Title
Age-Related Impairment on a Forced-Choice Version of the Mnemonic Similarity Task

Journal
Behavioral Neuroscience, 131(1)

ISSN
0735-7044

Authors
Huffman, Derek J
Stark, Craig EL

Publication Date
2017-02-01

DOI
10.1037/bne0000180

Behavioral Neuroscience
2017, Vol. 131, No. 1, 55–67

0735-7044/17/$12.00 http://dx.doi.org/10.1037/bne0000180

Age-Related Impairment on a Forced-Choice Version of the Mnemonic Similarity Task

Derek J. Huffman and Craig E. L. Stark
University of California, Irvine

Previous studies from our lab have indicated that healthy older adults are impaired in their ability to mnemonically discriminate between previously viewed objects and similar lure objects in the Mnemonic Similarity Task (MST). These studies have used either old/similar/new or old/new test formats. The forced-choice test format (e.g., "Did you see object A or object A′ during the encoding phase?") relies on different assumptions than the old/new test format (e.g., "Did you see this object during the encoding phase?"); hence, converging evidence from these approaches would bolster the conclusion that healthy aging is accompanied by impaired performance on the MST. Consistent with our hypothesis, healthy older adults exhibited impaired performance on a forced-choice test format that required discriminating between a target and a similar lure. We also tested the hypothesis that age-related impairments on the MST could be modeled within a global matching computational framework. We found that decreasing the probability of successful feature encoding in the models caused changes that were similar to the empirical data in healthy older adults. Collectively, our behavioral results using the forced-choice format extend the finding that healthy aging is accompanied by an impaired ability to discriminate between targets and similar lures, and our modeling results suggest that a diminished probability of encoding stimulus features is a candidate mechanism for memory changes in healthy aging. We also discuss the ability of global matching models to account for findings in other studies that have used variants on mnemonic similarity tasks.

Keywords: cognitive aging, memory, modeling

Previous research has established that healthy older adults exhibit impaired performance on tests of associative memory. For example, a meta-analysis revealed that tests of source memory are impaired to a greater degree than tests of item memory (Spencer & Raz, 1995). More generally, Naveh-Benjamin and colleagues developed and tested an associative deficit hypothesis to account for memory changes among healthy older adults (e.g., Naveh-Benjamin, 2000; Naveh-Benjamin, Guez, Kilb, & Reedy, 2004; Old & Naveh-Benjamin, 2008b; for a meta-analysis, see Old & Naveh-Benjamin, 2008a). Specifically, they reported a greater age-related impairment on tests of associative memory than tests of single-item memory. Other studies have noted a greater impairment on memory recall tests than on traditional item- recognition memory tests (i.e., targets vs. unrelated foils; Craik & McDowd, 1987; Danckert & Craik, 2013). Taken together, there is unequivocal evidence for an age-related impairment on tasks that tax recollection and associative memory, with a milder—and sometimes not statistically significant—deficit on tests of simpler item-recognition memory.

Previous studies from our lab and others have shown that there are conditions in which healthy older adults exhibit a clear impairment on item-recognition memory tasks. For example, our lab developed a Mnemonic Similarity Task (MST; formerly BPS-O) that assesses a participants' ability to discriminate between previously viewed objects (i.e., targets), similar lure objects, and unrelated foil objects (Kirwan et al., 2012; Kirwan & Stark, 2007; Stark, Stevenson, Wu, Rutledge, & Stark, 2015; Stark, Yassa, Lacy, & Stark, 2013; Yassa, Lacy, et al., 2011, Yassa, Mattfeld, Stark, & Stark, 2011). The ability to discriminate between targets and similar lures has been shown to be impaired in healthy older adults, with an apparent sparing of their ability to discriminate between targets and unrelated foils (Bennett, Huffman, & Stark, 2015; Reagh et al., 2016; Stark et al., 2013, 2015; Toner, Pirogovsky, Kirwan, & Gilbert, 2009; Yassa, Lacy, et al., 2011; Yassa, Mattfeld, et al., 2011). Importantly, working memory versions of the task have failed to find age-related differences in the ability to discriminate between targets and similar lures (e.g., Yassa, Lacy, et al., 2011), suggesting a mnemonic rather than a perceptual deficit driving the effect in healthy older adults in this task. Previous studies have typically used a test format in which participants are instructed to respond "old" to exact repetitions of items seen during the encoding phase, to respond "similar" to images which are similar to—but not exactly the same as—a previously viewed image, and to respond "new" to images that they have not seen in the context of the experiment (Bennett et al., 2015; Kirwan

This article was published Online First December 22, 2016.

Derek J. Huffman and Craig E. L. Stark, Department of Neurobiology and Behavior, Center for the Neurobiology of Learning and Memory, University of California, Irvine.

This research was supported by National Institute on Aging Grant R01 AG034613 and National Institutes of Health Grant R01 MH085828 awarded to Craig E. L. Stark. We thank Patricia Place for assistance with data collection. We thank Jessica German for assistance with data collection, analysis, and helpful conversations about Experiment 1. We thank Shauna Stark and Veronique Boucquey for helpful discussions.

Correspondence concerning this article should be addressed to Craig E. L. Stark, 320 Qureshey Research Laboratory, University of California, Irvine, Irvine, CA 92697-3800. E-mail: cestark@uci.edu

et al., 2012; Kirwan & Stark, 2007; Stark et al., 2013, 2015; Toner et al., 2009; Yassa, Lacy, et al., 2011; Yassa, Mattfeld, et al., 2011). Stark et al. (2015) also used a test format that instructed participants to respond "old" only to exact repetitions and to respond "new" to both similar lures and unrelated foils, including a version with confidence ratings. The results from these tests have consistently shown an age-related impairment in the ability to discriminate between targets and similar lures.

An unaddressed question is whether healthy older adults would exhibit impaired performance on a forced-choice version of the MST. There are two reasons why the forced-choice format could differ from an old/new (or old/similar/new) format that would impact our understanding of age-related decline in performance on the task. First, between-groups differences in response criteria can lead to apparent differences in accuracy (Green & Swets, 1966; Stanislaw & Todorov, 1999). While we have not observed differences in response criteria (Stark et al., 2015), the forced-choice format negates any differences directly and is a more powerful means to address this confound than the failure to observe an effect. Here, we used several versions of the forced-choice procedure, similar to previous reports (Holdstock et al., 2002; Jeneson, Kirwan, Hopkins, Wixted, & Squire 2010; Migo et al., 2014; Migo, Montaldi, Norman, Quamme, & Mayes, 2009; Tulving, 1981). In each of the test formats used here, we displayed one target object and one distractor object and participants were instructed to choose the exact object that they saw during the encoding phase; thus, we used a two-alternative forced-choice procedure. In the first test format, participants were shown a target object (A) and an unrelated foil (X), which we refer to as A-X. In the second test format, participants were shown a target object and its corresponding similar lure (A′), which we refer to as A-A′ (dubbed *FCC* by Migo et al., 2009). In the third test format, participants were shown a target object and a noncorresponding lure (B′, a lure that is similar to a different studied object [B]), which we refer to as A-B′ (dubbed *FCNC* by Migo et al., 2009).

The second motivation for using a forced-choice version of the MST in younger and healthy older adults is that the old/new test format and the forced-choice test format have been hypothesized to rely on different cognitive processes. For example, the dual-process complementary learning systems model has been used to advance the notion that patients with hippocampal damage will be impaired on the old/new test format with targets and similar lures and on the A-B′ test format but will be relatively spared on the A-A′ test format, the A-X test format, and the old/new test format with targets and unrelated foils (Norman & O'Reilly, 2003; Norman, 2010; also see Holdstock et al., 2002; Migo et al., 2009, 2014). A study of a single patient with selective hippocampal damage revealed impaired performance on the old/new test format with targets and similar lures and intact performance on the A-A′ test format, supporting the predictions from the model (Holdstock et al., 2002). However, other studies with a larger sample of patients with selective hippocampal damage have shown a similar impairment on both the A-A′ test format and the old/new test format with targets and similar lures (Bayley, Wixted, Hopkins, & Squire, 2008; Jeneson et al., 2010). Additionally, Jeneson et al. (2010) revealed that the patients were equally impaired on the A-A′ test format, the A-B′ test format, and the old/new test format with targets and similar lures. Although the results from patients with hippocampal damage are equivocal, the forced-choice test format can provide further insight into the organization of memory in younger adults and can elucidate the nature of memory changes that occur in the course of healthy aging.

Our primary aim was to investigate whether healthy older adults would exhibit impaired performance on a forced-choice version of the MST. Our second aim was to investigate whether younger and older adults would exhibit an effect of test format. To address these questions, we conducted two behavioral experiments. In Experiment 1, we included both younger and older adults and used the three test formats mentioned above: A-X, A-A′, and A-B′. In Experiment 2, we aimed to replicate our findings from Experiment 1 and to rule out the possibility that the presence of the A-X test format was artificially impairing performance on the A-B′ test format. In a within-subjects design, participants performed two study-test cycles (with independent stimulus sets), one that included all three test formats and one that did not include the A-X test format. Our third aim was to investigate whether a class of models from mathematical psychology—global matching models (e.g., Hintzman, 1984, 1988; Murdock, 1982, 1995)—could account for our empirical results in both younger and healthy older adults. Specifically, we tested the hypothesis that healthy aging could be modeled as an impaired ability to encode stimulus features. We conclude by discussing the application of global matching models to interpret the results of other experiments that have used variants on mnemonic similarity tasks.

Experiment 1

In Experiment 1, we developed a forced-choice version of the MST to investigate memory performance in younger adults and healthy older adults. The task uses a set of pictures of objects (Figure 1A) that each have a similar version that can be used as a lure item at test. Through extensive testing, we have previously demonstrated that the lures have a controlled range of false alarm rates when used during recognition tests (Stark et al., 2013, 2015; Yassa, Lacy, et al., 2011). In addition, performance on these similar lures is sensitive to hippocampal damage (Kirwan et al., 2012), to aging (Stark et al., 2013, 2015; Toner et al., 2009) and the age-related changes in both the activity of the dentate gyrus and CA3 subfields during aging (Yassa, Mattfeld, et al., 2011), and to disruptions of hippocampal circuitry (Bennett et al., 2015; Yassa, Mattfeld, et al., 2011). Together, these findings demonstrate the viability of this task as a sensitive and appropriate measure of age-related memory change and of hippocampal function.

We tested the hypothesis that the age-related impairment in mnemonically discriminating between previously viewed objects and similar lure objects would extend from the old/similar/new test format and the old/new test format to the forced-choice test format. The forced-choice test format relies on different assumptions than the old/similar/new test format and the old/new test format; therefore, if similar results are observed in a forced-choice test format, then it would provide further support for the notion that healthy aging is accompanied by impaired performance on the MST. We included three forced-choice test formats: (1) A-X, (2) A-A′, and (3) A-B′ (Figure 1B). In addition to assessing age-related changes, we tested whether there was an effect of test format on performance in both younger and older adults.

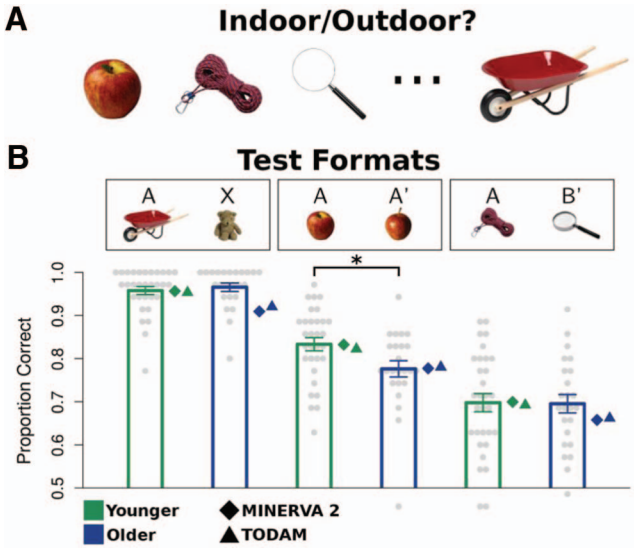

Figure 1. Investigation of performance in younger adults and healthy older adults in Experiment 1. (A) The encoding task was an incidental design in which participants indicated whether each item was an indoor or an outdoor item. (B) The test phase consisted of three test formats: A-X, A-A′, and A-B′. In both age groups, there was a significant effect of test format. A repeated-measures analysis of variance revealed an Age × Test format interaction, which was driven by better performance of younger adults on the A-A′ test format. The bar plots represent the mean proportion correct; the error bars represent the standard error of the mean; and the gray dots represent individual participants. We also investigated whether two global matching models, MINERVA 2 and TODAM, could account for the results in both younger and older adults. Both models were able to capture the A-X > A-A′ > A-B′ effect (data points represent mean proportion correct). After finding model parameters that provided a good fit to the data from younger adults (MINERVA 2: $L = 0.65$; TODAM: $p = .5$), we decreased the encoding probability to attempt to model the data from older adults (MINERVA 2: $L = 0.55$; TODAM: $p = .35$). In both models, decreasing the encoding parameter caused a decrease in proportion correct on all three test formats, but the largest change in performance was on the A-A′ test format. * $p < .05$. See the online article for the color version of this figure.

Method

Participants. Participants were 32 younger adults (18–28 years old) and 27 healthy older adults (64–85 years old). Older adults were screened to ensure that they did not have a memory impairment, similar to previous studies in our lab (e.g., Stark et al., 2013, 2015). Specifically, we ensured that participants scored in the normal range for their age group on the Mini-Mental Status Examination (Crum, Anthony, Bassett, & Folstein, 1993) and the Rey Auditory Verbal Learning Task (Rey, 1941). We excluded three older adults because they did not score within 1.5 *SD* of the mean for their age. Additionally, we excluded one younger adult due to very poor performance on our behavioral task (proportion correct ≈ 0.5 on all three test formats; more than 10 *SD* below the mean of the included participants on the A-X test format). Thus, 31 younger adults (26 female, 5 male) and 24 older adults (19 female, 5 male) were included in our analysis. Study and consent procedures were approved in accordance with the University of California, Irvine, internal review board (HS#2008–6128, "fMRI Studies of Memory Encoding and Retrieval").

Task design. Participants performed an incidental encoding task in which they indicated, via button press, whether they thought that the object in each picture was more of an "indoor" or an "outdoor" object (Figure 1A; Stark et al., 2013, 2015). The encoding phase consisted of 140 images, which were displayed for 2,000 ms with a 500-ms interstimulus interval. Following the encoding phase, participants performed a memory test, which contained three forced-choice test formats (top of Figure 1B): (1) A-X (i.e., a target and an unrelated foil), (2) A-A′ (i.e., a target and a corresponding similar lure), and (3) A-B′ (i.e., a target and a lure object from a different pair). On each test trial, one object was presented on the left side of the screen and one object was presented on the right side of the screen. Participants were told that on all trials they would view one image that they saw during the indoor/outdoor task and one new image. Moreover, they were told that on some trials the new image would be completely different than any of the images from the indoor/outdoor phase whereas on other trials the image would be similar to—but not exactly the same as—a previously viewed image from the indoor/outdoor phase. Participants were instructed to select, via button press, the exact image that they saw during the indoor/outdoor phase of the experiment. The images were displayed until the participant made a response or for 4 s, at which point the image disappeared and there was an unlimited response window. The target was randomly assigned to the left and right side of the screen on a trial-by-trial basis. The test formats were presented in a random, intermixed order (i.e., the task conditions varied on a trial-by-trial basis). Participants performed 35 trials of each test format.

Our lab has previously calculated empirical estimates of the mnemonic similarity of the stimuli that we used in the present experiment (Lacy, Yassa, Stark, & Stark, 2011; Stark et al., 2013; Yassa, Lacy, et al., 2011). Briefly, in over 100 participants, the mean proportion of times that participants responded "old" to a similar lure object was used as an index of mnemonic similarity (i.e., the higher the probability of responding "old" in response to a similar lure, the higher its mnemonic similarity). The stimuli were rank-ordered and divided into five "lure bins." In the present experiment, we balanced the similarity of the stimuli at two levels: (a) the stimulus set: the number of trials from each lure bin in each test format (7 stimuli per lure bin), and (b) the individual trial level: the lure bin of the target and distractor image. The former ensured that the similarity of targets and similar lures was balanced across the A-A′ and the A-B′ test formats for every subject. The latter addressed the potential issue of encoding versus retrieval difficulty of stimuli from different lure bins, which is particularly important for the A-B′ test format.

Data analysis. We calculated the proportion correct for each test format for each participant. To investigate whether there was an effect of test format, irrespective of age, we performed a separate one-way analysis of variance (ANOVA) in each age group. We performed planned tests to investigate whether performance was ranked in the following order: (1) A-X, (2) A-A′, and (3) A-B′. To investigate whether there was an age by test format interaction, we performed a mixed-design ANOVA (between-subjects variable: age group; within-subjects variable: test format). We performed planned tests to investigate whether performance

differed between younger and older adults on the A-A′ test format and on the A-B′ test format.

Results

We first investigated whether there was an effect of test format on performance in both age groups (Figure 1B). Separate one-way ANOVAs revealed a significant main effect of test format in both younger adults (YA; $F = 110.7$, $p < .001$) and older adults (OA; $F = 100.2$, $p < .001$). Planned comparisons revealed that both age groups performed better on the A-X test format than the A-A′ test format (YA: mean proportion correct A-X = 0.958, $SD = 0.052$, mean proportion correct A-A′ = 0.833, $SD = 0.085$, $t_{30} = 8.97$, $p < .001$, 95% confidence interval [CI] [0.096, 0.153]; OA: mean proportion correct A-X = 0.965, $SD = 0.048$, mean proportion correct A-A′ = 0.776, $SD = 0.093$, $t_{23} = 11.36$, $p < .001$, 95% CI [0.155, 0.224]) and better on the A-A′ test format than the A-B′ test format (YA: mean proportion correct A-B′ = 0.698, $SD = 0.118$, $t_{30} = 6.63$, $p < .001$, 95% CI [0.094, 0.177]; OA: mean proportion correct A-B′ = 0.695, $SD = 0.105$, $t_{23} = 3.58$, $p < .005$, 95% CI [0.034, 0.128]). These results suggest that there is an effect of test format in both age groups. Notably, although performance was worst on the A-B′ test format, performance was significantly better than chance (0.5) in both age groups (YA: $t_{30} = 9.33$, $p < .001$, 95% CI [0.654, 0.741]; OA: $t_{23} = 9.13$, $p < .001$, 95% CI [0.651, 0.739]).

We next investigated whether there was an effect of healthy aging on performance. A mixed-design ANOVA (between-subjects variable: age group; within-subjects variable: test format) revealed a significant Age group × Test format interaction ($F = 3.51$, $p = .033$). Planned comparisons revealed that younger adults performed significantly better than older adults on the A-A′ test format ($t_{53} = 2.37$, $p < .025$, 95% CI [0.009, 0.105]). Conversely, the difference between younger and older adults failed to reach significance for the A-B′ test format ($t_{53} = 0.081$, $p = .94$, 95% CI [−0.059, 0.064]; Figure 1B). These results extend the previous findings of an age-related decline in performance on the old/new and the old/similar/new test format with targets and similar lures (Bennett et al., 2015; Stark et al., 2013, 2015; Toner et al., 2009; Yassa, Lacy, et al., 2011; Yassa, Mattfeld, et al., 2011) to the A-A′ test format.

Previous studies that used the old/similar/new test format reported an age-related impairment in the ability to discriminate between targets and similar lures with intact discrimination between targets and unrelated foils (Bennett et al., 2015; Stark et al., 2013, 2015; Toner et al., 2009; Yassa, Lacy, et al., 2011; Yassa, Mattfeld, et al., 2011). Similarly, there was no evidence that younger adults performed better than healthy older adults on the A-X test format ($t_{53} = -0.58$, $p = .57$, 95% CI [−0.035, 0.020]). It is possible that an age-related difference on the A-X test format was obscured by a ceiling effect; however, when we compared the 15 worst-performing younger adults (i.e., median split) and the 12 worst-performing older adults (i.e., median split), the difference still failed to reach significance ($t_{25} = -0.56$, $p = .58$, 95% CI [−0.054, 0.031]).

Discussion

We investigated whether the previous reports of an age-related impairment on the MST (Bennett et al., 2015; Stark et al., 2013, 2015; Toner et al., 2009; Yassa, Lacy, et al., 2011, Yassa, Mattfeld, et al., 2011) would extend to the forced-choice test format. Our results revealed a significant Age × Test format interaction, which was driven by better performance in younger adults than healthy older adults on the A-A′ test format. These results suggest that the age-related impairment on the old/similar/new and the old/new test formats with targets and similar lures extends to the A-A′ test format.

We observed an effect of test format in younger adults and healthy older adults. In both age groups, performance was best on the A-X format, followed by the A-A′ format, followed by the A-B′ format. These findings support previous studies that have reported better performance on the A-A′ test format than the A-B′ test format (Hintzman, 1988; Jeneson et al., 2010; Migo et al., 2014; Tulving, 1981) while also raising the question of why performance is worse on the the A-B′ test format than the A-A′ test format. One possibility is that the presence of the A-X test format was affecting performance on the A-B′ test format. These results also raise the question of whether performance on the old/new test format with targets and similar lures more closely resembles performance on the A-A′ test format or the A-B′ test format. We investigated these questions in Experiment 2.

Experiment 2

In Experiment 2, we aimed to replicate the findings in younger adults from Experiment 1 as well as to rule out the possibility that the A-X test format artificially reduced A-B′ test format performance. Specifically, we thought that it was possible that the presence of the A-X test trials increased the propensity for participants to immediately select the first item that they viewed in the A-B′ test format. Thus, in Experiment 2, participants performed two study-test cycles (with distinct stimulus sets), one that included all three test formats and one that included only the A-A′ test format and the A-B′ test format.

Method

Participants. Participants were 21 younger adults (18–33 years old). We excluded one participant due to very poor performance on our behavioral task (proportion correct ≈ 0.5 on all of the test formats; more than 10 *SD* below the mean of the included participants on the A-X test format). Thus, 20 participants (14 female, 6 male) were included in our analysis.

Task design. The behavioral tasks in Experiment 2 were similar to Experiment 1. In Experiment 2, participants performed two encoding and two testing phases, each with a distinct stimulus set. Previous research in our lab has ensured that the two stimulus sets are very well matched in terms of similarity of the lure pairs (Stark et al., 2015). The encoding phases were identical to those in Experiment 1. The two test phases differed in the number of test formats used. The purpose of this manipulation was to address whether the A-X test trials were artificially reducing performance on the A-B′ test trials. Accordingly, one version used three test formats, as in Experiment 1, and the other version used two test formats: A-A′ and A-B′ (i.e., never showing an unrelated foil item as an option). In the two-test version, participants were instructed that on each trial they would view one image that was in the indoor/outdoor task and one image that was similar to—but not

exactly the same as—an image from the indoor/outdoor task. Moreover, they were instructed that on some trials the similar image would be from the same pair (e.g., if they studied an image of an apple they might see the exact apple and a similar apple) and on some trials the similar image would be from a different pair (e.g., if they studied an apple and an orange, they might see the exact apple and a similar orange). The order in which participants received the three-test version and the two-test format version was counterbalanced between participants. As in Experiment 1, participants performed 35 trials of each test format, and the order of the trials varied on a trial-by-trial basis. Thus, the test phase contained 35 fewer trials in the two-test version.

Data analysis. A one-way ANOVA was used to test for the presence of a main effect of test format in the three-test version. We performed planned tests to investigate whether performance was ranked in the following order: (1) A-X, (2) A-A′, and (3) A-B′. For the two-test version, we performed a planned test to investigate whether performance was better on the A-A′ test format than the A-B′ test format. We investigated whether performance was enhanced on the two-test version relative to the three-test version using separate paired *t* tests for the A-A′ and the A-B′ test formats.

Results

Main results. In the three-test condition, a one-way ANOVA revealed a significant main effect of test format ($F = 73.1$, $p < .001$; Figure 2A). Planned comparisons revealed significantly better performance on the A-X test format than the A-A′ test format (proportion correct A-X: $M = 0.959$, $SD = 0.057$; proportion correct A-A′: $M = 0.836$, $SD = 0.065$; $t_{19} = 9.36$, $p < .001$, 95% CI [0.095, 0.150]) and significantly better performance on the A-A′ test format than the A-B′ test format (proportion correct A-B′: $M = 0.756$, $SD = 0.088$; $t_{19} = 4.38$, $p < .001$, 95% CI [0.042, 0.118]). In the two-test version, a paired *t* test revealed significantly better performance on the A-A′ test format than the A-B′ test format (proportion correct A-A′: $M = 0.837$, $SD = 0.091$; proportion correct A-B′: $M = 0.746$, $SD = 0.106$; $t_{19} = 3.05$, $p < .01$, 95% CI [0.029, 0.154]; Figure 2B). Finally, paired *t* tests revealed no sign of a benefit for the two-test version over the three-test version for either the A-A′ test format ($M = 0.001$, $t_{19} = 0.078$, $p = .94$, 95% CI [−0.037, 0.040]) or the A-B′ test format ($M = -0.010$, $t_{19} = -0.46$, $p = .65$, 95% CI [−0.055, 0.035]). These results replicate the effect of test format that we observed in Experiment 1. Moreover, these results rule out the possibility that the A-X test format was artificially reducing performance on the A-B′ test format.

Comparison of performance on forced-choice and old/new test formats. The results of Experiment 1 and 2 provide clear evidence that performance is better on the A-A′ test format than the A-B′ test format. Previous reports have suggested that performance on the A-A′ test format can rely on familiarity to a greater degree than performance on the A-B′ test format and the old/new test format with targets and similar lures (Holdstock et al., 2002; Migo et al., 2009, 2014). Accordingly, we were interested in examining whether performance on the A-A′ test format was better than performance on the old/new test format with targets and similar lures. Similarly, we were interested in investigating whether performance was better on the A-X test format than performance on the old/new test format with targets and unrelated foils.

To compare old/new performance with performance on the forced-choice tests from Experiment 2, we reanalyzed data from 20 younger adults from a previous study from our lab (Experiment 4 in the work of Stark et al., 2015). As in Experiment 2, participants performed two study-test cycles with two unique stimulus sets. The encoding phase consisted of an indoor/outdoor judgment for each of 128 images of objects. One of the test formats used "gist" instructions (i.e., participants were instructed to respond "old" to similar lures) while the other test format used "veridical" instructions (i.e., participants were instructed to respond "new" to similar lures), and the order of the test formats was counterbalanced across participants. For the present analysis, we used the

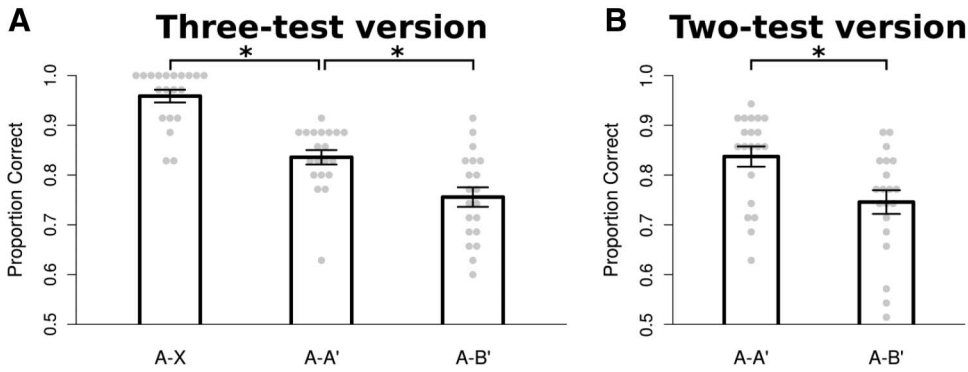

Figure 2. Investigation of performance on the three-test version and the two-test version in younger adults in Experiment 2. (A) A one-way analysis of variance revealed a significant effect of test format on performance ($F = 73.1$, $p < .001$). Paired *t* tests revealed significantly better performance on the A-A′ condition compared to the A-B′ condition for both the three-test version ($t_{19} = 4.38$, $p < .001$) and the two-test version ($t_{19} = 3.05$, $p < .01$). Paired *t* tests revealed no sign of a benefit for the two-test version over the three-test version for either the A-A′ format ($M = 0.001$, $t_{19} = 0.078$, $p = .94$) or the A-B′ format ($M = -0.010$, $t_{19} = -0.46$, $p = .65$). The bar plots represent the mean proportion correct; the error bars represent the standard error of the mean; and the gray dots represent individual participants. * $p < .05$.

data from the veridical condition because the test instructions were equivalent to our instructions for Experiment 2. The test phase consisted of three probe types: (a) targets (exact repetitions), (b) similar lures, and (c) unrelated foils. Participants were instructed to respond "old" only for exact repetitions and to respond "new" for both similar lures and for novel foils. After making the old/new decision, participants indicated the confidence of their response (very sure, somewhat sure, somewhat unsure, very unsure), resulting in eight confidence bins (ranging from "very sure old" to "very sure new"). Participants performed 64 trials of each probe type. Three participants were excluded due to a failure to distribute responses across the confidence bins (which resulted in poor model fit); thus 17 participants were included in the between-groups analysis.

The area under the receiver operating characteristic (ROC) curve—calculated from the old/new test format with confidence ratings—is mathematically equivalent to the proportion correct on the two-alternative forced-choice test format (Green & Moses, 1966; Green & Swets, 1966; Stanislaw & Todorov, 1999; Swets & Pickett, 1982). The preferred approach for estimating the area under the ROC curve is to use maximum-likelihood estimation to fit the z-transformed ROC curve—a measure referred to as A_z (Stanislaw & Todorov, 1999; Swets & Pickett, 1982). Importantly, A_z does not assume equal variance of the target and distractor (e.g., unrelated foil, similar lure) distributions. We used the function rocfit in Stata to compute A_z.

If the A-A′ test format enhances a participant's ability to rely on familiarity, then we should observe significantly better performance on the A-A′ test format (i.e., proportion correct) than on the old/new test format (i.e., A_z). Conversely, if performance on the A-A′ test format and the old/new test format rely on similar cognitive processes, then we should not observe a difference between proportion correct on the A-A′ format and A_z from the old/new format. Similarly, if the A-X test format enhances a participant's ability to rely on familiarity, then we should observe significantly better performance on the A-X test format than on the old/new test format.

The difference between A_z for targets versus unrelated foils and proportion correct on the A-X test format failed to reach significance (A_z: $M = 0.946$, $SD = 0.035$; proportion correct A-X: $M = 0.959$, $SD = 0.057$; $t_{35} = -0.76$, $p = .45$, 95% CI [−0.045, 0.020]; Figure 3A). Additionally, the difference between A_z for targets versus similar lures and proportion correct on the A-A′ test format failed to reach significance (A_z: $M = 0.860$, $SD = 0.070$; proportion correct A-A′: $M = 0.836$. $SD = 0.065$; $t_{35} = 1.11$, $p = .28$, 95% CI [−0.020, 0.070]; Figure 3B). Taken together, these results suggest that the old/new test format and the forced-choice test format recruit similar cognitive processes. In contrast, A_z for targets versus similar lures was significantly greater than proportion correct on the A-B′ test format (proportion correct A-B′ = 0.756 ± 0.088 [mean, standard deviation], $t_{35} = 3.96$, $p < .001$, 95% CI [0.051, 0.158]). While there were minor differences between the stimulus sets used in these two experiments, these results are at least consistent with the notion that that the old/new test format with targets versus similar lures is more closely related to the A-A′ test format. Finally, for comparison to the present experiments, a paired t test revealed significantly greater A_z for targets versus unrelated foils than A_z for targets versus similar lures ($t_{16} = 6.93$, $p < .001$, 95% CI [0.060, 0.112]).

Discussion

In Experiment 2, we replicated the effect of test format in a group of younger adults. Moreover, we ruled out the possibility that impaired performance on the A-B′ test format was driven by the presence of the A-X test trials embedded in the task. Accordingly, the results from Experiments 1 and 2 provide consistent evidence for an effect of test format. Moreover, this effect was present in both younger adults and healthy older adults. We next investigated whether performance on the old/new test format with

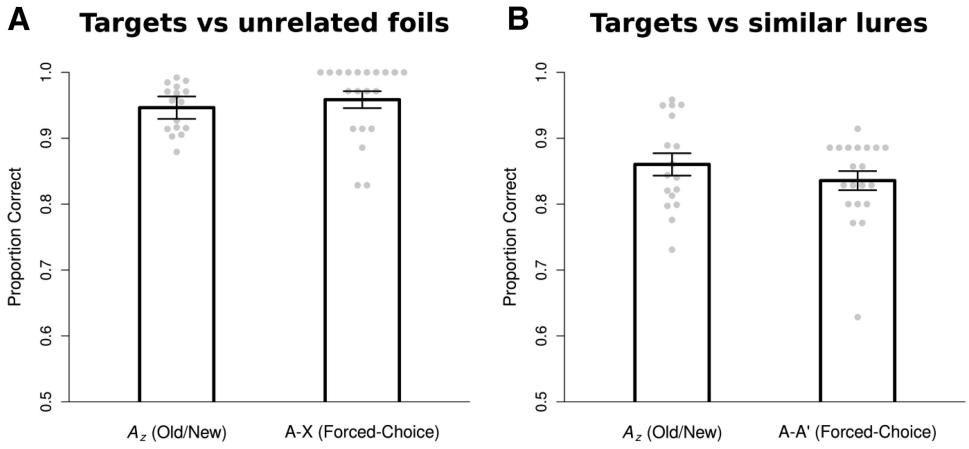

Figure 3. Between-groups performance was similar on forced-choice and old/new test formats. (A) A_z and two-alternative forced-choice performance were similar for targets versus unrelated foils ($t_{35} = -0.76$, $p = .45$). (B) A_z and two-alternative choice performance were similar for targets versus similar lures ($t_{35} = 1.11$, $p = .28$). In contrast, A_z for targets versus similar lures was significantly greater than performance on the A-B′ test format ($t_{35} = 3.96$, $p < .001$). The bar plots represent the mean proportion correct; the error bars represent the standard error of the mean; and the gray dots represent individual participants.

targets and similar lures more closely resembles performance on the A-A′ test format or the A-B′ test format.

Performance on two-alternative forced-choice tests is mathematically equivalent to the area under the ROC curve from old/new tests with confidence ratings (Green & Moses, 1966; Green & Swets, 1966; Stanislaw & Todorov, 1999; Swets & Pickett, 1982). We reanalyzed published data from our lab (Experiment 4 in the work of Stark et al., 2015) to examine whether there were differences in performance between the forced-choice test format and the old/new test format with confidence ratings. Specifically, if the forced-choice test format allowed participants to rely on familiarity to a greater extent than the old/new test format (Holdstock et al., 2002; Migo et al., 2009, 2014; Norman & O'Reilly, 2003), then we should observe significantly better performance on the forced-choice test format than A_z from the old/new test format. Conversely, if the two test formats rely on similar cognitive processes, then we should not observe a difference between the two test formats.

There was no evidence for a difference in performance between the old/new test format and the forced-choice test format for the discrimination between targets and unrelated foils (i.e., A-X) or for the discrimination between targets and similar lures (i.e., A-A′). We observed significantly worse performance on the A-B′ test format than the old/new test format with targets and similar lures. Taken together, these results suggest that the forced-choice format does not improve the discrimination between targets and unrelated foils nor the discrimination between targets and similar lures. We note that there were minor differences between the stimulus sets used in these experiments; however, the results are at least consistent with the notion that forced-choice formats rely on the same cognitive processes and do not receive familiarity-related enhancements in performance (cf. Khoe, Kroll, Yonelinas, Dobbins, & Knight, 2000; Bayley et al., 2008; but see Jeneson et al., 2010). Previous studies have shown similar performance on the A-X test format and the old/new test format with targets and unrelated foils (Green & Moses, 1966; Khoe et al., 2000; Smith & Duncan, 2004). Bayley et al. (2008) showed that performance was similar on the A-A′ test format and the old/new test format with targets and similar lures in patients with selective hippocampal damage and in healthy control participants (but see Jeneson et al., 2010).

Global Matching Models

Previous reports have shown that a class of models from mathematical psychology, global matching models, can account for better performance on the A-A′ test format than the A-B′ test format (Hintzman, 1988, 2001; also see Clark & Gronlund, 1996). Accordingly, we were interested in investigating whether global matching models could be used to account for our results in both younger and healthy older adults. We tested the hypothesis that decreasing the probability of successful feature encoding would cause a similar pattern of results to the empirical data in healthy older adults using two examples of this class of model: MINERVA 2 and TODAM. Our goal here is not to advocate for or against these models writ large, but to understand how this general class of memory models might account for the observed results.

Method

MINERVA 2. MINERVA 2 (Hintzman, 1984, 1988) is a member of a class of mathematical psychology models referred to as *global matching models*. MINERVA 2 is a multiple-trace or exemplar-based model, meaning that a new memory trace is added to an existing memory matrix every time that an item is encoded. In MINERVA 2, items are represented as vectors, each feature of which is set to -1, 0, or 1 with equal probability (i.e., 1/3). Similar lures were generated for each target by redrawing from the original features with probability δ. In the present report, we used $\delta = 0.16$, meaning that on average 16% of the features were redrawn from the original distributions. This resulted in approximately 11% of the features changing values. During encoding, each feature is encoded with probability of L and not encoded with probability $1 - L$. The encoding phase results in a memory matrix, T, which contains M rows (i.e., memory traces) and N columns (i.e., features). Our implementation relied on the equations presented in (Hintzman, 1984, 1988) and our simulations used $M = 35$, similar to our empirical test formats. The first equation provides an estimate of the similarity of a probe (p; i.e., a test item) to a given trace (T_i; i.e., one of the items in memory):

$$s_i = \left(\frac{p \cdot T_i}{n_i}\right)^3 \tag{1}$$

where n_i is the number of features that are relevant to the comparison of the probe and a given trace (a feature is relevant if it is nonzero in either p or T_i). Thus, the portion of the equation within the parentheses is a normalized dot product. The cubing function causes the similarity function to be nonlinear, which allows retrieval to be "quite selective" (Hintzman, 1984, 1988). The global match, g, of the trace is given by the summed similarity across all stored traces (where there are M traces in the memory matrix):

$$g = \sum_{i=1}^{M} s_i. \tag{2}$$

While MINERVA 2 uses a multiple-trace storage operation, the retrieval operation is the global match of a probe to all of the contents in memory. Thus, MINERVA 2 is a global matching model by the nature of its retrieval process. We modeled MINERVA 2 in R.

TODAM. TODAM (Theory of Distributed Associative Memory; Murdock, 1982) is a different global matching model. In contradistinction to MINERVA 2, which is a multiple-trace or exemplar-based model, TODAM is a distributed or prototype-based memory model, meaning that memories are stored in a single, composite memory vector (e.g., a prototype). Thus, while these models share the assumption that memory retrieval is a global matching process, the memory storage mechanisms of the models are very different. While most versions of TODAM have focused on associative memory tasks (e.g., Murdock, 1982), it can also be used as an item-only model (e.g., Murdock, 1995). Our implementation relied on the version of TODAM presented by Murdock (1995).

As in MINERVA 2, items are represented as vectors. In TODAM, each feature of an item vector is a random draw from a normal distribution with mean 0 and standard deviation $\sqrt{1/N}$, where N is the number of features. Occasionally the numerator is set to a value other than 1 (this parameter is referred to as P by

Murdock, 1982); however, setting the value to 1 causes the vectors to be of approximately unit length, which is useful because similarity is calculated using the dot product (i.e., the dot product between two vectors of unit length is between −1 and 1, similar to a normalized dot product). The following equation was used to generate a similar lure item ($f_{j'}$) for a given target item (f_j; Murdock, 1995):

$$f_{j'} = \rho f_j + (\sqrt{1-\rho^2}) g_j \quad (3)$$

where ρ represents the similarity of items to each other and g_j represents an independent random vector. The expected value of the similarity, defined as the dot product, between a target item and its similar lure is ρ. The memory vector for the item-only version of TODAM was calculated with the following equation (Kahana, 2012, p. 105; Murdock, 1995):

$$m_t = \alpha m_{t-1} + B_t f_t \quad (4)$$

where α is a forgetting parameter (which can also be thought of as a retention parameter because 0 represents complete erasure of previous memories, whereas 1 indicates that the new memory is added to the memory vector from the previous trial without any forgetting); m_{t-1} represents the memory vector from the previous trial; and f_t represents the item that is presented at time t in the encoding phase. B_t is a diagonal matrix with entries drawn from a Bernoulli distribution with probability p, where p represents the probability that a feature is encoded (Kahana, 2012, p. 105)—that is, each feature is encoded with probability p and not encoded with probability $1 - p$ (Murdock, 1995). Accordingly, p is isomorphic to L in MINERVA 2. As implied by the subscript t, B_t is trial unique. For item memory, the model has four parameters: (a) α, the forgetting/retention rate; (b) N, the number of features in each item; (c) p, the probability of encoding a feature; and (d) ρ, the similarity between a target item and its lure. The purpose of α is to emphasize recent items relative to items that were presented further in the past (cf. Kahana, 2012, p. 105). To preserve similarity to MINERVA 2 (and because we did not investigate list position effects), α was set to 1. Thus, in our application, both models have three parameters: (a) the number of features, (b) the probability of encoding a feature, and (c) the similarity between targets and similar lures. Additionally, we used a list length of 35 items as in our MINERVA 2 simulations and in our empirical test formats. The global match, g, of a probe to the memory vector was calculated with the following equation:

$$g = p \cdot m \quad (5)$$

where p represents a probe item, and m represents the memory vector. In contrast to MINERVA 2 (Equation 1), the item-only version of TODAM uses a linear similarity function. Also, because TODAM uses a single, composite memory vector, the global match is simply defined as the similarity—that is, the dot product—between the probe and the memory vector. Thus, TODAM is a global matching model by the nature of both its storage and its retrieval operations. The standard instantiations of TODAM use closed-form equations to calculate measures such as d'; however, we were interested in the effect of test format on forced-choice performance. Thus, we used a computational, rather than a mathematical, approach. We modeled TODAM in GNU Octave.

Simulation of forced-choice performance. As in our empirical data, we were interested in simulating performance from three different test formats: (1) A-X, (2) A-A′, and (3) A-B′. To simulate the A-X test format, we calculated the proportion of times that the global match, g (Equations 2 and 5), for a target item (A) exceeded that of an unrelated foil (X), using the following equation (cf. Hintzman, 1988):

$$Pr\{A > X\} = \frac{1}{M}\sum_{i=1}^{M} [I(g_{A_i} > g_{X_i}) + 0.5 \cdot I(g_{A_i} = g_{X_i})] \quad (6)$$

where M is the list length and $I(\cdot)$ is the indicator function that sets the value to 1 if the statement is true and to 0 otherwise. The second part of the equation simulates random guessing if the two items generate the same global match. To simulate the A-A′ test format, we calculated the proportion of times that the global match for a target item (A) exceeded that of its lure item (A′) using Equation 6. Similarly, to simulate the A-B′ test format, we calculated the proportion of times that the global match for a target item (A) exceeded that of a similar lure item from a different pair (B′) using Equation 6. In both models, we simulated 10,000 participants and we found parameter values that provided a good fit to the empirical data for the younger adults. To test the hypothesis that healthy aging is accompanied by impaired encoding, we investigated the effect of decreasing the probability of successfully encoding each feature in both models (parameters L and p in MINERVA 2 and TODAM, respectively).

Results

MINERVA 2. We investigated whether MINERVA 2 could account for our empirical findings of Experiment 1 in both younger and older adults. Specifically, we tested the hypothesis that aging could be modeled as a decreased probability of accurately encoding stimulus features. Such an account would be consistent with a number of neurocognitive models of aging that stress a role for the degradation of the medial temporal lobe associated with aging (for a review, see Stark & Stark, in press). We began by finding model parameters that achieved similar values to the mean values of our empirical data in younger adults ($N = 20$, $L = 0.65$, and $\delta = 0.16$; Figure 1B). We next investigated whether decreasing the encoding parameter, L, would cause a similar pattern of deficits as we observed in healthy older adults. We incrementally decreased the encoding parameter until the model achieved similar performance to older adults on the A-A′ test format, which was the format with a significant age group difference in the empirical data. We found that $L = 0.55$ met this condition. We then investigated performance on the A-X and the A-B′ test format using $L = 0.55$ and we found that the differences in performance relative to the $L = 0.65$ model were smaller than those observed for the A-A′ test format (Figure 1B). Thus, at least for certain model parameters, MINERVA 2 can predict a disproportionate change in performance on the A-A′ test format by simply changing the L parameter. These results support the hypothesis that healthy aging is accompanied by an impaired ability to encode stimulus features. We next asked whether a different global matching model, which relies on a very different storage mechanism, would predict a similar pattern of results.

TODAM. As in the MINERVA 2 simulation, we began by finding parameters that achieved similar values to the mean values of our empirical data in younger adults ($N = 400$, $p = .5$, and ρ =

0.7; Figure 1B). We next investigated whether decreasing the encoding parameter, p, would cause a similar pattern of deficits as we observed in healthy older adults. We incrementally decreased the encoding parameter until the model achieved similar performance to older adults on the A-A′ test format, which was the format with a significant age group difference in the empirical data. We found that $p = .35$ met this condition. We then investigated performance on the A-X and the A-B′ test format using $p = .35$ and we found that the differences in performance relative to the $L = 0.5$ model were smaller than those observed for the A-A′ test format (Figure 1B). Thus, at least for certain model parameters, TODAM can predict a disproportionate change in performance on the A-A′ test format by simply changing the p parameter. Given that MINERVA 2 and TODAM use very different storage mechanisms, these results provide additional support for both the global matching framework and for the hypothesis that healthy aging is accompanied by an impaired ability to encode item features.

Why do the models predict better performance on the A-A′ test format than the A-B′ test format? Previous reports showed that both MINERVA 2 and TODAM predict better performance on the A-A′ test format relative to the A-B′ test format (Clark & Gronlund, 1996; Hintzman, 1988, 2001). As discussed by Hintzman (1988, 2001), variability increases the overlap between target and distractor (e.g., similar lure, unrelated foil) distributions in MINERVA 2. One source of variability in MINERVA 2 (and TODAM) is encoding variability. In the standard version of MINERVA 2, each feature is encoded with probability L; hence, on average $L \times N \times 2/3$ nonzero features are encoded for each item, where N is the total number of features and it is multiplied by 2/3 because on average one third of the features are equal to zero. Because the number of encoded nonzero features is variable, there are trials where the number of nonzero features that are encoded is greater than $L \times N \times 2/3$ and trials where the number of nonzero features that are encoded is less than $L \times N \times 2/3$.

We hypothesized that removing trial-by-trial encoding variability in MINERVA 2 would reduce the A-A′ test format advantage. We tested this hypothesis by altering the model to encode a fixed number of features on each trial (note, a similar approach would be more difficult in TODAM because the features are drawn from a normal distribution rather than from $\{-1, 0, 1\}$). First, we set the number of nonzero features to be equal on each trial. In this version of the model we increased N from 20 to 21 to allow an equal number of -1, 0, and 1 features (i.e., seven each) and we set $L = 9/14$. We verified that this had no effect on performance of the model (see "With Encoding Variability" in Figure 4). Next, we eliminated encoding variability by forcing the model to encode 9 of the 14 nonzero features. Thus, the only difference between these two models is the presence of encoding variability. We observed an increase in proportion correct for all formats, and we observed a reduction of the A-A′ test format advantage over the A-B′ test format (see Figure 4). These results suggest that one possible reason for worse performance on the A-B′ test format relative to the A-A′ test format is that for some trials participants happen to encode more features for the original B item than the original A item. Because the lures are correlated with the original target item, this results in greater summed similarity for the B′ item than the A item. Under the condition in which there is not variability in the number of features that are encoded for each A and B item, there is less of a difference in performance between the A-A′ test format and the A-B′ test format.

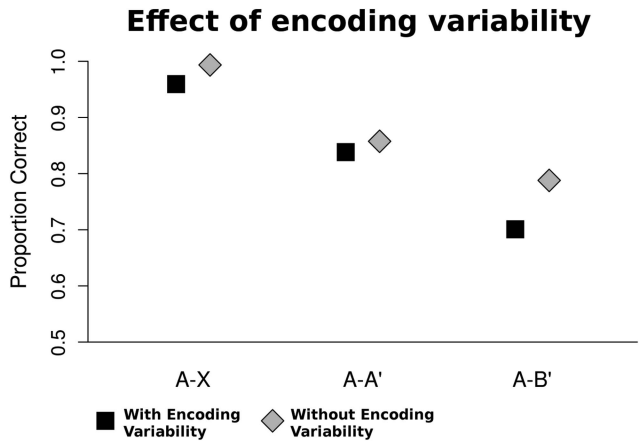

Figure 4. The removal of encoding variability in MINERVA 2 reduced the magnitude of the difference between proportion correct on the A-A′ format and the A-B′ format. Data points represent the mean proportion correct.

Encoding variability reduced the A-A′ test format advantage but it did not eliminate the advantage. As discussed by Hintzman (1988, 2001), there are other sources of variability that contribute to the A-A′ test format advantage. For example, within the stimulus set used for the encoding phase, some stimuli happen to be more similar to other stimuli, which results in certain trials in which the B item more closely resembles other items in the encoding set than the A item. Because the model assumes that memory strength is determined by the match of the test item to all of the contents of memory, this results in a greater global match of the B′ item than the A item (Hintzman, 1988, 2001). Interestingly, in our simulations, we found that list length modulated the strength of the effect of encoding variability on the A-A′ test format advantage. Specifically, for shorter list lengths (e.g., four items), the elimination of encoding variability accounted for more of the difference between the two test formats than for longer list lengths (e.g., 35 items). In fact, for short list lengths, the elimination of encoding variability was sufficient to nearly eliminate the difference between the two test formats, suggesting that as more items are encoded there is a greater chance of a B′ item providing a better global match than the A item (due to similarity to other items in the stimulus set). Thus, MINERVA 2 suggests that there are a number of potential sources of variability that contribute to enhanced A-A′ test format performance compared to the A-B′ test format, including encoding variability and the similarity between items in the study list.

Discussion

Similar to Hintzman (1988, 2001), we showed that MINERVA 2 could account for the observed effect of test format. Moreover, we showed that a different global matching model, TODAM (Murdock, 1982, 1995), can also account for the observed effect of test format (cf. Clark & Gronlund, 1996). We used MINERVA 2 to provide a possible explanation of the A-A′ test format advantage (also see Hintzman, 1988, 2001). Specifically, we showed that

removing trial-by-trial encoding variability reduced the magnitude of the A-A′ test format advantage. Thus, the model suggests that one possible reason for the A-A′ test format advantage is that there are trials on which a participant happens to encode more details than other trials, which causes certain lures (B′) to contain a stronger global match than a noncorresponding target item (A)—that is, the global match for the lures is shifted along with the global match of the target item due to the similarity between them. Additionally, we found that the list length contributed to the effect of encoding variability, such that the elimination of encoding variability had a larger effect for short list lengths. Specifically, for short lists, the removal of encoding variability nearly eliminated the A-A′ test format advantage. Our simulations with longer list lengths suggested that there are more trials in which the B item (and by extension the B′ item) is more similar to other items in the encoding set than the A item. Accordingly, both encoding variability and variability in between-items similarity in the encoding list could contribute to better performance on the A-A′ test format than the A-B′ test format.

After finding parameters that provided a good fit of the empirical data in younger adults, we investigated whether decreasing the probability of encoding stimulus features would cause a similar pattern of results to the empirical data in healthy older adults. In MINERVA 2 and TODAM, decreasing the encoding probability caused the largest change in performance on the A-A′ test format, which was the test format on which we observed an age-related change. It is noteworthy, however, that both models predicted a change on the other test formats as well, which suggests that the most sensitive test format for detecting differences in encoding was the A-A′ test format. MINERVA 2 and TODAM rely on very different assumptions regarding how memories are stored—namely, MINERVA 2 is a multiple-trace model while TODAM is a distributed memory model. The fact that both models predicted the largest change on the A-A′ test format as a result of decreasing the encoding parameter supports the global matching framework and suggests that a possible explanation for the observed age-related changes is a decrease in the probability of encoding stimulus features.

General Discussion

The Effect of Test Format on Performance

We investigated the effect of test format on recognition memory performance in younger and healthy older adults. In Experiment 1, we used three test formats: (1) A-X, (2) A-A′, and (3) A-B′. In both age groups, performance was best on the A-X format, followed by the A-A′ format, followed by the A-B′' format. The results from Experiment 2 replicated the results of Experiment 1 and provided no evidence to suggest that the A-X test format artificially reduced performance on the A-B′ test format. Thus, we consistently observed better performance on the A-A′ test format than the A-B′ test format. The findings in younger adults replicate the effects from a study that used images of scenes (Tulving, 1981). Other reports have shown enhanced performance on the A-A′ test format compared to the A-B′ test format in young adults (Hintzman, 1988; but see Migo et al., 2009), healthy middle-aged/older adults (mean age: 61.2 years; Jeneson et al., 2010), healthy older adults (mean age 71 years; Migo et al., 2014), and in patients with selective hippocampal damage (Jeneson et al., 2010). Moreover, performance on the A-A′ test format has been shown to be better than performance on the A-B′ test format across a variety of encoding and stimulus conditions: single presentations of images of objects (Experiment 1 and Experiment 2) and of scenes (Tulving, 1981), multiple encoding trials of images of objects (color images: Jeneson et al., 2010; black and white silhouettes: Jeneson et al., 2010; Migo et al., 2014; but see Migo et al., 2009), and judgments of the number of times that words were presented during the encoding phase (Hintzman, 1988). As in previous reports, we found that two global matching models, MINERVA 2 and TODAM, can account for the effect of test format (Hintzman, 1988, 2001; cf. Clark & Gronlund, 1996). As we discussed above, the models predict that encoding variability and variability in between-items similarity in the encoding list could contribute to better performance on the A-A′ test format than the A-B′ test format (also see Hintzman, 1988, 2001).

Molitor, Ko, Hussey, and Ally (2014) used eye tracking to infer differences in encoding. Their results suggest that differences in the number fixations during encoding are predictive of subsequent false alarms to similar lures (using an old/similar/new test format). Future studies can use similar techniques to test the prediction from MINERVA 2 that encoding variability modulates the differences between the A-A′ test format and the A-B′ test format. For example, MINERVA 2 predicts that incorrect trials on the A-B′ test format would be associated with a lower A to B fixation ratio than correct trials (i.e., somewhat counterintuitively, better encoding of the original B item would lead to an increased tendency to select the B′ item at test due to a stronger global match). Future studies could also match the number of fixations between the A and B items and then compare performance to the A-A′ test format (also matching the number of fixations during encoding across the A-B′ test format and the A-A′ test format). This would address an untested prediction from the model that minimizing encoding variability (in particular, encoding differences between the A item and the B item) would diminish the differences in performance on the A-B′ test format relative to the A-A′ test format. It is important that such studies match the "lure bin" of the A and the B items on each test trial, as we have done here, to address the potential issue of encoding versus retrieval difficulty of stimuli from different lure bins (see Task design of Experiment 1).

Forced-Choice and Old/New Test Formats Reveal a Stable Age-Related Impairment of Performance on the MST

We investigated whether the age-related impairment on the old/new and old/similar/new versions of the MST (Bennett et al., 2015; Stark et al., 2013, 2015; Toner et al., 2009; Yassa, Lacy, et al., 2011; Yassa, Mattfeld, et al., 2011) would extend to the forced-choice test format. Our results revealed a significant Age × Test format interaction, which was driven by better performance in younger adults than healthy older adults on the A-A′ test format. These results suggest that the age-related impairment on the old/similar/new and the old/new test formats with targets and similar lures extends to the A-A′ test format. We were admittedly surprised that there was not a significant difference in performance on the A-B′ test format; however, we suggest that the results from our model-based approach and the results from Experiment 2 provide

possible explanations for the significant age-related difference on the A-A′ test format but not on the A-B′ test format, which we will discuss in turn.

The models predict that decreasing the probability of successful feature encoding would cause a decrease in the proportion correct on all three test formats; however, both models predict that the largest change would be on the A-A′ test format, similar to our empirical results in healthy older adults. These results suggest that one explanation for the observed age-related change in performance on the A-A′ test format but not the A-B′ test format is that the A-A′ test format is more sensitive to changes in the ability to encode stimulus features. Notably, the model "incorrectly" predicted a change on the other two test formats; however, it is possible that future studies could alter the sensitivity of the other test formats to reveal an age-related change in performance. For example, other reports have used four-alternative forced-choice test formats (Migo et al., 2009, 2014), which provide a larger dynamic range than the two-alternative format. Although performance on the A-B′ test format was above chance in both younger and older adults, it is possible that the four-alternative format would be more sensitive to detecting an age-related impairment. Additionally, larger sample sizes might be necessary to uncover age-related differences on the A-B′ test format.

Future studies could investigate list position effects in younger and healthy older adults. Such studies could help elucidate whether changing the forgetting parameter in TODAM (or modifying MINERVA 2 to contain a forgetting parameter) would provide a better account of the data in healthy older adults. The forgetting parameter emphasizes recent items relative to items that were presented further in the past (cf. Kahana, 2012, p. 105). Notably, a previous study in our lab used a continuous recognition test (old/similar/new format) to investigate the effect of the number of intervening items on the ability to correctly respond "similar" to similar lure items (Experiment 2 in the work of Stark et al., 2015). The results of this experiment revealed a main effect of age (i.e., worse performance in healthy older adults), a main effect of lag (worse performance with more intervening items between the original encoding and the presentation of a lure item), but no sign of an interaction between lag and age group. These results suggest that there are age-related differences in encoding (given the main effect of age) but not age-related differences in the rate of forgetting or the rate interference (given the null interaction effect). Future studies that are optimized for detecting lag effects could be paired with simulations using TODAM (and a modified version of MINERVA 2 that incorporates a forgetting parameter) to test whether changes to the encoding parameter, the forgetting parameter, or to both parameters provide a better account of the data.

The results from the area under the ROC curve analysis in Experiment 2 can be brought to bear on our findings in healthy older adults. Specifically, it appears that the discrimination between targets and similar lures in the old/new test format most closely resembles the A-A′ test format. Previous studies from our lab and others have revealed an age-related impairment in the discrimination between targets and similar lures across a variety of test formats, including old/similar/new (Bennett et al., 2015; Stark et al., 2013, 2015; Toner et al., 2009; Yassa, Lacy, et al., 2011; Yassa, Mattfeld, et al., 2011), old/new (Stark et al., 2015), and old/new with confidence ratings (Stark et al., 2015). Moreover, these effects maintained across a variety of encoding conditions—for example, incidental encoding, intentional encoding, continuous recognition (Stark et al., 2015). The finding that performance in younger adults was similar between the old/new test format with targets and similar lures and the A-A′ test format suggests that the age-related change in performance on the old/new test format with targets and similar lures should be accompanied by an age-related impairment on the A-A′ test format, consistent with our results.

We have previously reported a relationship between age and similar-lure discrimination in a life-span sample that used the old/similar/new test format (Bennett et al., 2015; Stark et al., 2013). Similarly, in a group of healthy older adults, Migo et al. (2014) revealed a relationship between age and proportion correct on the A-A′ test format but not between age and proportion correct on the A-B′ test format. The converging findings across these studies (including Experiment 1) support the notion that the old/similar/new test format and A-A′ test formats are similarly affected by aging. Altogether, there is a stable age-related impairment in the ability to discriminate between targets and similar lures, which we argue is caused by a mnemonic rather than a decision-based difference between younger and healthy older adults. Indeed, the A-A′ test format eliminates any possible shifts in decision criterion across groups, thus obviating concerns raised by Loiotile and Courtney (2015).

Application of Global Matching Models to Interpret Other Studies That Used Variants on MSTs

We investigated whether global matching models could account for the results of other studies that have used mnemonic similarity tasks. Reagh and Yassa (2014) used a variant on the MST to investigate the effect of stimulus repetition on memory for images of objects. They reported that stimulus repetition—three presentations compared to one presentation—improved discrimination between targets and unrelated foils and increased the false alarm rate to similar lures (using an old/new test format). They concluded that repetition improves generalization while impairing mnemonic discrimination. Moreover, they suggested that stimulus repetition can induce competition between memory traces which would cause a loss of details from memory. Subsequently, Loiotile and Courtney (2015) used signal detection theory to show that while repetition increased the false alarm rate to similar lures, it also enhanced discrimination between targets and similar lures (as measured by d_a). They also showed that repetition improved performance on the A-A′ test format. The results from these studies were initially puzzling, and we were curious whether they could be accounted for within a global matching framework. To test this possibility, we modeled their tasks using MINERVA 2 and we found that it can account for the data from both studies. Specifically, MINERVA 2 predicts that repetition will cause: (a) better discrimination between targets and unrelated foils (as measured by an ROC analysis), (b) an increased false-alarm rate to similar lures (cf. Hintzman, 1988, 2001; Hintzman, Curran, & Oppy, 1992), (c) better discrimination between targets and similar lures (as measured by an ROC analysis), and (d) improved A-A′ test format performance.

The key insight from MINERVA 2 is that stimulus repetition increases the global match of similar lure items by increasing the number of traces that match the similar lure—that is, three traces of A will generate a larger global match in response to A′ than only a single trace. As a corollary, MINERVA 2 predicts that

encoding the same exact details of an item three times would also increase the false alarm rate to a similar lure, suggesting that an increased false alarm rate to similar lures does not necessarily indicate a loss of details from memory. Furthermore, while repetition increases the global match of the similar lure distribution, it also decreases the overlap between the target and similar lure distributions. Therefore, the model predicts that comparisons between the target distribution and the similar lure distribution will be more discriminable for items that are presented three times than items that are presented one time (i.e., based on an ROC analysis or performance on the A-A′ test format). Altogether, the findings from our simulations highlight the notion that formal models can be used to constrain the interpretation of behavioral results. Thus, although global matching models have been challenged by a number of findings (for a review, see Clark & Gronlund, 1996), we believe that they provide useful tools for interpreting the results of studies that manipulate stimulus similarity.

Conclusion

Previous research has shown that there are clear age-related impairments on tasks that tax recollection and associative memory with a more mild impairment on tests of simpler item-recognition memory (Craik & McDowd, 1987; Danckert & Craik, 2013; Naveh-Benjamin, 2000; Naveh-Benjamin et al., 2004; Old & Naveh-Benjamin, 2008a, 2008b; Spencer & Raz, 1995). Other studies have shown that healthy older adults reliably exhibit an impairment on item-recognition memory tests that require discriminating between targets and similar lures (Bennett et al., 2015; Stark et al., 2013, 2015; Toner et al., 2009; Yassa, Lacy, et al., 2011; Yassa, Mattfeld, et al., 2011). Our results suggest that healthy older adults are similarly impaired on the forced-choice discrimination between an object and its similar lure. Taken together, there is clear evidence that memory tests that require a high degree of fidelity are impaired in healthy older adults. Our modeling results suggest that healthy aging causes an impaired ability to encode stimulus features, which causes fewer details to be encoded on each trial. These results provide a potential mechanistic interpretation of previous results that does not emphasize differences in cognitive processes but instead emphasizes differences in the mnemonic resolution required to solve the task (cf. Cowell, Bussey, & Saksida, 2010).

References

Bayley, P. J., Wixted, J. T., Hopkins, R. O., & Squire, L. R. (2008). Yes/no recognition, forced-choice recognition, and the human hippocampus. *Journal of Cognitive Neuroscience, 20,* 505–512. http://dx.doi.org/10.1162/jocn.2008.20038

Bennett, I. J., Huffman, D. J., & Stark, C. E. L. (2015). Limbic tract integrity contributes to pattern separation performance across the lifespan. *Cerebral Cortex, 25,* 2988–2999. http://dx.doi.org/10.1093/cercor/bhu093

Clark, S. E., & Gronlund, S. D. (1996). Global matching models of recognition memory: How the models match the data. *Psychonomic Bulletin & Review, 3,* 37–60. http://dx.doi.org/10.3758/BF03210740

Cowell, R. A., Bussey, T. J., & Saksida, L. M. (2010). Components of recognition memory: Dissociable cognitive processes or just differences in representational complexity? *Hippocampus, 20,* 1245–1262. http://dx.doi.org/10.1002/hipo.20865

Craik, F. I. M., & McDowd, J. M. (1987). Age differences in recall and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 474–479. http://dx.doi.org/10.1037/0278-7393.13.3.474

Crum, R. M., Anthony, J. C., Bassett, S. S., & Folstein, M. F. (1993). Population-based norms for the Mini-Mental State Examination by age and educational level. *Journal of the American Medical Association, 269,* 2386–2391. http://dx.doi.org/10.1001/jama.1993.03500180078038

Danckert, S. L., & Craik, F. I. M. (2013). Does aging affect recall more than recognition memory? *Psychology and Aging, 28,* 902–909. http://dx.doi.org/10.1037/a0033263

Green, D. M., & Moses, F. L. (1966). On the equivalence of two recognition measures of short-term memory. *Psychological Bulletin, 66,* 228–234. http://dx.doi.org/10.1037/h0023645

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Hoboken, NJ: Wiley.

Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments & Computers, 16,* 96–101. http://dx.doi.org/10.3758/BF03202365

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review, 95,* 528–551. http://dx.doi.org/10.1037/0033-295X.95.4.528

Hintzman, D. L. (2001). Similarity, global matching, and judgments of frequency. *Memory & Cognition, 29,* 547–556. http://dx.doi.org/10.3758/BF03200456

Hintzman, D. L., Curran, T., & Oppy, B. (1992). Effects of similarity and repetition on memory: Registration without learning? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 667–680. http://dx.doi.org/10.1037/0278-7393.18.4.667

Holdstock, J. S., Mayes, A. R., Roberts, N., Cezayirli, E., Isaac, C. L., O'Reilly, R. C., & Norman, K. A. (2002). Under what conditions is recognition spared relative to recall after selective hippocampal damage in humans? *Hippocampus, 12,* 341–351. http://dx.doi.org/10.1002/hipo.10011

Jeneson, A., Kirwan, C. B., Hopkins, R. O., Wixted, J. T., & Squire, L. R. (2010). Recognition memory and the hippocampus: A test of the hippocampal contribution to recollection and familiarity. *Learning & Memory, 17,* 63–70. http://dx.doi.org/10.1101/lm.1546110

Kahana, M. J. (2012). *Foundations of human memory*. New York, NY: Oxford University Press.

Khoe, W., Kroll, N. E. A., Yonelinas, A. P., Dobbins, I. G., & Knight, R. T. (2000). The contribution of recollection and familiarity to yes-no and forced-choice recognition tests in healthy subjects and amnesics. *Neuropsychologia, 38,* 1333–1341. http://dx.doi.org/10.1016/S0028-3932(00)00055-5

Kirwan, C. B., Hartshorn, A., Stark, S. M., Goodrich-Hunsaker, N. J., Hopkins, R. O., & Stark, C. E. L. (2012). Pattern separation deficits following damage to the hippocampus. *Neuropsychologia, 50,* 2408–2414. http://dx.doi.org/10.1016/j.neuropsychologia.2012.06.011

Kirwan, C. B., & Stark, C. E. L. (2007). Overcoming interference: An fMRI investigation of pattern separation in the medial temporal lobe. *Learning & Memory, 14,* 625–633. http://dx.doi.org/10.1101/lm.663507

Lacy, J. W., Yassa, M. A., Stark, S. M., & Stark, C. E. L. (2011). Distinct pattern separation related transfer functions in human CA_3/dentate and CA_1 revealed using high-resolution fMRI and variable mnemonic similarity. *Learning & Memory, 18,* 15–18. http://dx.doi.org/10.1101/lm.1971111

Loiotile, R. E., & Courtney, S. M. (2015). A signal detection theory analysis of behavioral pattern separation paradigms. *Learning & Memory, 22,* 364–369. http://dx.doi.org/10.1101/lm.038141.115

Migo, E., Montaldi, D., Norman, K. A., Quamme, J., & Mayes, A. (2009). The contribution of familiarity to recognition memory is a function of test format when using similar foils. *Quarterly Journal of Experimental*

Psychology: Human Experimental Psychology, 62, 1198–1215. http://dx.doi.org/10.1080/17470210802391599

Migo, E. M., Quamme, J. R., Holmes, S., Bendell, A., Norman, K. A., Mayes, A. R., & Montaldi, D. (2014). Individual differences in forced-choice recognition memory: Partitioning contributions of recollection and familiarity. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 67,* 2189–2206. http://dx.doi.org/10.1080/17470218.2014.910240

Molitor, R. J., Ko, P. C., Hussey, E. P., & Ally, B. A. (2014). Memory-related eye movements challenge behavioral measures of pattern completion and pattern separation. *Hippocampus, 24,* 666–672. http://dx.doi.org/10.1002/hipo.22256

Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review, 89,* 609–626. http://dx.doi.org/10.1037/0033-295X.89.6.609

Murdock, B. B. (1995). Similarity in a distributed memory model. *Journal of Mathematical Psychology, 39,* 251–264. http://dx.doi.org/10.1006/jmps.1995.1026

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. http://dx.doi.org/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. http://dx.doi.org/10.1037/0882-7974.19.3.541

Norman, K. A. (2010). How hippocampus and cortex contribute to recognition memory: Revisiting the complementary learning systems model. *Hippocampus, 20,* 1217–1227. http://dx.doi.org/10.1002/hipo.20855

Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. *Psychological Review, 110,* 611–646. http://dx.doi.org/10.1037/0033-295X.110.4.611

Old, S. R., & Naveh-Benjamin, M. (2008a). Differential effects of age on item and associative measures of memory: A meta-analysis. *Psychology and Aging, 23,* 104–118. http://dx.doi.org/10.1037/0882-7974.23.1.104

Old, S. R., & Naveh-Benjamin, M. (2008b). Memory for people and their actions: Further evidence for an age-related associative deficit. *Psychology and Aging, 23,* 467–472. http://dx.doi.org/10.1037/0882-7974.23.2.467

Reagh, Z. M., Ho, H. D., Leal, S. L., Noche, J. A., Chun, A., Murray, E. A., & Yassa, M. A. (2016). Greater loss of object than spatial mnemonic discrimination in aged adults. *Hippocampus, 26,* 417–422. http://dx.doi.org/10.1002/hipo.22562

Reagh, Z. M., & Yassa, M. A. (2014). Repetition strengthens target recognition but impairs similar lure discrimination: Evidence for trace competition. *Learning & Memory, 21,* 342–346. http://dx.doi.org/10.1101/lm.034546.114

Rey, A. (1941). Lexamen psychologique dans les cas dencephalopathie traumatique. *Archives de Psychologie, 28,* 286–340.

Smith, D. G., & Duncan, M. J. J. (2004). Testing theories of recognition memory by predicting performance across paradigms. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 615–625. http://dx.doi.org/10.1037/0278-7393.30.3.615

Spencer, W. D., & Raz, N. (1995). Differential effects of aging on memory for content and context: A meta-analysis. *Psychology and Aging, 10,* 527–539. http://dx.doi.org/10.1037/0882-7974.10.4.527

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments & Computers, 31,* 137–149. http://dx.doi.org/10.3758/BF03207704

Stark, S. M., & Stark, C. E. L. (in press). The aging hippocampus: Linking animal and human research. In R. Cabeza, L. Nyberg, & D. Park (Eds.), *Cognitive neuroscience of aging* (2nd ed.) New York, NY: Oxford University Press.

Stark, S. M., Stevenson, R., Wu, C., Rutledge, S., & Stark, C. E. L. (2015). Stability of age-related deficits in the mnemonic similarity task across task variations. *Behavioral Neuroscience, 129,* 257–268. http://dx.doi.org/10.1037/bne0000055

Stark, S. M., Yassa, M. A., Lacy, J. W., & Stark, C. E. L. (2013). A task to assess behavioral pattern separation (BPS) in humans: Data from healthy aging and mild cognitive impairment. *Neuropsychologia, 51,* 2442–2449. http://dx.doi.org/10.1016/j.neuropsychologia.2012.12.014

Swets, J. A., & Pickett, R. M. (1982). *Evaluation of diagnostic systems*. New York, NY: Academic Press.

Toner, C. K., Pirogovsky, E., Kirwan, C. B., & Gilbert, P. E. (2009). Visual object pattern separation deficits in nondemented older adults. *Learning & Memory, 16,* 338–342. http://dx.doi.org/10.1101/lm.1315109

Tulving, E. (1981). Similarity relations in recognition. *Journal of Verbal Learning & Verbal Behavior, 20,* 479–496. http://dx.doi.org/10.1016/S0022-5371(81)90129-8

Yassa, M. A., Lacy, J. W., Stark, S. M., Albert, M. S., Gallagher, M., & Stark, C. E. (2011). Pattern separation deficits associated with increased hippocampal CA3 and dentate gyrus activity in nondemented older adults. *Hippocampus, 21,* 968–979.

Yassa, M. A., Mattfeld, A. T., Stark, S. M., & Stark, C. E. L. (2011). Age-related memory deficits linked to circuit-specific disruptions in the hippocampus. *Proceedings of the National Academy of Sciences of the United States of America, 108,* 8873–8878. http://dx.doi.org/10.1073/pnas.1101567108

Received August 5, 2016
Revision received November 9, 2016
Accepted November 18, 2016 ■