100% accuracy in the repeated testing session was 31.1 (middle-aged), 36.7 (young-old), and 41.5 (old-old), which did not differ significantly from the corresponding number of study–test trials in the repeated study session (all $t$s $< 1.64$). This shows that repeated study did not produce faster and more efficient learning than repeated testing. As in Experiment 1, the difference in the overall number of study–test trials in repeated study versus repeated testing sessions was treated as a covariate in all of the following analyses.

**Final recall phase.** Final recall data are shown in the lower left panel of Figure 4. Importantly, there was a clear testing effect for all three age groups. The 3 (age) × 2 (condition) ANOVA yielded significant main effects of age and condition, but not an Age × Condition interaction (see Table 2). The middle-aged adults showed better performance than the young-old adults, who in turn showed better overall performance than the old-old adults (.84 vs. .72 vs. .63). Unlike Experiment 1, in this study all participants produced a testing effect (.80 vs. .66), confirming the predicted benefit of feedback. Although the Age × Condition interaction was not significant, the pattern showed a greater difference in the testing effect with age (i.e., repeated testing produced greater final recall than repeated study, and this effect was numerically larger in older adults relative to middle-aged adults). As shown below, this trend was also validated in the correlational analyses.

**Correlational analyses and intrusion error analyses.** As shown in the scatter plots in Figure 5, there was a general decrease in final recall performance as age increased in both repeated study and repeated testing sessions, as supported by the significant Pearson correlation coefficients between final recall and age. The age-related decrease was shallower for the repeated testing session than for the repeated study session. The correlation between the testing effect and participants' age was positive ($r = +.25$, $p = .11$), rather than negative as in Experiment 1, confirming the trend observed in Figure 4. The lack of significance is likely due to less statistical power in this experiment. Thus, with the aid of feedback, older adults now benefited more from repeated testing, relative to repeated study. We again performed item analyses by conditionalizing the final recall performance for each face–name pair on the number of intra-list and extra-list intrusion errors that participants made during acquisition for that item. In contrast to Experiment 1, the correlation between participants' age and final recall performance when they made one or more intrusion errors in the acquisition phase was not significant in the repeated study or repeated testing session for intra-list (−.12 and –.14) and extra-list intrusion errors (−.07 and –.01). These findings suggest that the presence of feedback decreased the influence of the intrusion errors that contributed to the reversal of the testing effect in the old-old individuals observed in Experiment 1.

**Combined analyses for Experiments 1 and 2.** We focus only on the main effect and interactions associated with feedback here. As expected, in the acquisition phase there was a main effect of feedback on the proportion of cumulative recall, $F(1, 134) = 8.82$, $MSE = 0.19$, $\eta_p^2 = .06$, and on the number of study–test trials (42.1 vs. 44.8), $F(1, 134) = 4.22$, $MSE = 105.66$, $\eta_p^2 = .12$ (see Figure 3), with no reliable interactions associated with feedback. In the final recall phase (see Figure 4, left panels), there was a main effect of feedback, $F(1, 133) = 15.35$, $MSE = 0.08$, $\eta_p^2 = .10$; a Condition × Feedback interaction, $F(1, 133) = 10.69$, $MSE = 0.03$, $\eta_p^2 = .07$; and most importantly, an Age × Condition × Feedback interaction, $F(2, 133) = 4.61$, $MSE = 0.03$, $\eta_p^2 = .07$. As predicted, follow-up analyses on the three-way interaction yielded a highly reliable Condition × Feedback interaction for old-old adults, $F(1, 44) = 15.01$, $MSE = 0.04$, $\eta_p^2 = .25$, but not for

middle-aged, $F(1, 44) = .09$, $MSE = 0.02$, $\eta_p^2 = .002$, or young-old adults, $F(1, 43) = 1.87$, $MSE = 0.04$, $\eta_p^2 = .04$.[4]

**Metamemory predictions.** We examined participants' actual versus predicted performance as a function of repeated study and repeated testing as a function of feedback. Figure 6 displays the mean proportions of face–name pairs that participants predicted they would recall in the final recall test as a function of condition. In general, in Experiment 1 (i.e., without feedback) participants did not vary their predictions as a function of repeated study and repeated testing. However, when they received feedback (i.e., Experiment 2), they accurately predicted that repeated testing would lead to superior retention later. We submitted participants' predicted and actual final recall to a 3 (age) × 2 (condition) × 2 (feedback) × 2 (actual vs. predicted performance) ANOVA. The significant main effect of measure, $F(1, 133) = 30.15$, $MSE = 0.05$, $\eta_p^2 = .19$, indicated that the participants generally underestimated their final recall performance (predicted = .42 vs. actual = .63). The difference was not modulated by other variables (all $F$s $< 1.87$). Interestingly, the Condition × Feedback interaction was significant, $F(1, 133) = 17.73$, $MSE = 0.02$, $\eta_p^2 = .12$. Regardless of participants' age, when they received the feedback during the acquisition, they recognized the benefit of repeated testing (.59), relative to repeated study (.47), $t(43) = 3.51$, $d = .75$. However, when they did not receive feedback, they predicted similar performance for repeated testing (.35) and repeated study sessions (.38), $t(95) = 1.63$, $d = .24$, replicating Karpicke and Roediger's (2008) findings with young adults.

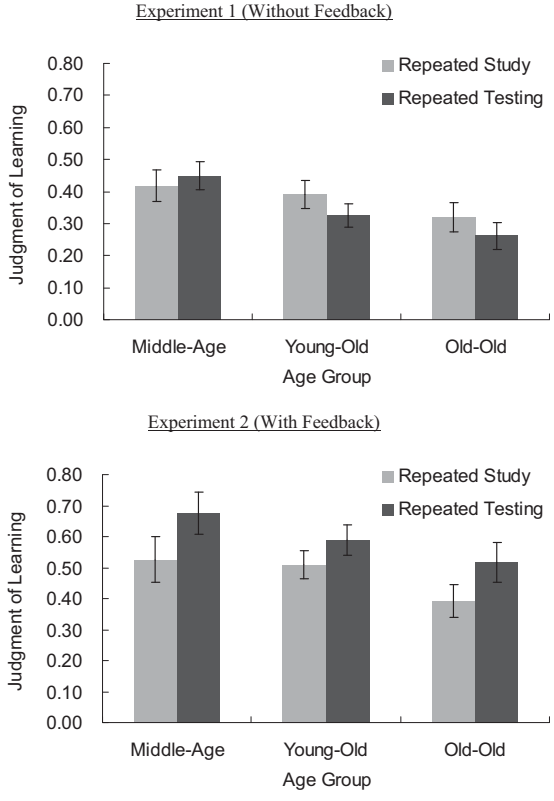


*Figure 6.* Mean predicted proportion of face–name pairs recalled in repeated study and repeated testing sessions in the final-test phase for the three age groups.

## General Discussion

Using a variant of Karpicke and Roediger's (2008) repeated study and repeated testing procedures, we replicated the pattern they obtained with young adults in both experiments with our middle-aged adults; that is, better delayed recall performance in the repeated testing condition than in the repeated study condition. This suggests that Karpicke and Roediger's results generalize for conditions in which (a) face–name pairs were used as study materials (see also Carpenter & DeLosh, 2005), (b) repeated study and repeated testing sessions were manipulated within participants, (c) the number of study items was sharply reduced from 40 to eight, (d) the retention interval was substantially shortened from 1 week to 1.5 hr, and (e) a group of middle-aged adults who were not college students was tested. However, the same pattern was not observed for old-old adults, who produced lower overall memory performance. These individuals performed better in the repeated study condition compared to the repeated testing condition when they did not receive feedback during the acquisition phase. Nevertheless, old-old adults demonstrated the expected testing effect when they received feedback in Experiment 2. It is noteworthy that using a quite different paradigm, Kang, McDermott, and Roediger (2007) found evidence in young adults of the importance of feedback in obtaining a testing effect with text materials. Overall, these results indicate that the benefits of repeated testing are particular sensitive to the difficulty of the material and the presence of feedback. Of course, in retrospect, this is quite reasonable because the benefits of retrieval compared to simply restudying are going to be dependent upon the ability to retrieve the correct information, which may be difficult for some materials and in some populations. Feedback insures that the correct information is available. Finally, including feedback in Karpicke and Roediger's paradigm extends its ecological validity, because it is unlikely that in most everyday situations people would test their memory repeatedly without checking on the accuracy of their responses.

Turning to the details of the paradigm used in the present study, when the pairs were correctly recalled once in the repeated testing session, they were dropped from the study list, and thus participants did not receive any restudy opportunities (even though they were tested in subsequent cycles). The lack of feedback produced a greater disadvantage in older adults, because they were more likely to confuse which names occurred with which faces during subsequent tests due to their declarative memory deficits. Indeed, the conditional analyses indicated that final recall of the face–

[4] Given that participants went through all four study–test cycles even after correctly recalling all the face–name pairs in the repeated testing session but stopped once they correctly recalled all face–name pairs in the repeated study session, the mean number of study–test trials was not the same in repeated study and repeated testing sessions. Despite this, on the basis of the number of study–test trials, one might predict that participants across the three age groups would show the testing effect because they took fewer study and test trials to reach 100% accuracy in the repeated study session than in the repeated testing session. However, old-old adults clearly showed better memory in repeated study sessions than in repeated testing sessions when there was no feedback, indicating that the differential number of study and test trials could not fully account for the Feedback × Age × Condition interaction. If anything, the difference in number of trials across the two conditions minimizes the strength of the crossover interaction.

name pairs that had produced intrusion errors during acquisition declined as a function of age only in the repeated testing session when there was no feedback provided during the acquisition phase. It is noteworthy that feedback did not simply reduce the intra-list intrusion errors in the older adults, because they were comparable across no feedback and feedback conditions, but it appears to protect these individuals from being confused by their own intrusion errors during the final recall test.

Using feedback to minimize the influence of intrusion errors on subsequent final recall performance has also been demonstrated in previous studies using an errorless learning paradigm. Baddeley and Wilson (1994; see also Wilson, Baddeley, Evans, & Shiel, 1994) had their participants perform a stem-response learning task (e.g., TH for THUMB). There were multiple study cycles prior to the final test. In the errorful condition, participants were asked to identify the correct answer to the stem by trial and error. Hence, they were likely to generate several erroneous guesses before coming up with the target across the study cycles. In the errorless condition, participants were given a correct answer immediately after the presentation of the stem. After several study trials, participants were tested for the stem–response association in a final cued-recall test. The older adults showed better acquisition and final recall performance in the errorless condition than in the errorful condition. Although this effect was not large in these studies due to older adults' ceiling performance, the benefit of errorless learning has been reported in other studies (e.g., Anderson & Craik, 2006) and has been generalized to other tasks, such as recognition memory (e.g., Lubinsky, Rich, & Anderson, 2009). The importance of minimizing errors during retrieval is consistent with the present observation that feedback reestablishes the benefits of repeated testing in an older adult group with relatively poor declarative memory performance.

Providing feedback during acquisition (i.e., in Experiment 2) not only restored the effectiveness of repeated testing for older adults but also increased metacognitive accuracy for all participants. Similar to Karpicke and Roediger's (2008) findings with young adults, when middle-aged and older adults were not provided with feedback during the acquisition phase, they failed to recognize the benefit of repeated testing, as indicated by their equivalent predicted performance for repeated study and repeated testing sessions. However, when middle-aged and older adults were given feedback, both groups were more accurate in predicting the benefits of testing that mirrored their actual final recall performance, showing that feedback improves metacognitive accuracy (see also Butler, Karpicke, & Roediger, 2008; Roediger, Agarwal, Kang, & Marsh, 2010). Although feedback improved relative metacognitive accuracy, these older adults still underestimated their absolute final recall performance. This seems odd in the face of previous studies, in which it has often been reported that older adults overestimate their absolute memory performance (e.g., Connor, Dunlosky, & Hertzog, 1997). However, because older adults report that face–name pairs are more difficult to recall than other types of materials (e.g., Cohen & Faulkner, 1986), they may simply be relying on the heuristic that they will have difficulty with these materials independent of learning conditions. Indeed, Woo, Schmitter-Edgecombe, and Fancher (2008) also reported that older adults underestimated their retention of face–name associations when they used a memory task different from the one here (i.e., a three-alternative forced-choice, face-cued name recognition task). Thus, it appears that the materials used are quite important in modulating the metacognitive judgments of older adults.

Although we demonstrated the importance of feedback in test-enhanced learning, other test instructions, such as encouraging participants not to guess (i.e., providing an answer only when they were sure that it would be correct), might also increase the likelihood of the old-old participants to produce a testing effect, even in a no-feedback condition. This could not be directly examined in the present experiments, because we did not provide instructions regarding guessing; the participants guessed at their own rate. Interestingly, the proportion of extra-list intrusion was quite similar across different age groups in and across Experiments 1 and 2 (all $F$s $< 1.47$; see Table 3), indicating that participants did not differ in their use of guessing strategies, regardless of their ages and regardless of whether they were given feedback. Of course, it is possible that extra-list intrusions per se may not sufficiently reflect the use of guessing strategies. As depicted in Table 3, the overall omission rate was higher when there was no feedback than when there was feedback in the acquisition phase, $t(138) = 3.45$ and $t(138) = 3.95$, for repeated study and repeated testing sessions, respectively. Although this may indicate that participants were less likely to pass when they received feedback, it is not clear whether it was due to willingness to guess or to having better memory as they were given the feedback. Clearly, the role of guessing in test-enhanced learning (especially without feedback) is an important issue for further study.[5]

## Conclusion

In the present study, we used a modification of Karpicke and Roediger's (2008) repeated study and repeated testing procedures; we replicated their testing effect and extended the observation to face–name learning for our middle-aged adults. However, older adults benefited from repeated testing only when they received feedback. When they did not receive feedback, they showed better delayed recall after they acquired the pairs in a repeated study procedure than in a repeated testing procedure. This failure to find a testing effect was due to older adults making more errors during acquisition. Essentially, testing caused them to learn the erroneous pairings they created during the acquisition phase. We consider this a possible boundary constraint for the use of repeated testing procedures (see also Kang et al., 2007). In this situation, feedback should be included to minimize the opportunity for older adults to be confused by their own intrusion errors. Although feedback might not reduce the likelihood of making such errors during the acquisition phase, it appears to protect them against errors influencing the subsequent final recall. Although the benefit of repeated testing has been typically shown in young adults (even for face–name learning; see Carpenter & DeLosh, 2005), it may not necessarily occur in those young adults who have lower memory

[5] We presented the occupation labels in addition to the face–name pairs, at study and at test, to provide supportive context for the encoding/retrieval of older adults. One could argue that it was the occupation–name association rather than the face–name association that older adults learned and had problems remembering in the experiments. However, this possibility did not undermine the major point of our present study: the role of feedback in test-enhanced learning in older adults.

monitoring abilities or in any participant group when conditions of initial learning are difficult and when feedback is not provided.

## References


Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger, H. L., & McDermott, K. B. (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology, 22,* 861–876. doi:10.1002/acp.1391

Anderson, N. D., & Craik, F. I. M. (2006). The mnemonic mechanisms of errorless learning. *Neuropsychologia, 44,* 2806–2813. doi:10.1016/j.neuropsychologia.2006.05.026

Armitage, S. G. (1945). An analysis of certain psychological tests used for the evaluation of brain injury. *Psychological Monographs, 60* (1, Serial No. 177), 1–48.

Baddeley, A., & Wilson, B. A. (1994). When implicit learning fails: Amnesia and the problem of error elimination. *Neuropsychologia, 32,* 53–68. doi:10.1016/0028-3932(94)90068-X

Balota, D. A., Duchek, J. M., & Paullin, R. (1989). Age-related differences in the impact of spacing, lag and retention interval. *Psychology and Aging, 4,* 3–9. doi:10.1037/0882-7974.4.1.3

Balota, D. A., Duchek, J. M., Sergent-Marshall, S. D., & Roediger, H. L. III (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,* 19–31. doi:10.1037/0882-7974.21.1.19

Balota, D. A., & Faust, M. (2001). Attention in dementia of the Alzheimer's type. In F. Boller & S. Cappa (Eds.), *Handbook of neuropsychology* (2nd ed., pp. 51–80). New York, NY: Elsevier Science.

Barresi, B. A., Obler, L. K., & Goodglass, H. (1998). Dissociation between proper name and common noun learning. *Brain and Cognition, 37,* 21–23.

Berg, L., McKeel, D. W., Jr., Miller, J. P., Storandt, M., Rubin, E. H., Morris, J. C., . . . Saunders, A. M. (1998). Clinicopathologic studies in cognitively healthy aging and Alzheimer's disease: Relation of histologic markers to dementia severity, age, sex, and apolipoprotein E genotype. *Archives of Neurology, 55,* 326–335. doi:10.1001/archneur.55.3.326

Botwinick, J., & Storandt, M. (1973). Age differences in reaction time as a function of experience, stimulus intensity, and preparatory interval. *Journal of Genetic Psychology, 123,* 209–217.

Burke, W. J., Miller, J. P., Rubin, E. H., Morris, J. C., Coben, L. A., Duchek, J. M., . . . Berg, L. (1988). Reliability of the Washington University Clinical Dementia Rating. *Archives of Neurology, 45,* 31–32.

Butler, A. C., Karpicke, J. D., & Roediger, H. L. (2008). Correcting a metacognitive error: Feedback increases retention of low-confidence correct responses. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 918–928. doi:10.1037/0278-7393.34.4.918

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

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

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition, 20,* 633–642.

Cohen, G. (1990). Why is it difficult to put names to faces? *British Journal of Psychology, 81,* 287–297.

Cohen, G., & Burke, D. M. (1993). Memory for proper names: A review. *Memory, 1,* 249–263. doi:10.1080/09658219308258237

Cohen, G., & Faulkner, D. (1986). Memory for proper names: Age differences in retrieval. *British Journal of Developmental Psychology, 4,* 187–197.

Connor, L. T., Dunlosky, J., & Hertzog, C. (1997). Age-related differences in absolute but not relative metamemory accuracy. *Psychology and Aging, 12,* 50–71. doi:10.1037/0882-7974.12.1.50

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. A. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology: General, 128,* 309–331. doi:10.1037/0096-3445.128.3.309

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,* 189–198. doi:10.1016/0022-3956(75)90026-6

Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology, 81,* 392–399. doi:10.1037/0022-0663.81.3.392

Goodglass, H., & Kaplan, E. (1983). *Boston Naming Test.* Philadelphia, PA: Lea & Febiger.

Grober, E., Buschke, H., Crystal, H. A., Bang, S., & Dresner, R. (1988). Screening for dementia by memory testing. *Neurology, 38,* 900–903.

Groninger, L. D., & Murray, K. N. (2004). Reminiscence, forgetting, and hypermnesia using face–name learning: Isolating the effects using recall and recognition memory measures. *Memory, 12,* 351–365. doi:10.1080/09658210344000044

Helder, E., & Shaughnessy, J. J. (2008). Retrieval opportunities while multitasking improve name recall. *Memory, 16,* 896–909. doi:10.1080/09658210802360611

James, L. E. (2004). Meeting Mr. Farmer versus meeting a farmer: Specific effects of aging on learning proper names. *Psychology and Aging, 19,* 515–522. doi:10.1037/0882-7974.19.3.515

James, L. E., Fogler, K. A., & Tauber, S. K. (2008). Recognition memory measures yield disproportionate effects of aging on learning face–name associations. *Psychology and Aging, 23,* 657–664. doi:10.1037/a0013008

Kang, S. H. K., McDermott, K. B., & Roediger, H. L. (2007). Test format and feedback modulate the effect of testing on memory retention. *European Journal of Cognitive Psychology, 19,* 528–558. doi:10.1080/09541440601056620

Karpicke, J. D., Butler, A. C., & Roediger, H. L. (2009). Metacognitive strategies in student learning: Do students practice retrieval when they study on their own? *Memory, 17,* 471–479. doi:10.1080/09658210802647009

Karpicke, J. D., & Roediger, H. L. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language, 57,* 151–162. doi:10.1016/j.jml.2006.09.004

Karpicke, J. D., & Roediger, H. L. (2008, February 15). The critical importance of retrieval for learning. *Science, 319,* 966–968. doi:10.1126/science.1152408

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review, 14,* 219–224.

Lubinsky, T., Rich, J. B., & Anderson, N. D. (2009). Errorless learning and elaborative self-generation in healthy older adults and individuals with amnestic mild cognitive impairment: Mnemonic benefits and mechanisms. *Journal of the International Neuropsychological Society, 15,* 704–716. doi:10.1017/S1355617709990270

McDaniel, M. A., Einstein, G. O., & Jacoby, L. L. (2008). New considerations in aging and memory: The glass may be half full. In. F. I. M. Craik & T. A. Salthouse (Eds.), *The handbook of aging and cognition* (3rd ed., pp. 251–310). Hove, England: Psychology Press.

McDaniel, M. A., & Fisher, R. P. (1991). Tests and test feedback as learning sources. *Contemporary Educational Psychology, 16,* 192–201. doi:10.1016/0361-476X(91)90037-L

McDermott, K. B. (2006). Paradoxical effects of testing: Repeated retrieval attempts enhance the likelihood of later accurate and false recall. *Memory & Cognition, 34,* 261–267.

McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., & Stadlan, E. M. (1984). Clinical diagnosis of Alzheimer's disease: Report

of the NINCDS-ADRDA work group under the auspices of the Department of Health and Human Services Task Force on Alzheimer's disease. *Neurology, 34,* 939–944.

McWeeny, K. H., Young, A. W., Hay, D. C., & Ellis, A. W. (1987). Putting names to faces. *British Journal of Psychology, 78,* 143–149.

Morris, J. C. (1993). The Clinical Dementia Rating (CDR): Current version and scoring rules. *Neurology, 43,* 2412–2414.

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,* 1170–1187. doi:10.1037/0278-7393.26.5.1170

Naveh-Benjamin, M., Guez, J., Kilb, A., & Reedy, S. (2004). The associative memory deficit of older adults: Further support using face–name associations. *Psychology and Aging, 19,* 541–546. doi:10.1037/0882-7974.19.3.541

Naveh-Benjamin, M., Shing, Y.-L., Kilb, A., Werkle-Bergner, M., Lindenberger, U., & Li, S.-C. (2009). Adult age differences in memory for name–face associations: The effects of intentional and incidental learning. *Memory, 17,* 220–232. doi:10.1080/09658210802222183

Psychological Image Collection at Stirling [Photographic database]. (n.d.). Retrieved from http://pics.psych.stir.ac.uk/

Neuschatz, J. S., Preston, E. L., Toglia, M. P., & Neuschatz, J. S. (2005). Comparison of the efficacy of two name-learning techniques: Expanding rehearsal and name–face imagery. *American Journal of Psychology, 118,* 79–101.

Rendell, P. G., Castel, A. D., & Craik, F. I. M. (2005). Memory for proper names in old age: A disproportionate impairment? *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 58*A, 54–71.

Roediger, H. L., Agarwal, P. K., Kang, S. H. K., & Marsh, E. J. (2010). Benefits of testing memory: Best practices and boundary conditions. In G. M. Davies & D. B. Wright (Eds.), *New frontiers in applied memory* (pp. 13–49). Brighton, United Kingdom: Psychology Press.

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

Thurstone, L. E., & Thurstone, T. G. (1949). *Examiner manual for the SRA Primary Mental Abilities Test.* Chicago, IL: Science Research Associates.

Troyer, A. K., Hafliger, A., Cadieux, M. J., & Craik, F. I. M. (2006). Name and face learning in older adults: Effects of level of processing, self-generation, and intention to learn. *Journals of Gerontology, Series B: Psychological Sciences and Social Sciences, 61,* P67–P74.

Tulving, E. (1967). The effects of presentation and recall of material in free-recall learning. *Journal of Verbal Learning and Verbal Behavior, 6,* 175–184. doi:10.1016/S0022-5371(67)80092-6

Verhaeghen, P., & Marcoen, A. (1996). On the mechanisms of plasticity in young and older adults after instruction in the method of loci: Evidence for an amplification model. *Psychology and Aging, 11,* 164–178. doi:10.1037/0882-7974.11.1.164

Wechsler, D. (1987). *Manual: Wechsler Memory Scale–Revised.* San Antonio, TX: Psychological Corporation.

Wechsler, D. (1997). *Wechsler Adult Intelligence Scale: Administration and scoring manual* (3rd ed.). San Antonio, TX: Psychological Corporation.

Wheeler, M. A., & Roediger, H. L. (1992). Disparate effects of repeated testing: Reconciling Ballard's (1913) and Bartlett's (1932) results. *Psychological Science, 3,* 240–245. doi:10.1111/j.1467-9280.1992.tb00036.x

Wilson, B. A., Baddeley, A., Evans, J., & Shiel, A. (1994). Errorless learning in the rehabilitation of memory-impaired people. *Neuropsychological Rehabilitation, 4,* 307–326. doi:10.1080/09602019408401463

Woo, E., Schmitter-Edgecombe, M., & Fancher, J. B. (2008). Memory prediction accuracy in younger and older adults: A cross-sectional and longitudinal analysis. *Neuropsychology, Development, and Cognition, Section B: Aging, Neuropsychology and Cognition, 15,* 68–94. doi:10.1080/13825580701626936

## Appendix

### Experimental Stimuli

| First name | Last name | Occupation | Gender | Age |
|---|---|---|---|---|
| Set 1 | | | | |
| Annie | Jamison | teacher | female | old |
| Laura | Hamilton | jeweler | female | old |
| Megan | Higgins | waitress | female | young |
| Jenny | Morrison | florist | female | young |
| Aaron | Vaughn | geologist | male | old |
| Kyle | Sullivan | pilot | male | old |
| Alex | Massey | architect | male | young |
| Jacob | Armstrong | cashier | male | young |
| Set 2 | | | | |
| Betty | Price | seamstress | female | old |
| Colleen | Weber | secretary | female | old |
| Natalie | Lowe | nurse | female | young |
| Kate | Campbell | beautician | female | young |
| Anthony | Foster | plumber | male | old |
| Paul | Thompson | electrician | male | old |
| Ben | Gianico | painter | male | young |
| John | Dobbins | student | male | young |

Received November 13, 2009
Revision received March 26, 2010
Accepted April 12, 2010 ■