Mem Cogn (2015) 43:760–774
DOI 10.3758/s13421-014-0499-6

Retrieval practice and spacing effects in young and older adults: An examination of the benefits of desirable difficulty

Geoffrey B. Maddox · David A. Balota

Published online: 24 January 2015

Abstract In the present study, we examined how the function relating continued retrieval practice (e.g., one, three, or five tests) and long-term memory retention is modulated by desirable difficulty (R. A. Bjork, 1994). Of particular interest was how retrieval difficulty differed across young and older adults and across manipulations of lag (Exp. 1) and spacing (Exp. 2). To extend on previous studies, the acquisition phase response latency was used as a proxy for retrieval difficulty, and our analysis of final-test performance was conditionalized on acquisition phase retrieval success, to more directly examine the influence of desirable difficulty on retention. The results from Experiment 1 revealed that continued testing in the short-lag condition led to consistent increases in retention, whereas continued testing in the long-lag condition led to increasingly smaller benefits in retention for both age groups. The results from Experiment 2 revealed that repeated spaced testing enhanced retention relative to taking one spaced test, for both age groups; however, repeated massed testing only enhanced retention over taking one test for young adults. Across both experiments, the response latency results were overall consistent with an influence of desirable difficulty on retention. The discussion focuses on the role of desirable difficulty during encoding in producing the benefits of lag, spacing, and testing.

Keywords Testing effect · Spacing effect · Retrieval practice aging · Refreshing

G. B. Maddox · D. A. Balota
Washington University in St. Louis, St. Louis, MO, USA

G. B. Maddox (✉)
Rhodes College, 110 Clough Hall, Memphis, TN 38112, USA
e-mail: maddoxg@rhodes.edu

Healthy aging is marked by broad declines in episodic memory (Arking, 1998; Balota, Dolan, & Duchek, 2000; Salthouse, 1996). Given the substantial increase in our aging population, there is a clear need to identify ways of improving memory that are effective across diverse populations and variable aging trajectories (e.g., Hertzog, Kramer, Wilson, & Lindenberger, 2009). One technique that is effective across varying contexts and populations, *spaced retrieval practice*, combines the mnemonic benefits of spacing and testing (see Balota, Duchek, & Logan, 2007, and Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006, for reviews). Indeed, spaced retrieval practice has been shown to enhance performance in healthy older adults (e.g., Balota, Duchek, & Paullin, 1989), individuals with Alzheimer's disease (e.g., Balota, Duchek, Sergent-Marshall, & Roediger, 2006; Camp, Foss, Stevens, & O'Hanlon, 1996), and individuals suffering from amnesia (Schacter, Rich, & Stamp, 1985).

Given the effectiveness of spaced retrieval, it is important to better understand why this technique improves memory and how it can be used most effectively. Multiple mechanisms have been developed to account for spaced retrieval, including the combined study-phase retrieval and encoding variability account (e.g., Greene, 1989; Raaijmakers, 2003; see Cepeda et al., 2006, for a review). Here, we focus on one account of the benefits of spaced retrieval, *desirable difficulty* (R. A. Bjork, 1994), which suggests that more effortful retrieval will lead to greater strengthening in the underlying memory trace than less effortful retrieval assuming that successful retrieval occurs in both situations. Specifically, spaced practice produces better performance than massed practice because the second retrieval event in the spaced condition is relatively more difficult thereby producing better retention. The present study extends on previous examinations of desirable difficulty in two ways. First, we operationally define desirable difficulty in a way that utilizes response latencies on retrieval trials during the acquisition phase. Second, we examine the influence of desirable difficulty on *retention* by accounting for differences across conditions in acquisition performance that may later influence final-test performance.

With these methodological extensions in hand, we can more carefully address the following two questions: First, how does the efficiency of retrieval practice differ across healthy young and older adults given past research indicating age differences in optimal spacing schedules (e.g., Maddox, Balota, Coane, & Duchek, 2011)? Second, how can the efficiency of retrieval practice be maximized by varying the spacing interval (i.e., lag) that occurs between retrieval attempts? In other words, how does the relationship between long-term memory performance and additional retrieval practice (e.g., one vs. three vs. five tests) differ as a function of lag?

Before introducing the present experiments, we more fully consider the two methodological extensions in the present study. We will first examine how desirable difficulty may influence final-test performance in a retrieval practice paradigm and will discuss the way desirable difficulty has often been operationally defined. We will then consider how spaced retrieval practice may differentially influence encoding versus retention of material.

Spaced retrieval practice and desirable difficulty

As we noted earlier, the desirable difficulty account of the spacing effect (R. A. Bjork, 1994) suggests that performance in the spaced condition benefits more than the massed condition from successful retrieval, because retrieval events in the former condition are more difficult than retrieval events in the latter condition. Similarly, the desirable difficulty account can help explain the more general benefit of testing over restudying (see Rowland, 2014, for a review) and the benefit of spaced study over massed study when incorporated with other proposed mechanistic accounts (e.g., the study-phase retrieval and reminding accounts; see Benjamin & Tullis, 2010, for a review).

One concern with considering past spaced retrieval studies as assessments of the desirable difficulty account is how retrieval difficulty during the encoding phase is operationally defined. The retrieval difficulty of a condition is typically inferred on the basis of the lag separating study and test events (i.e., retrieval following a short lag is easier than retrieval following a long lag; e.g., E. L. Bjork & Bjork, 2011; R. A. Bjork, 2013; Clark & Bjork, 2014; Pyc & Rawson, 2009) or utilizing other experimental manipulations hypothesized to induce different levels of difficulty during the encoding process (e.g., Sungkhasettee, Friedman, & Castel, 2011; Yue, Castel, & Bjork, 2013).

In studies that have examined the benefit of spaced retrieval practice, two studies have examined acquisition phase response latencies in addition to cued recall (Karpicke & Roediger, 2007; Logan & Balota, 2008). The results from both studies revealed slower mean response latency on the first retrieval attempt following a long lag relative to a short lag, which suggests that it may be particularly important to introduce desirable difficulty on the first retrieval attempt as a means for enhancing long-term memory performance. Thus, there is precedent for using response latencies as a proxy for retrieval difficulty, but it is not true that retrieval difficulty must always be correlated with the spacing interval separating retrieval attempts. In some instances, the relationship may actually be curvilinear. For example, the difference in retrieval difficulty between lag 1 and lag 5 conditions may be less than the difference between lag 5 and lag 9 conditions, even though the objective change in lag size is constant. This might occur if lag 1 and lag 5 intervals are within an individual's working memory capacity, but lag 9 is beyond their working memory capacity (cf. Bui, Maddox, & Balota, 2013). Similarly, if multiple populations are examined in a single study (as in the present experiments), retrieval difficulty and the difference in difficulty between spacing conditions may shift across groups. In the present study, older adults may experience a larger difference in retrieval difficulty between lag conditions than do young adults, given age-related changes in episodic memory (see Balota et al., 2000). Hence, one might expect larger benefits of spacing in older than in younger adults.

Thus, it is critical to jointly examine acquisition phase response latency and accuracy to more directly assess the influence of desirable difficulty on long-term memory for the specific lags and populations examined in a given study. Although the manipulation of lag as a proxy for changes in desirable difficulty appears a priori to be a reasonable assumption, it is also critical to have an independent measure of this construct. Therefore, in the present study, we used acquisition phase response latency as a proxy for retrieval difficulty, such that longer response latencies would indicate greater retrieval difficulty than shorter response latencies.

The influence of spaced retrieval practice on encoding versus retention

Typically, past research has emphasized the extent to which spaced retrieval maximizes final-test performance. However, differences during acquisition performance are typically observed across spaced retrieval conditions as well as differences in final-test performance, which complicates the understanding of how spaced retrieval influences retention and final retrieval above and beyond its influence on the encoding process. In order to illustrate this issue, consider the hypothetical situation in which ten items are assigned to a short-lag condition and ten items are assigned to a long-lag condition. If six items are retrieved on the final test from the long-lag condition and four items are retrieved from the short-lag condition, one might argue that spaced retrieval practice produced a 20% benefit in benefit on final recall performance (60% vs. 40% correct). However, if during acquisition, participants correctly retrieved all ten items in the short-lag condition but only eight items correct in the long-lag condition, then the benefit of retrieval practice on *retention* would be 35% (6/8 items, or 75% in the long-lag condition vs. 4/10 items or 40% in the short-lag condition). This, of course, assumes that there is little if any hypermnesia (see Erdelyi & Becker, 1974; Payne & Roediger, 1987). A similar concern may also arise when comparing age groups. If young adult memory is more intact than older adult memory, one would predict a larger difference in accuracy between spacing conditions for older adults than for young adults during the acquisition phase due to age-related changes in episodic memory (see Balota et al., 2000). If one does not account for these differences during acquisition, it may appear that the benefit of spaced retrieval practice differs across age groups.

With these points in mind, in the present study we emphasized conditional final-test performance. In doing so, we aimed to minimize the influence of acquisition phase differences (between spacing conditions and between age groups) on final-test performance to better isolate the influence of spaced retrieval on retention.

Present study

Armed with the two methodological extensions to past spaced retrieval studies (i.e., conditional final recall performance and measuring response latency during acquisition), in Experiment 1 we examined the benefits of retrieval practice for items that were retrieved once, three times or five times either at a short or long lag. Figure 1 displays a partial-list structure in which both spacing and retrieval practice are within-participants manipulations.

On the basis of prior studies and the assumptions described earlier relating retrieval difficulty and spacing intervals, one would expect a priori that continued testing should produce a benefit in final-test performance in the long-lag condition, but not in the short-lag condition. This assumes that repeated retrieval with a more difficult, longer lag should produce continued benefits, whereas little benefit should be gained from additional retrieval practice with the easier, shorter lag. Importantly, however, the present study afforded a direct measure of retrieval difficulty (i.e., response latency), instead of simply assuming differences in retrieval difficulty across conditions. As we will see, this additional measure provides important insights into the retrieval difficulty encountered during the acquisition phase in each lag condition.

Importantly, in the present study we also compared young and older adult memory performance as a function of lag and number of retrieval attempts. As a result of the well-established difference in episodic memory between young and older adults, one would expect the difference between the short and long lags in retrieval difficulty during acquisition to be greater for older than for younger adults. However, examining conditional final-recall performance (along with response latencies during encoding) will afford a more direct measure of retrieval difficulty on retention, without the potential confounding influence of differences in retrieval success during encoding.

3 Retrievals

HORSE -- jumped

--

--

--

HORSE -- ?????

APPLE -- evil

--

APPLE -- ?????

HORSE -- ?????

APPLE -- ?????

--

APPLE -- ?????

HORSE -- ?????

Fig. 1 Partial schedule for items receiving three retrieval attempts in the lag 1 (e.g., APPLE–evil) and lag 3 (e.g., HORSE–jumped) conditions

Experiment 1

Method

Participants The young adults were undergraduates at Washington University in St. Louis and received partial course credit or monetary remuneration ($15 or $20 for the short and long retention intervals, respectively) for their participation. The older adults were healthy, community-dwelling individuals who provided their own transportation to the testing site. For their participation, the older adults received monetary remuneration ($20). The participants in each age group were equally divided between the short and long retention interval (RI) conditions (see Table 1 for demographics). Age, years of education, and Shipley vocabulary scores were significantly different between the young and older adults (*p*s

Table 1 Mean (and *SD*) age (in years), education (in years), and Shipley vocabulary score as a function of age, retention interval, and experiment

		Young		Older	
		Short RI	Long RI	Short RI	Long RI
Experiment 1	*N*	49	49	42	42
	Age	20.33 (2.46)	20.76 (2.89)	73.81 (5.20)	75.66 (7.49)
	Education	14.39 (1.92)	14.48 (1.53)	15.79 (2.58)	15.21 (2.93)
	Shipley	33.71 (2.64)	32.96 (3.05)	35.24 (3.55)	35.95 (3.43)
Experiment 2	*N*	24	–	24	–
	Age	19.00 (1.02)	–	70.08 (6.19)	–
	Education	13.75 (1.70)	–	16.54 (2.21)	–
	Shipley	30.25 (3.49)	–	35.92 (2.55)	–

< .005). An additional group of young adults (*n*s = 1 and 3 for short and long RIs, respectively) and older adults (*n*s = 5 and 4 for short and long RIs, respectively) were excluded from the analysis due to low performance on the final test (i.e., unconditional mean accuracy less than 5%), and two additional young adults in the long-RI condition were excluded for not completing the second experimental session.

Design A 2 (Age) × 2 (RI) × 2 (Lag: 1 vs. 3) × 3 (Number of Tests: 1 vs. 3 vs. 5) mixed-factor design was used, with Age and RI as between-participants factors and Lag and Number of Tests as within-participants factors. The short RI was 5 min for both age groups. Because of large age-related differences in long-term retention, and in an attempt to minimize differences due to the scaling of final-test performance (see Salthouse, 2000, for a discussion of Variable × Age interactions), the long RI was 1 h for older adults and one day for younger adults. These RIs were selected on the basis of pilot testing, and to foreshadow our results, the use of different durations successfully equated young and older adult retention following the long RI. The lag between study and test trials was either a single trial (lag 1) or three trials (lag 3), and items were tested one time (one test), three times (three tests), or five times (five tests) without feedback.

Materials A continuous paired-associate task was used for the acquisition phase of the memory task (see Fig. 1 for an example). Fifty-six low-associate word pairs (e.g., APPLE–evil) were selected from the USF Free Association Norms (Nelson, McEvoy, & Schreiber, 1998) that had been used in prior spaced retrieval studies (e.g., Maddox et al., 2011). Word pairs shared some features that made them more easily associable (e.g., WHISKEY–water) or could be used to form a sentence (e.g., HORSE–jumped). These stimuli were divided into seven sets of eight pairs that were counterbalanced across lists, with each pair occurring equally often in each of the within-participants conditions. Stimuli were statistically equated across the sets for word length, frequency, orthographic neighborhood, and phonological neighborhood (Balota et al., 2007), and pairs were equated on backward associative strength across the stimulus sets (*p*s > .10).

The critical conditions were interleaved, and the average serial list position was equated across all conditions (*p*s > .70), as was the average serial list position for the first and the last tests across the six testing conditions (*p*s > .70). Thus, the average RI was constant for all conditions. In total, the acquisition phase included 218 trials, consisting of 192 trials for the critical conditions, 18 filler trials, and eight trials that were equally split between primacy and recency buffer items. Of the 192 critical condition trials, 48 were encoding trials (e.g., HORSE–jumped), and 144 were retrieval practice trials (e.g., HORSE–?????). Filler trials were included to ensure that average serial list position was equated across the critical conditions. The final cued-recall test presented the cue for each critical pair (e.g., HORSE–?????). Encoding trials were presented at a 4.5-s rate, whereas the cues during the acquisition and final-test retrieval trials were presented until participants responded with an answer or by stating that they did not know the answer. Thus, participants were not required to recall an item to complete the cued-recall trial unless they were confident of their response. In both the acquisition and final-test phases, participants were asked to speak their answers aloud. The experimenter indicated when the participant produced the response, via keypress, and then typed the participant's response on a second screen.

Procedure Participants first completed a brief practice phase, which included encoding and retrieval practice trials, before the acquisition phase of the memory task. After the practice phase, participants were instructed to learn the word pairs for a final test and were also aware that they would be tested on the pairs throughout the

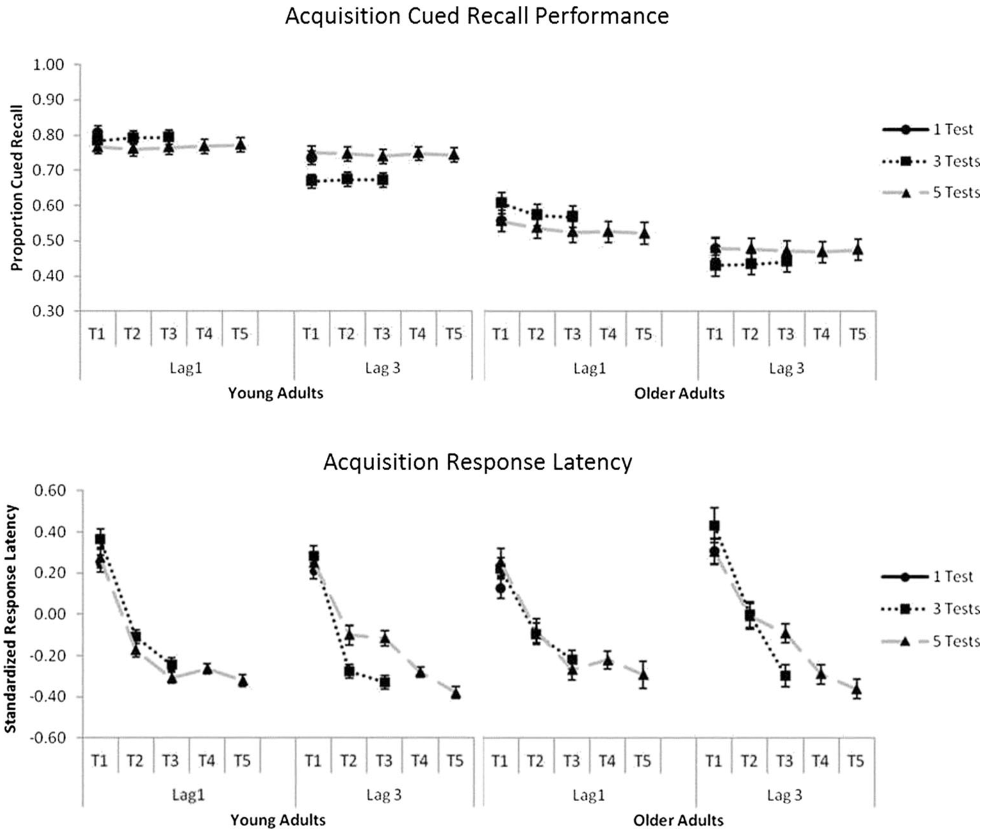

Fig. 2 Mean proportions of cued recall (top panel) and mean standardized response latencies (bottom panel) during the acquisition phase in Experiment 1, as a function of age, lag, number of tests, and test number (e.g., T1 = first retrieval attempt, T2 = second retrieval attempt, etc.). Error bars represent ±1 *SEM*

acquisition phase. On retrieval trials, participants were asked to speak their answers aloud as quickly and accurately as possible, and the experimenter typed the response immediately upon vocalization by the participant.[1] Following the acquisition phase, participants completed 5 min of a distractor trivia task in which trivia questions were presented at a rate of one question every 10 s. The procedure following the distractor task differed as a function of RI group. For participants in the short-RI condition, the final cued-recall test for the memory task occurred immediately following the trivia task. Again, participants were presented with the cue word (e.g., HORSE–????) for each critical pair one at time and made an oral response that the experimenter entered into the computer. Participants then completed a battery of cognitive tasks, the Shipley vocabulary task, and a demographic questionnaire before being dismissed (see Maddox, 2013, for a full discussion and analysis of the cognitive battery). Participants in the long-RI condition proceeded with the cognitive battery, Shipley vocabulary task, and demographic questionnaire following the trivia task. After completing all of the other tasks, older adults completed the final cued-recall test for the memory task before being dismissed. Young adults were dismissed following completion of the other tasks and were asked to return 24 h later to complete the final cued-recall test.

Results

Acquisition data were collapsed across the RI groups, because the acquisition phase was the same for all participants, and indeed we found no differences as a function of RI group (*p*s > .25).

Although there are numerous ways to examine acquisition performance, the present set of analyses focused on the first and last retrieval attempts in each of the multiple-retrieval-

[1] Research assistants were trained by the first author and received extensive practice prior to data collection. To avoid experimenter bias, the research assistants were unaware of the hypotheses related to response latency. Moreover, the procedure utilized in the present study emphasized an immediate buttonpress upon vocalization of participant's response, which triggered a second screen, on which the response was entered.

attempt conditions. This approach allowed for assessments of the stability in accuracy across retrieval attempts in each lag condition and the extent to which response latency was influenced by retrieval attempts and lag.

Accuracy during acquisition Mean proportions of correct recall are shown in Fig. 2 as a function of age, lag, number of tests, and test number. We can note three observations from this figure. First, as expected, young adult performance was higher than older adult performance. Second, performance was higher in the lag 1 condition than in the lag 3 condition, and the difference in performance between lag conditions was greater for older than for young adults. Third, performance remained relatively stable across retrieval attempts in both lag conditions and age groups.

The results from the 2 (Age) × 2 (Lag) × 2 (Number of Tests: 3 vs. 5) × 2 (Test Number: first vs. last) mixed-factor analysis of variance (ANOVA) yielded main effects of age, $F(1, 180) = 89.19$, $p < .001$, $\eta^2_{\mathrm{p}} = .33$, and lag, $F(1, 180) = 50.68$, $p < .001$, $\eta^2_{\mathrm{p}} = .22$, as well as significant Lag × Number of Tests, $F(1, 180) = 21.35$, $p < .001$, $\eta^2_{\mathrm{p}} = .11$, and Lag × Test Number, $F(1, 180) = 4.72$, $p = .031$, $\eta^2_{\mathrm{p}} = .03$, interactions.

Importantly, the Lag × Test Number interaction was qualified by a significant Age × Lag × Test Number interaction, $F(1, 180) = 11.72$, $p = .001$, $\eta^2_{\mathrm{p}} = .06$, which is displayed in Fig. 2. Separate Lag × Test Number ANOVAs were conducted for each age group. An analysis of young adult accuracy only revealed a single main effect of lag, $F(1, 97) = 21.01$, $p < .001$, $\eta^2_{\mathrm{p}} = .18$. An analysis of older adult accuracy revealed main effects of lag, $F(1, 83) = 28.42$, $p < .001$, $\eta^2_{\mathrm{p}} = .26$, and test number, $F(1, 83) = 5.67$, $p = .02$, $\eta^2_{\mathrm{p}} = .06$, as well as a Lag × Test Number interaction, $F(1, 83) = 12.27$, $p = .001$, $\eta^2_{\mathrm{p}} = .13$. Follow-up t tests revealed a single significant difference ($M_{\mathrm{diff}} = .03$) between the first and final tests in the lag 1 condition for older adults, $t(83) = 3.48$, $p = .001$. We observed no difference in performance between the first and final tests in the lag 3 condition for the older adults, or in either lag condition for the young adults (ps > .20). Thus, it appears that older adults produced forgetting across repeated tests for items that were initially retrieved at a short lag. This may have been due to the fact that initial retrieval success is not as strong an indicator of encoding quality following a short lag as following a long lag for older adults. If one can maintain the item across the longer lag for the initial retrieval event, then the item is sufficiently well encoded to be produced across the remaining retrieval events for the older adults. The results from the younger adults suggest that they are not susceptible to this forgetting.

Standardized response latency on successful retrieval during the acquisition phase In the present and all subsequent analyses of response latencies, only latencies from trials on which retrieval was successful were included. All latencies beyond three *SD*s from the mean were excluded from the analysis (<1%). Because older adults were overall slower than young adults, and because this difference in speed can compromise the interpretation of interactions, the response latencies were converted to z scores based on each participant's mean and standard deviation of raw reaction times to correct trials (see Faust, Balota, Spieler, & Ferraro, 1999). For purposes of the ANOVA, missing response latency data were estimated per person per condition by a triangulation procedure in which the relationship in performance between conditions at the group level was used in relation to individual performance at the participant level to provide an estimate for the missing data. Specifically, the relevant conditional means for participants who had at least one observation per cell was taken in proportion to the grand mean for those same participants. In turn, this proportion was used to estimate a given participant's missing cell(s) by multiplying the [conditional mean/grand mean] proportion for all participants and the participant's grand mean. Importantly, the patterns of results were similar when analyses were conducted only on participants who had observations for all cells (see Maddox, 2013, for details)

The standardized mean response latencies on correct trials are displayed in Fig. 2 as a function of age, lag, number of tests, and test number. We can note three observations from the figure. First, response latency decreased across retrieval attempts. Second, the decrease in response latencies between the first and last tests was larger for the lag 3 condition than for the lag 1 condition. Third, the difference between lag conditions in speeding across retrieval attempts was larger for older than for young adults.

The 2 (Age) × 2 (Lag) × 2 (Number of Tests: 3 vs. 5) × 2 (Test Number: first vs. last) mixed-factor ANOVA revealed main effects of number of tests, $F(1, 180) = 4.17$, $p = .043$, $\eta^2_{\mathrm{p}} = .02$, which reflected a small difference in response latencies between the three- and five-test conditions ($M = .03$ vs. $-.03$, respectively), and test number, $F(1, 180) = 473.19$, $p < .001$, $\eta^2_{\mathrm{p}} = .72$, which reflected the speeding of response latencies between the first and last tests ($M = .30$ vs. $-.31$, respectively). Additionally, the Lag × Test Number interaction was significant, $F(1, 180) = 6.94$, $p = .009$, $\eta^2_{\mathrm{p}} = .04$

Importantly, the results from the ANOVA again yielded a significant Age × Lag × Test Number interaction, $F(1, 180) = 4.65$, $p = .032$, $\eta^2_{\mathrm{p}} = .03$, which is displayed in the bottom half of Fig. 2. To examine this interaction, separate 2 (Lag) × 2 (Test Number) ANOVAs were conducted for young and older adults. Analysis of the young adult performance revealed significant effects of lag and test number (ps < .05), but no interaction, whereas the older adult performance revealed a significant effect of test number ($p < .001$) and again a significant Lag × Test Number interaction, $F(1, 83) = 8.32$, $p = .005$, $\eta^2_{\mathrm{p}} = .09$. The significant Lag × Test Number interaction reflected significantly slower response latencies on the first test for the lag 3 than for the lag 1 condition ($M = .37$ vs. .24, respectively, $p = .034$) and a numerical reversal of this pattern on the final test ($M = –.33$ vs. –.25, respectively, $p = .154$). This finding is consistent with R. A. Bjork's (1994) desirable difficulty account, in which items that are retrieved with more difficulty will be strengthened to a greater extent than items retrieved with less difficulty. In the present results, the time that it took to initially retrieve items provides evidence that lag 3 produced a relatively more difficult retrieval event than did lag 1. As a result, the trace may have been strengthened to a greater extent, and consequently could be retrieved faster on the final retrieval attempt in the lag 3 condition than in the lag 1 condition.

Taken together, the three-way interactions among age, lag, and test number observed in both accuracy and response latencies suggest that the underlying memory trace continues to be strengthened by subsequent tests (as indicated by faster response latencies), even when accuracy remains stable. Moreover, the results suggest that the two lag conditions were more distinct for older than for young adults, as indicated by the significant Lag × Test Number interactions for older but not for young adults in the follow-up analyses. Again, this is consistent with different forgetting functions for young and older adults.

Final-test phase performance As we noted earlier, because we were interested in the influences of testing, lag, and age on retention, the present analyses emphasized conditional recall performance. Conditional recall was calculated for items that were correctly retrieved on their final retrieval attempt during the acquisition phase (see Maddox, 2013, for a complete analysis of unconditional performance, which generally accorded with the present analyses).

Conditional final-test accuracy Mean proportions of conditional recall are shown in Fig. 3 as a function of age, RI, lag, and number of tests. We can note three observations in this figure. First, retention was greater for young than for older adults after a short RI, but was comparable across the age groups following the long RI. This confirms that we were successful in matching the young and older adults at the long RI by increasing the RI more for the former than for the latter group. Second, continued retrieval during the acquisition phase led to increased retention for young and older adults when tests were spaced by a single item, regardless of RI. A similar increase in retention was observed across age groups and RIs in the lag 3 condition, when pairs were tested three times versus one time, but no additional benefit was observed in the five-test condition relative to the three-test condition. Third, older adults produced a larger lag effect than did young adults following both RIs.

These observations were supported by the results of a 2 (Age) × 2 (RI) × 2 (Lag) × 3 (Number of Tests) mixed-factor ANOVA. The main effects of RI, $F(1, 178) = 13.86$, $p < .001$, $\eta^2_p = .26$, lag, $F(1, 178) = 65.78$, $p < .001$, $\eta^2_p = .27$, and number of tests, $F(2, 356) = 55.96$, $p < .001$, $\eta^2_p = .24$, were significant, along with a marginal effect of age, $F(1, 178) = 3.53$, $p = .062$, $\eta^2_p = .019$. A reliable interaction between age and RI, $F(1, 178) = 17.69$, $p < .001$, $\eta^2_p = .09$, indicated a significant age difference following the short RI ($p < .001$), but similar performance between groups following the long RI ($p = .134$). The Lag × Number of Tests interaction was also significant, $F(2, 356) = 4.52$, $p = .012$, $\eta^2_p = .03$. This interaction reflected significant increases in performance in the lag 1 condition as the number of tests during acquisition increased, (Ms = .34, .45, and .55 for the one-test, three-test, and five-test conditions, respectively; ps < .001), whereas performance in the lag 3 condition increased from the one-test to the three-test condition ($M = .46$ vs. .61, $p < .001$), but did not increase further with five tests ($M = .61$, $p > .90$). Finally, the Age × Lag interaction was significant, $F(1, 178) = 12.65$, $p <$

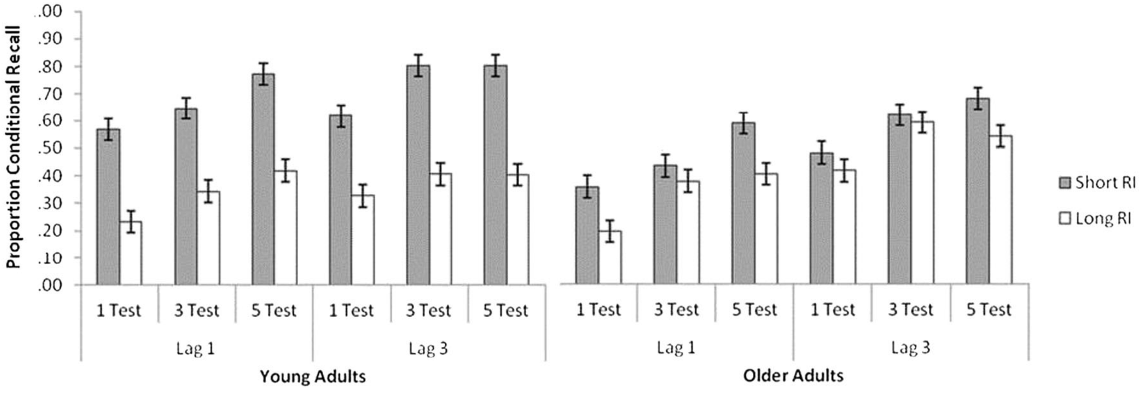

Fig. 3 Mean proportions of conditional cued recall on the final test in Experiment 1, as a function of age, retention interval (RI), lag, and number of tests. Error bars represent ±1 *SEM*

.001, $\eta^2_p = .07$, which reflected a larger lag effect for older adults ($M = .16$) than for young adults ($M = .06$).

Discussion

The final-recall results from Experiment 1 are inconsistent with original predictions from the desirable difficulties perspective that had assumed differences in retrieval difficulty between lag conditions and age groups. Specifically, these assumptions led to the prediction that a benefit of continued retrieval across tests would be observed in the lag 3 condition, but not in the lag 1 condition. However, our results revealed significant increases in retention with each increase in testing in the lag 1 condition (i.e., increased retention when increasing from one to three to five tests). In contrast, for the lag 3 condition, we found an increase in retention between three tests as compared to one test during acquisition, but no comparable increase in retention between three tests and five tests. Importantly, the present experiment afforded a measure of desirable difficulty during encoding—that is, response latency—and hence can provide some direct evaluation of this prediction.

When considering the ways in which response latency may reflect the retrieval difficulty associated with each condition during encoding, one might expect difficulty on the first retrieval attempt to be particularly useful in predicting long-term retention. Specifically, past research has suggested that a long initial lag produces increased long-term memory relative to a short initial lag, regardless of the subsequent form of spacing (i.e., equal spaced vs. expanding retrieval; Karpicke & Roediger, 2007). Thus, one factor likely to influence final-test performance is the response latency on the first retrieval attempt during acquisition (i.e., a slower response latency indicates more difficult retrieval). As is shown at the bottom of Fig. 2, response latencies on the first test were slower in the three-test condition than in the one-test condition ($M_{\text{diff}} = .11$), $t(182) = 2.61$, $p = .010$, and there was no difference in response latencies between the three-test and five-test conditions ($M_{\text{diff}} = .06$), $t(181) = 1.31$, $p = .190$. Hence, one would predict that conditional accuracy on the final test should be greater in the three- and five-test conditions than in the one-test condition, and indeed, the final-recall results indicated that taking three tests produced a benefit over taking one test in both lag conditions.[2] Moreover, response latencies were similar across lag conditions for young adults ($M_{\text{diff}} = .05$) [Lag: $F(1, 97) = 1.41$, $p = .237$, $\eta^2_p = .014$], but were faster in the lag 1 condition than in the lag 3 condition for older adults ($M_{\text{diff}} = .13$) [Lag: $F(1, 83) = 4.66$, $p = .034$, $\eta^2_p = .05$]. Thus, the lag effect should be larger for older than for younger adults in conditional final-test performance, which was observed.

In sum, measuring acquisition response latency and conditional final-test recall provided better leverage in examining the influence of desirable difficulty on long-term memory performance than did relying solely on assumptions about the retrieval difficulty of our various manipulations. Indeed, our results were largely consistent with R. A. Bjork's (1994) desirable difficulty framework.

Experiment 2

Experiment 2 was motivated to extend the results from Experiment 1 to an examination of massed versus spaced retrieval practice, which should produce a more extreme manipulation of spacing than the lag manipulation used in Experiment 1. On the basis of the extant literature, one would predict that continued testing with an ineffective lag—namely massed retrieval—should produce little benefit in final-test performance relative to continued testing with a more effective lag.

It is possible, however, that continued massed retrieval may enhance long-term retention as compared to a single massed retrieval attempt, as a result of mechanisms other than desirable difficulty. Specifically, Experiment 2 was also motivated by an intriguing age-related difference in the benefits of refreshing. Johnson, Reeder, Raye, and Mitchell (2002) reported that young adults may benefit more from an immediate retrieval attempt than do older adults, a process referred to as *refreshing*, despite older adults being slower to retrieve items on adjacent trials than are young adults. Interestingly, Maddox et al. (2011) also found that younger adults appeared to benefit more from an immediate test than did older adults. Provided that older adults are slower than younger adults to retrieve items on massed testing trials, it does not appear that the benefit of refreshing reflects desirable difficulty. Indeed, Johnson et al. suggested that refreshing yields prolonged activation of the item, and in turn, that this prolonged activation benefits young adult memory performance to a greater extent than older adult memory performance. Thus, one might expect young adults'

[2] Both three and five tests led to similar levels of retention in the lag 3 condition, but this was not true in the lag 1 condition. Specifically, taking five tests in the lag 1 condition produced significantly better performance on the final test than did taking three tests, which might reflect an additional benefit obtained from additional retrieval practice. Thus, the influence of repeated exposure to material via testing may compensate for less-effective spacing intervals.

long-term memory performance to benefit more from continued massed retrieval than would older adults' performance.

Given our specific interests in addressing how desirable difficulty is operationally defined and in accounting for differences between conditions in acquisition phase accuracy, Experiment 2 included only two levels of testing (one vs. three tests) and one RI (5 min). Because RI did not interact with spacing and testing in Experiment 1, we only tested a short RI in Experiment 2, which still allowed us to examine conditional final-test performance and acquisition response latency as a proxy for retrieval difficulty. These changes in methodology had the additional benefit of reducing the overall list length and increasing older adult performance to be closer to younger adult performance than we had observed in Experiment 1.

Method

Participants The young adults were undergraduates at Washington University in St. Louis and received partial course credit or monetary remuneration ($10) for their participation. The older adults were healthy, community-dwelling adults and received monetary compensation ($15) for their participation. Their demographics are displayed in Table 1.

Design A 2 (Age) × 2 (Lag: 0 vs. 4) × 2 (Number of Tests: 1 vs. 3) mixed-factor design was used in Experiment 2. Age was a between-participants factor, and Lag and Number of Tests were within-participants factors. The RI was 5 min for both age groups.

Materials A subset of 32 low-associate word pairs was selected from Experiment 1. These stimuli were divided into four sets of eight pairs, and the sets were counterbalanced across lists such that each pair occurred once in each of the within-participants conditions. A continuous paired-associate task was again used for the acquisition phase of the memory task. The average serial list position was equated across conditions (*p*s > .75). In total, the acquisition phase consisted of 139 trials, of which 96 trials were critical condition trials, 35 were filler trials, and eight trials were equally split between primacy and recency buffer items. Of the 96 critical condition trials, 32 were encoding trials and 64 were retrieval practice trials. Filler trials were included to

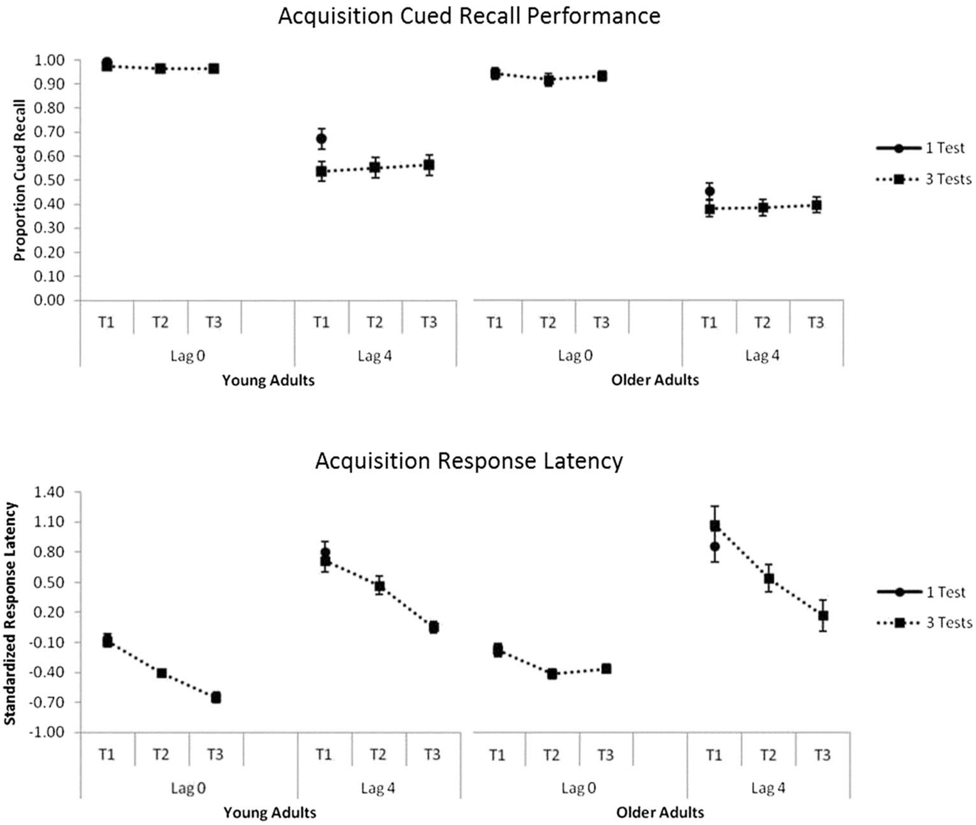

Fig. 4 Mean proportions of cued recall (top panel) and mean standardized response latencies (bottom panel) during the acquisition phase in Experiment 2, as a function of age, lag, number of tests, and test number (i.e., T1 = first retrieval attempt, T2 = second retrieval attempt, and T3 = third retrieval attempt). Error bars represent ±1 *SEM*

ensure that the average serial list position was equated across the critical conditions. Thus, the average RI was constant for all conditions.

Procedure The procedure was the same as in Experiment 1, with three exceptions: (a) only single-test and three-test conditions were included; (b) a single, 5-min RI was used; and (c) the lags were changed to 0 and 4, to investigate the effects of massed versus spaced testing.

Results

The present set of analyses again emphasized performance on the first and last retrieval attempts in each of the multiple-retrieval-attempt conditions as a way of assessing the stability of retrieval accuracy across testing events and the degree to which response latencies decreased across test events as a function of lag.

Acquisition memory accuracy Mean proportions of correct recall for young and older adults are shown in Fig. 4 as a function of lag, number of tests, and test number. Accuracy from the first and last retrieval attempt in the three test conditions were submitted to a 2 (Age) × 2 (Lag) × 2 (Test Number) mixed-factor ANOVA. The results revealed main effects of age and lag, *p*s < .005, that were further qualified by a significant Age × Lag interaction, $F(1, 46) = 5.86$, $p = .020$, $\eta^2_p = .11$. The interaction revealed statistically equivalent performance across age groups in the lag 0 condition, $p = .173$, but a significant difference in lag 4 performance between young (.55) and older (.39) adults, $p = .003$, reflecting the greater forgetting rate in older adults across the longer lags, consistent with the Experiment 1 results.

Standardized response latencies on successful retrieval attempts during the acquisition phase Mean standardized response latencies for young and older adults are shown in the bottom panel of Fig. 4 as a function of lag, number of tests, and test number. Again, response latency data from the first and last retrieval attempts in the three-test condition were submitted to a 2 (Age) × 2 (Lag) × 2 (Test Number) mixed-factor ANOVA. All main effects were significant, *p*s < .05, in addition to a significant Lag × Test Number interaction, $F(1, 46) = 10.87$, $p = .002$, $\eta^2_p = .19$. Moreover, the three-way interaction was significant, $F(1, 46) = 6.17$, $p = .017$, $\eta^2_p = .12$. Separate analysis of the young adult response latencies revealed main effects of lag and test number, *p*s < .001, with no interaction, $p > .50$. Analysis of the older adult response latencies also revealed significant main effects of lag and test number, *p*s < .001, and a significant Lag × Test Number interaction, $F(1, 23) = 14.12$, $p = .001$, $\eta^2_p = .38$. As is shown in Fig. 4, response latencies for older adults decreased from the first to the second retrieval attempt in the lag 0 condition ($p = .002$), but remained stable from the second to the third retrieval attempt ($p > .40$). In contrast, response latencies significantly decreased across all retrieval attempts in the lag 4 condition, *p*s < .05

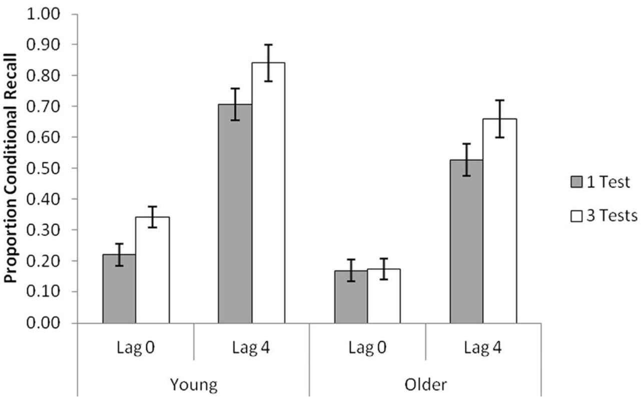

Fig. 5 Mean proportions of conditional cued recall on the final test in Experiment 2, as a function of age, lag, and number of tests. Error bars represent ±1 *SEM*

Conditional final-test memory accuracy Figure 5 displays the mean proportions of conditional recall as a function of age, lag, and number of tests. Performance was higher for young than for older adults (*M* = .53 vs. .38, respectively); lag 4 items were remembered better than lag 0 items (*M* = .68 vs. .23, respectively); and taking three tests led to better retention than did taking a single test (*M* = .50 vs. .41, respectively).

The 2 (Age) × 2 (Lag) × 2 (Number of Tests) mixed-factor ANOVA on conditional final recall yielded main effects of age, $F(1, 46) = 15.44$, $p < .001$, $\eta^2_p = .25$, lag, $F(1, 46) = 272.33$, $p < .001$, $\eta^2_p = .86$, and number of tests, $F(1, 46) = 8.03$, $p = .007$, $\eta^2_p = .15$. Although the three-way interaction was not significant, separate analyses were conducted to examine the benefit of additional testing for each lag condition, given the a priori predictions based on age differences in refreshing discussed above. Our analysis of lag 0 performance revealed main effects of age and number of tests (*p*s < .05), which were further qualified by a significant Age × Number of Tests interaction, $F(1, 46) = 4.20$, $p = .046$, $\eta^2_p = .08$. This interaction reflected a significant increase in performance when testing was increased from one to three tests in the massed condition for young adults ($p = .005$), but not for older adults ($p > .90$). Regarding lag 4 performance, the ANOVA revealed main effects of age and number of tests (*p*s < .05), but no interaction, $p > .95$. As predicted, these results indicated that young adults benefited from repeated testing when refreshing was engaged in the lag 0 condition. Older adults did not produce this benefit.

Discussion

The results from Experiment 2 are clear. First, as predicted, both age groups benefited from spaced retrieval and continued testing in the lag 4 condition. However, only young adults benefited in terms of conditional accuracy from continued testing in the lag 0 condition, which is consistent with the previously reported age differences in refreshing described above.

With respect to the influence of retrieval difficulty on long-term memory performance, there was a clear relationship between acquisition phase response latency and retention. A significant three-way interaction was observed in acquisition phase response latencies among age, lag, and test number. A follow-up analysis of this interaction revealed main effects of lag and test number, but no interaction in young adult performance, whereas an analysis of older adult performance revealed effects of lag and test number, as well as a significant Lag × Test Number interaction. First, consider how the young adult retrieval latencies are related to final-recall performance. The lack of an interaction between lag and test number in acquisition response latency suggests that the benefits of continued retrieval practice over taking a single test should be equivalent across lag conditions, and indeed, additive effects of lag and number of tests were observed in retention for the younger adults. Turning to the older adult data, the significant interaction between lag and test number in older adult response latencies leads one to expect a larger benefit of repeated testing in the lag 4 condition than in the lag 0 condition, which was also observed. In sum, the present emphasis on acquisition phase response latency as a proxy for retrieval difficulty provided a more precise way of examining the desirable difficulty account. Indeed, it appears that overall, the results are consistent with the benefits of desirable difficulty.

General discussion

The present experiments have extended past studies investigating spacing and retrieval practice in two ways. First, the response latency for correctly retrieved items during acquisition was used as a metric of retrieval difficulty, to allow for a more precise assessment of R. A. Bjork's (1994) desirable difficulty account. Second, final-test performance was examined only for those items that were correctly retrieved (i.e., that received the benefit of retrieval practice) during acquisition. In this way, the influence of spaced retrieval on retention was isolated better from the effects of spaced retrieval on encoding than in previous studies.

Importantly, with these methodological extensions, in the present experiments we were able to more carefully examine two questions regarding the benefits of spaced retrieval practice across age groups and RIs. First, how does lag modulate the extent to which continued testing improves long-term memory? Second, how does the function relating lag and continued testing to final-test performance differ across young and older adults? We will first address these two questions before considering the extent to which our results are consistent with Bjork's desirable difficulty account.

Retrieval practice as a function of lag

The results from Experiment 1 provided information regarding the function relating continued testing and lag to final-test performance. Similar to previous studies (e.g., Karpicke & Roediger, 2007; Wheeler & Roediger, 1992), the results revealed a long-term retention benefit with increased testing when comparing a single test condition to a three test condition in both the lag 1 condition (11% benefit) and lag 3 condition (15% benefit). More importantly, the inclusion of a third level of testing (i.e., five tests) extended previous studies and revealed a difference in the function relating continued testing and lag to final-test performance. Specifically, retention continued to increase with additional retrieval practice in the short-lag condition (10% from three to five tests) but did not increase in the long-lag condition (0% from three to five tests). Thus, the benefits of additional retrieval practice appear to asymptote after three successful retrieval events in the long-lag condition but not in the short-lag condition.

It is important to note that a similar pattern of data has recently been observed in paradigm utilizing feedback and learning to criterion with young adults (Rawson & Dunlosky, 2011). Specifically, Rawson and Dunlosky reported results from a cued-recall paradigm that suggested the most efficient way of scheduling study events is to learn material to an initial criterion of three correct retrieval events and then schedule three relearning sessions in which material is retrieved to a criterion of one correct retrieval. This is a critical observation, because the present study emphasized the influence of spaced retrieval on retention (i.e., conditional accuracy) rather than unconditional final-test performance. Thus, there is converging evidence across different methodologies that the benefits of spaced retrieval practice asymptote after an item has been tested a specific number of times (i.e., three times) following a relatively long lag. Continuing to test material beyond this optimal number provides relatively little additional benefit in terms of retention. It is also important to note that both methodologies assessed final-test performance when differences in acquisition performance were minimized between conditions. Of course, ultimately, the precise number of tests needed to maximize the efficiency of retrieval practice is likely to be influenced by the precise lag used, the retention interval, the difficulty of the materials to be learned and the ability of the learner.

The present results also have implications for one of the leading accounts of the spacing effect, namely the combined study-phase retrieval and encoding variability account (e.g., Greene, 1989; Raaijmakers, 2003). This account proposes that the benefit of spaced study results from increased encoding variability for repetitions separated by time or intervening items relative to repetitions that are studied consecutively. Often, the account suggests that participants must retrieve the first presentation of an item when it is later re-studied to obtain the full mnemonic benefit of spaced study (e.g., Madigan, 1969; Melton, 1967; Thios & D'Agostino, 1976). Our results are generally consistent with this account when examining the single-test and three-test conditions (see Fig. 3; Exp. 1). However, the five-test condition included in Experiment 1 suggests that the benefits of encoding variability may asymptote (i.e., continued retrieval after initial encoding variability may yield minimal increases in long-term performance), and that continued retrieval practice may help compensate for the use of a less variable encoding condition (i.e., lag 1). Of course, some caution must be exerted when attributing the benefits of repeated retrieval to encoding variability versus retrieval difficulty, because these conditions are naturally confounded (i.e., more variable encoding is predicted to lead to increased retrieval difficulty). Thus, future work should attempt to isolate the contributions of these two mechanisms to long-term memory performance.

Age-related differences and the benefits of spaced retrieval practice

One particularly interesting aspect of Experiment 1 is the reliable Age × Lag interaction in accuracy, which reflected a larger lag effect in conditional accuracy for older than for young adults. As can be seen in Fig. 6, a much larger lag effect was observed for older than for young adults following a long RI, as compared to the short RI. At this level, it appears that older adults actually benefited more from the lag 3 condition in terms of conditional accuracy than did young adults. This pattern is particularly important, because young and older adult performance was equated in the lag 1 condition (as is shown in Fig. 6). Hence, it appears that the increased lag effect in older adults may be due to age-related differences in retrieval effort and desirable difficulty (R. A. Bjork, 1994). Specifically, because older adults have lower performance during the long-lag condition during encoding, those items that did survive may have benefited more from desirable difficulty in the older adult group than in the young adult group and, hence, may have produced a stronger long-term trace. Indeed, this interpretation is consistent with the age differences observed across lag conditions in acquisition response latencies, discussed above.

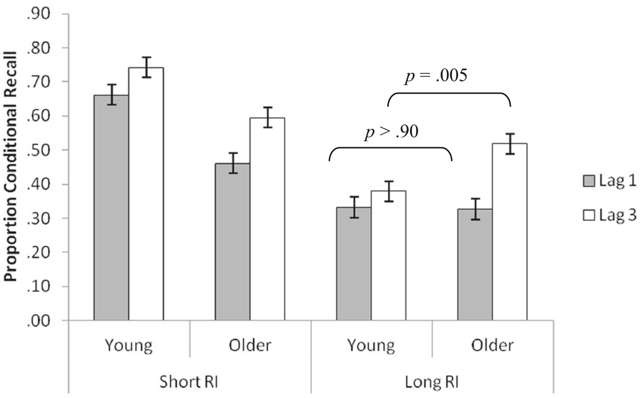

Fig. 6 Proportions of conditional cued recall on the final test in Experiment 1, as a function of age, retention interval (RI), and lag. Error bars represent ±1 *SEM*

With respect to Experiment 2, the results indicated that both age groups benefited from continued testing in the long-lag condition. However, young and older adults differed in the benefit of continued massed retrieval practice. Namely, young adults benefited from continued testing, but older adults failed to show a similar benefit. This benefit of repeated testing in young adults in the massed condition led to a reduction in the overall spacing effect, and hence contributed to the observation that older adults produced a larger lag effect than did young adults, which is consistent with the results observed at the long RI in Experiment 1. More importantly, the benefits of the immediate retrieval event in young but not older adults provide further support for age differences in a refreshing mechanism (Johnson et al., 2002; Maddox et al., 2011).

Although the present results are useful for examining the benefits of spaced retrieval for items that successfully incurred retrieval practice during encoding, an emphasis on conditional analyses would overlook overall differences between the age groups and lag conditions in performance during acquisition. Indeed, the benefit of a long lag during acquisition is offset by reduced acquisition performance. Thus, in future studies researchers may wish to extend recent work reported by Rawson and Dunlosky (2011) to an older adult population, as a means of examining the benefits of criterion level learning and the benefits of continued testing with feedback in this group (see Pyc & Balota, 2013, for such a study).

Continued retrieval practice and desirable difficulty

As we discussed in the introduction, the benefits of various spaced retrieval schedules have been tied to the degree to which a given schedule produces desirably difficult retrieval (e.g., R. A. Bjork, 1994) during the learning phase. On the basis of Bjork's concept of desirable difficulty, we originally predicted that continued testing would improve final-recall performance in the long-lag condition, but would produce relatively little improvement in the short-lag condition, given that longer lags should lead to more difficult retrieval attempts than short lags. This pattern of data was not observed in the present experiments.

In our discussion of each experiment, we more closely examined acquisition phase response latencies as a proxy for retrieval difficulty, rather than relying on a priori assumptions about differences in difficulty across conditions and participant groups. This approach allowed for a more precise assessment of retrieval difficulty, and ultimately revealed general support for R. A. Bjork's (1994) account. Indeed, this approach led to a different conclusion than would have been reached had we simply assumed retrieval difficulty differences as a function of lag condition and age group. Hence, these results emphasize the importance of measuring difficulty in order to examine the influences of desirable difficulty.

Jointly considering acquisition phase response latencies and conditional final recall accuracy in the present study also provided evidence that an item continues to strengthen and increase in accessibility after it has been successfully retrieved on an initial test event, despite little change in accuracy across subsequent test events during the acquisition phase. Specifically, the speeding of response latencies across later retrieval attempts suggests that items continue to strengthen with each additional test, which may be viewed as consistent with the bifurcation model of the testing effect (Halamish & Bjork, 2011; Kornell, Bjork, & Garcia, 2011; Storm, Friedman, Murayama, & Bjork, 2014). This model rests on two core assumptions. First, to-be-remembered items are distributed across a continuous "memory strength" dimension that influences the probability of successful encoding. Second, restudying a list of items after initial encoding will shift the entire distribution along the memory strength dimension, whereas testing without feedback will bifurcate the distribution, such that successfully retrieved items are substantially increased in memory strength, and unsuccessfully retrieved items retain their original memory strength. Critically, the act of testing will increase the memory strength for successfully retrieved items to a greater extent than the increase in memory strength obtained by restudying all items. Once the distribution of items is bifurcated, it is not always clear whether items continue to be strengthened with additional retrieval practice, because all items have memory strength that surpasses the threshold for successful retrieval. However, the acquisition phase response latencies observed in the present study suggest that strengthening continues to occur even when it is otherwise undetectable in the accuracy measures.

Limitations of the present study

Two noteworthy limitations of the present study should be considered in future work. First, in the present study we used different retention intervals across young and older adult groups in Experiment 1, to equate overall performance. Although this was successful (see Salthouse, 2000, for a similar procedure), this manipulation may have also allowed for other mechanisms to contribute to the observed results (e.g., the one-day RI for young adults may have allowed for greater consolidation of material than did the 1-h RI for older adults). Second, in order to minimize the noise variance associated with age-related differences in motor control and computer use (e.g., Hickman, Rogers, & Fisk, 2007; Nair et al., 2007), which could influence the response latency measure, the experimenter immediately typed in the response, and this was used as this was used as the proxy for retrieval difficulty. Although the results were quite systematic in the present study, it is possible that the experimenters slowed their responses for older adults relative to those for younger adults, which has been observed when caregivers speak to older adults (e.g., Kemper, 1994). Of course, such an effect would have to be more subtle as a function of condition, since in the present study we used standardized response latencies, which controlled for overall age-related slowing.

Conclusions

The results from the present study underscore the importance of conditional analyses in understanding the differing effects of spaced retrieval practice on the learning and retention of material. Moreover, the use of acquisition response latencies as a proxy for retrieval difficulty provided a more accurate assessment of the desirable difficulty account (R. A. Bjork, 1994) than would by simply assuming differences in retrieval difficulty across conditions, and also provided evidence that items continue to strengthen with additional retrieval practice, even when those changes are not evident in cued-recall performance. Overall, the present results provide support for an important role of desirable difficulty in accounting for the benefits of spacing and repeated testing in younger and older adults.

Author note Portions of this article are based on the dissertation by G.B.M. We thank Jan Duchek, Mark McDaniel, and Roddy Roediger for helpful comments at various stages of this work. Additionally, we thank Clare Misko and Zack Pinsky for their help with data collection. This research was supported by NIA Training Grant No. AG00030.

References

Arking, R. A. (1998). Perspectives on aging. In R. A. Arking (Ed.), *Biology of aging* (2nd ed., pp. 2–36). Sunderland, MA: Sinauer.

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

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., Duchek, J. M., & Logan, J. M. (2007a). Is expanded retrieval practice a superior form of spaced retrieval? A critical review of the extent literature. In J. S. Nairne (Ed.), *The foundations of remembering: Essays in honor of Henry L. Roediger III* (pp. 83–105). New York, NY: Psychology Press. doi:10.3758/MC.38.1.116

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

Benjamin, A. S., & Tullis, J. (2010). What makes distributed practice effective? *Cognitive Psychology, 61,* 228–247. doi:10.1016/j.cogpsych.2010.05.004

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). Cambridge, MA: MIT Press.

Bjork, R. A. (2013). Desirable difficulties perspective on learning. In H. Pashler (Ed.), *Encyclopedia of the mind* (pp. 242–244). Thousand Oaks, CA: Sage.

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). New York, NY: Worth.

Bui, D. C., Maddox, G. B., & Balota, D. A. (2013). The roles of working memory and intervening task difficulty in determining the benefits of repetition. *Psychonomic Bulletin & Review, 20,* 341–347.

Camp, C. J., Foss, J. W., Stevens, A. B., & O'Hanlon, A. M. (1996). Improving prospective memory task performance in person with Alzheimer's disease. In M. Bandimonte, G. O. Einstein, & M. A. McDaniel (Eds.), *Prospective memory: Theory and applications* (pp. 351–367). Mahwah, NJ: USum Associates.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132,* 354–380. doi:10.1037/0033-2909.132.3.354

Clark, C. M., & Bjork, R. A. (2014). When and why introducing difficulties and errors can enhance instruction. In V. A. Benassi, C. E. Overson, & C. M. Hakala (Eds.), *Applying the science of learning in education: Infusing psychological science into the curriculum* (pp. 20–30). Washington, DC: Society for the Teaching of Psychology.

Erdelyi, M. H., & Becker, J. (1974). Hypermnesia for pictures: Incremental memory for pictures but not for words in multiple recall trials. *Cognitive Psychology, 6,* 159–171.

Faust, M. E., Balota, D. A., Spieler, D. H., & Ferraro, F. R. (1999). Individual differences in information-processing rate and amount: Implications for group differences in response latency. *Psychological Bulletin, 125,* 777–799. doi:10.1037/0033-2909.125.6.777

Greene, R. L. (1989). Spacing effects in memory: Evidence for a two process account. *Journal of Experimental Psychology Learning Memory and Cognition, 15,* 371–377. doi:10.1037/0278-7393.15.3.371

Halamish, V., & Bjork, R. A. (2011). When does testing enhance retention? A distribution-based interpretation of retrieval as a memory modifier. *Journal of Experimental Psychology Learning Memory and Cognition, 37,* 801–812.

Hertzog, C., Kramer, A. F., Wilson, R. S., & Lindenberger, U. (2009). Enrichment effects on adult cognitive development. Can the functional capacity of older adults be preserved and enhanced? *Psychological Science in the Public Interest, 9,* 1–65.

Hickman, J. M., Rogers, W. A., & Fisk, A. D. (2007). Training older adults to use new technology. *Journal of Gerontology, 62B,* 77–P84.

Johnson, M. K., Reeder, J. A., Raye, C. L., & Mitchell, K. J. (2002). Second thoughts versus second looks: An age-related deficit in reflectively refreshing just-active information. *Psychological Science, 13,* 64–67. doi:10.1111/1467-9280.00411

Karpicke, J. D., & Roediger, H. L., III. (2007). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval promotes long-term retention. *Journal of Experimental Psychology Learning Memory and Cognition, 33,* 704–719. doi:10.1037/0278-7393.33.4.704

Kemper, S. (1994). Elderspeak: Speech accommodations to older adults. *Aging, Neuropsychology, and Cognition, 1,* 17–28.

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

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

Maddox, G. B. (2013). *The efficiency of retrieval practice as a function of spacing and intrinsic value in young and older adults (Unpublished doctoral dissertation)*. Louis, St. Louis, MO: Washington University in St.

Maddox, G. B., Balota, D. A., Coane, J. H., & Duchek, J. M. (2011). The role of forgetting rate in producing a benefit of expanded over equal spaced retrieval in young and older adults. *Psychology and Aging, 26,* 661–670.

Madigan, S. A. (1969). Intraserial repetition and coding processes in free recall. *Journal of Verbal Learning and Verbal Behavior, 8,* 828–835.

Melton, A. W. (1967). Repetition and retrieval from memory. *Science, 158,* 532.

Nair, S. N., Czaja, S. J., Sharit, J. (2007). A multilevel modeling approach to examining individual differences in skill acquisition for a computer-based task. *Journals of Gerontology, 62B,* 85–96.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms [Database]. Retrieved from w3.usf.edu/FreeAssociation/

Payne, D. G., & Roediger, H. L., III. (1987). Hypermnesia occurs in recall but not recognition. *American Journal of Psychology, 100,* 145–166.

Pyc, M. A., & Balota, D. A. (2013). *Catastrophic interference? The influence of lag and testing on retention in young and older adults*. Toronto, Ontario, Canada: Paper presented at the Annual Meeting of the Psychonomic Society.

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

Raaijmakers, J. G. W. (2003). Spacing and repetition effects in human memory: Application of the SAM model. *Cognitive Science, 27,* 431–452. doi:10.1207/s15516709cog2703_5

Rawson, K. A., & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? *Journal of Experimental Psychology: General, 140,* 283–302. doi:10.1037/a0023956

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140,* 1432–1463. doi:10.1037/a0037559

Salthouse, T. A. (1996). The processing speed theory of adult age differences in cognition. *Psychological Review, 103,* 403–428. doi:10.1037/0033-295X.103.3.403

Salthouse, T. A. (2000). Methodological assumptions in cognitive aging research. In F. I. M. Craik & T. A. Salthouse (Eds.), *Handbook of cognitive aging* (2nd ed., pp. 467–498). Mahwah, NJ: Erlbaum.

Schacter, D. L., Rich, S. A., & Stamp, M. S. (1985). Remediation of memory disorders: Experimental evaluation of the spaced-retrieval technique. *Journal of Clinical and Experimental Neuropsychology, 7,* 70–96.

Storm, B. C., Friedman, M. C., Murayama, K., & Bjork, R. A. (2014). On the transfer of prior tests or study events to subsequent study. *Journal of Experimental Psychology Learning Memory and Cognition, 40,* 115–124.

Sungkhasettee, V. W., Friedman, M. C., & Castel, A. D. (2011). Memory and metamemory for inverted words: Illusions of competency and desirable difficulties. *Psychonomic Bulletin & Review, 18,* 973–978. doi:10.3758/s13423-011-0114-9

Thios, S. J., & D'Agostino, P. R. (1976). Effects of repetition as a function of study-phase retrieval. *Journal of Verbal Learning and Verbal Behavior, 15,* 529–536.

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.

Yue, C. L., Castel, A. D., & Bjork, R. A. (2013). When disfluency is—and is not—a desirable difficulty: The influence of typeface clarity on metacognitive judgments and memory. *Memory & Cognition, 41,* 229–241. doi:10.3758/s13421-012-0255-8