Examining the contributions of desirable difficulty and reminding to the spacing effect

Geoffrey B. Maddox¹ · Mary A. Pyc² · Zachary S. Kauffman¹ · Jessica D. Gatewood¹ · Aubrey M. Schonhoff¹

Published online: 25 July 2018
© Psychonomic Society, Inc. 2018

Abstract

Although substantial evidence indicates that spacing repeated study events with intervening material generally enhances memory performance relative to massing study events, the mechanism underlying this benefit is less clear. Two experiments examined the role of reminding difficulty during the acquisition of material in modulating final memory performance for spaced repetitions utilizing recognition (Experiment 1) and recall tests (Experiment 2). Specifically, participants studied a list of words presented one or two times separated by one or five items. On each trial participants reported whether the item had been previously presented (i.e., repetition detection judgment), and the response latency served as a proxy for reminding difficulty such that longer response latencies reflected more difficult reminding. A third experiment extended this paradigm with the inclusion of a massed condition and novel lag conditions (three and ten items). Results revealed significant lag effects in final test performance across experiments despite comparable repetition detection difficulty between lag conditions during acquisition. Moreover, results from within-participant point-biserial analyses and mediation analyses converged on overall performance measures in suggesting that repetition detection difficulty failed to modulate final test performance in the current paradigm. Discussion considers the implications of the current results for mechanisms proposed to underlie the benefits of spaced study and spaced retrieval practice.

Keywords Spacing effect · Desirable difficulty · Reminding · Encoding variability

Introduction

Substantial evidence indicates that material repeatedly studied with intervening time or other study material is typically remembered better on a later memory test than material repeatedly studied in massed fashion (i.e., the *spacing effect*; Ebbinghaus, 1885). Moreover, research suggests that the function relating the precise spacing interval (i.e., lag) between repetitions of an item and memory performance is an inverted U-shape. Specifically, final test performance initially increases as the lag between repetitions of an item increases, but at a certain point, continued increases in lag result in reduced performance (Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006). Lag and spacing effects are ubiquitous and have consequently been the

subject of numerous reviews (e.g., Crowder, 1976; Delaney, Verkoeijen, & Spirgel, 2010; Dempster, 1996; Maddox, 2016). However, the underlying mechanism for the benefit of spaced study is still somewhat unclear.

The current study examined one leading explanation for spacing and lag effects that combines a reminding process (e.g., Hintzman, 2004, 2010) with the concept of desirable difficulty (Bjork, 1994) to explain the inverted U-shape function of lag effects (e.g., the reminding account from Benjamin & Tullis, 2010). Critically, each of the two constituent mechanisms included in Benjamin and Tullis' reminding account has been useful in explaining cognitive phenomena beyond the lag effect. For example, the desirable difficulty account (Bjork, 1994) suggests that more effortful processing during the acquisition of information yields enhanced memory relative to items that require less effortful processing and has been useful in considering the benefits of the testing effect (e.g., Roediger & Karpicke, 2006; Soderstrom, Kerr, & Bjork, 2016) and the generation effect (Slamecka & Graf, 1978). The reminding account (Hintzman, 2004, 2010) suggests that studied items have the propensity to trigger retrieval of earlier instances of the same or related items. Past research indicates

✉ Geoffrey B. Maddox
maddoxg@rhodes.edu

¹ Department of Psychology, Rhodes College, 2000 N. Parkway, Memphis, TN 38112, USA

² Dart NeuroScience, LLC, San Diego, CA, USA

that items with successful reminding events are better remembered on final memory tests than repeated items for which reminding was not successful (e.g., Jacoby, 1974; Wahlheim, Maddox & Jacoby, 2014). Moreover, evidence suggests that reminding may help explain how individuals judge the number of times an item was encountered during the learning phase of an experiment (e.g., Hintzman, 2004), report the order in which two different items appeared (Hintzman, 2010; Jacoby, Wahlheim, & Yonelinas, 2013), and discriminate the study list in which an item was presented (e.g., Jacoby & Wahlheim, 2013). Thus, the current study aimed to examine the utility of the reminding mechanism incorporating desirable difficulty in accounting for the lag effect while potentially providing greater insight into how these separate mechanisms may operate beyond the current phenomenon of interest. To address this aim, we first review the reminding mechanism broadly before considering the additional contribution of desirable difficulties to explaining the lag effect.

Repetition detection and reminding

One precursor to the reminding account originally proposed to explain the benefit of spaced versus massed study was formalized by Thios and D'Agostino (1976) in the form of the study-phase retrieval account. This account suggested that the benefit of spaced versus massed study would only be observed if the item was detected as having previously been studied when presented a second time during acquisition. In support of this account, Madigan (1969; see also Hintzman, Summers, & Block, 1975) asked participants to learn a series of words presented either one or two times, complete a recall test, and then provide frequency judgments for each recalled word (i.e., report the number of times each recalled item was presented during the acquisition). Notably, Madigan only observed a spacing effect for items that were correctly reported as having been repeatedly studied during acquisition. No spacing effect was observed for repeated items reported as having been studied only one time. Similar results have been observed when unique items belonging to the same semantic category are presented during acquisition (e.g., orange and apple; Jacoby, 1974) such that the benefit of spaced study is obtained when participants detect the repetition of the category across items. Thus, evidence suggests that detecting repetitions and relationships between similar items might be critical for obtaining the benefit of spaced versus massed study. Nonetheless, it is important to note that simply detecting the repetition of an item cannot account for spacing and lag effects. Specifically, study-phase retrieval should be most successful for massed items, and the probability of successful study-phase retrieval should decrease as the lag between repetitions increases. Thus, the function relating lag to memory performance is predicted to decrease monotonically, which should in turn yield better memory for massed over spaced items. To address this shortcoming, recent theories have

posited a role for additional mechanisms to be combined with a reminding mechanism so that the joint influence of reminding and other mechanisms (e.g., desirable difficulties; Bjork, 1994) yields the nonmonotonic function of the lag effect.

Reminding and the role of desirable difficulty

To address the otherwise monotonically decreasing function relating lag to final test performance predicted by the study-phase retrieval mechanism, Benjamin and Tullis (2010; see also Tullis, Benjamin, & Ross, 2014) offered an alternative account that combined Hintzman's (2004) reminding mechanism with Bjork's (1994) desirable difficulty account. In his account, Bjork proposed that the effort involved in successfully retrieving information from memory during acquisition serves to modulate subsequent storage and retrieval of that information. For example, assume that one is trying to remember and then successfully retrieve two different past events. The event that requires more effort for successful retrieval will be strengthened in memory to a greater extent than the event that requires relatively less effort. In the context of spacing and lag effects, reminding or retrieval of an item's first presentation should become increasingly difficult as the lag between repetitions increases. When considered separately, the desirable difficulty account predicts that the function relating lag and memory will increase monotonically. To yield the inverted U-shape function relating lag to memory typically observed (e.g., Cepeda et al., 2006), consider the way in which these mechanisms may operate concurrently. As the lag between repetitions increases, the effort involved in the reminding or retrieval of an item's original presentation will increase, which in turn will produce the increasing portion of the inverted U-shape function. Simultaneously, as the lag between repetitions increases, the probability of successful reminding will decrease, which will produce the decreasing portion of the inverted U-shape function. Thus, memory performance will increase as lag increases until the point at which reminding fails, after which the benefit of spaced study will not be obtained, producing the inverted U-shape function. Critically, the desirable difficulty hypothesis does not identify the dimensions of difficulty throughout the learning process that are critical for promoting effective learning. Specifically, desirable difficulties during the learning process should lead learners to use "more elaborate encoding processes and more substantial and varied retrieval processes" (Bjork, 1994; p 192), which stands in contrast with experiences of difficulty that do not yield such a change. As originally proposed, the desirable difficulty hypothesis provided examples of encoding manipulations that introduce difficulty (e.g., spacing, retrieval practice) but did not provide an explicit way of manipulating difficulty during the learning process to produce optimal difficulty. Thus, subsequent research has attempted to operationalize difficulty in a way in which the hypothesis can be evaluated.

Two ways of operationally defining difficulty that have been most explored include acquisition accuracy and response latency to detect repetitions. The obvious choice for a measure of difficulty is learning phase accuracy. To the extent that study-phase retrieval or reminding occurs successfully, one can infer that various lag conditions are differentially difficult. As Jacoby (1974) noted, one must be concerned with differences in item difficulty that confound condition difficulty. In other words, reminding or study-phase retrieval will be less successful for more difficult items regardless of the learning condition, and in turn, any differences in long-term memory observed between different lags or spacing conditions may simply reflect differences in item difficulty. Importantly, Jacoby's paradigm utilized a category recognition structure in which the reminding judgment was whether the item currently being studied (e.g., cherry) belonged to the same semantic category as the immediately preceding item (e.g., apple; a one-back condition) or any previously studied item (n-back condition). Results were not contaminated by item difficulty effects but nonetheless revealed a role for reminding difficulty as indexed by accuracy. In addition to acquisition phase accuracy to detect repetitions, which may not always provide the most sensitive indicator of learning difficulty, researchers have considered response latency for detecting repetitions as another proxy. Indeed, several notable studies have considered response latency for the detection of repetitions or for the retrieval of a previously studied item as an index of desirable difficulty (e.g., Glover, 1989; Karpicke & Roediger, 2007; Logan & Balota, 2008; Maddox & Balota, 2015; Pyc & Rawson, 2009). Critically, response latency has typically been confounded with lag in previous studies, so it is unclear to what extent memory improvement for the long lag, more difficult retrieval conditions can be attributed to the desirable difficulty account (Bjork, 1994) versus an alternative account of the spacing effect (e.g., encoding variability; see Glenberg, 1979; Delaney et al., 2010; and Maddox, 2016 for reviews). The current study was designed to assess the unique contributions of response latency during the acquisition of repeated study items as a proxy for retrieval difficulty above and beyond the contribution of other theoretical mechanisms proposed to underlie the spacing and lag effects. In this way we attempted to extend on previous research by operationally defining difficulty during the acquisition of material in order to explicitly predict (and subsequently test) this mechanism's effect on long-term memory.

Current study

The reminding account posited by Benjamin and Tullis (2010) incorporates a reminding mechanism (e.g.,

Hintzman 2004) and Bjork's (1994) desirable difficulty account to explain the non-monotonic function relating lag to final test performance. Past studies have often assumed that longer lags lead to more difficult reminding, but as has been previously noted (e.g., Maddox & Balota, 2015), this does not necessarily have to be true. For example, two lags that fall within the working memory capacity of the targeted population may not differ meaningfully in difficulty. In this situation, one would not expect a difference in performance despite differences in lag conditions.

Although reminding has typically been regarded as an automatic process, evidence has shown that reminding can be brought under conscious control (e.g., Wahlheim et al., 2014). To more fully examine differences in controlled reminding difficulty between lag conditions, participants in the current study completed a continuous recognition task in which they studied a list of items presented either one or two times. Thus, the continuous recognition response served as a repetition detection judgment on each trial (i.e., participants indicated whether the item was being presented for the first or second time), and, in turn, the continuous recognition task was used as a proxy for explicit reminding. Items were separated by various lags, and repetition detection response latencies on repetition trials during acquisition served as a proxy for reminding retrieval difficulty (c.f., Karpicke & Roediger, 2007; Logan & Balota, 2008; Maddox & Balota, 2015; Pyc & Rawson, 2009). The use of repetition detection response latency as a proxy for desirable difficulty yields two predictions that will be examined in the current experiments. These predictions are made with the assumption that the reminding difficulty account is the sole explanation of spacing and lag effects as a means for providing a strong test of this mechanism. However, as has been previously noted (see Delaney et al., 2010; Greene, 1989; Maddox, 2016), it is likely that the magnitude of the spacing effect is influenced by contributions from multiple mechanisms (e.g., the deficient processing account, Greeno, 1967; Rundus, 1971) even if those individual mechanisms cannot fully account for the consistent findings observed in the spacing effect literature. The first prediction states that in the absence of differences in difficulty across lag conditions during acquisition, no lag effect should be observed in the final test performance. It is only when lag conditions differ in difficulty during the acquisition of material that lag should influence subsequent memory performance. The second prediction states that at an item level, one should observe an increase in final test performance as response latency for items correctly detected as repetitions during acquisition increases regardless of the lag separating repetitions.

Experiment 1

Method

Participants and design Forty-four Rhodes College undergraduate students¹ received partial course credit for their participation. A single factor within-participants design was used in which the lag between repetitions of an item was one or five intervening items (Lag 1 vs. Lag 5, respectively).

Materials A continuous recognition task was used for the acquisition phase of the memory task. Sixty words were selected as critical stimuli and were divided into two sets that were counterbalanced across lag conditions. Stimuli were statistically equated across sets for word length and frequency (Balota et al., 2007; $p_s > .20$). An additional set of 18 words served as once-presented filler items, and the lexical characteristics of these items were statistically equivalent to the characteristics of the critical items, $p_s > .15$. Filler items were included to ensure that average serial position across lag conditions was equated in terms of first and second presentations ($p_s > .55$). Thus, there was no confound in average retention interval across lag conditions. In total, the acquisition phase included 142 trials consisting of 120 trials for the critical conditions, 18 once-presented filler trials, and four trials that were equally split between primacy and recency buffer items. Finally, 78 additional items were selected as recognition test lures from the same database. These items shared similar lexical characteristics to the target items previously studied during acquisition ($p_s > .50$).

Procedure Participants were instructed to study a list of words for a final recognition test and that some items would be presented once and other items would be presented twice. Items were presented at a 4.5-s rate, and on each trial participants indicated via key press as quickly and accurately as possible whether or not they had previously studied the item. They used the remaining time to study the item for the final test. Following acquisition, participants completed a demographics

questionnaire (approximately 1 min). Participants then completed a final recognition test in which items were presented individually on the screen until an “old” or “new” response was entered by the participant.

Results

Acquisition Phase Repetition Detection Accuracy Proportion correct for repetition detection judgments is presented in the top panel of Table 1 as a function of presentation and lag. There are three observations to note in this table. First, accuracy was comparable across lag conditions. Second, repetition detection accuracy was higher on an item’s second presentation compared to its first presentation. Third, the increased accuracy observed on an item’s second presentation was similar across lag conditions. These observations were supported by results of a 2 (Presentation: 1st vs. 2nd) $\times$ 2 (Lag: 1 vs. 5) within-participants ANOVA. Results revealed a main effect of presentation, $F(1, 43) = 9.56$, $p = .003$, $\eta^2_p = .18$, reflecting more accurate judgments on an item’s second presentation ($M = .91$) than on its first presentation ($M = .88$). There was no effect of lag ($p > .55$) or interaction between factors ($p > .10$), suggesting that the difficulty of repetition detection was similar across lag conditions. To further substantiate the claim that repetition detection difficulty was similar across lag conditions and to examine the repetition event in isolation, proportion correct repetition detection for the second presentation was submitted to a t test which failed to reveal a significant difference, $t(43) = .75$, $p = .460$. These results suggest that the selected lag conditions yielded comparably difficult repetition detection events.

Acquisition phase repetition detection response latency

Mean response latency for correct repetition detection judgments is presented in the middle panel of Table 1 as a function of presentation and lag. Outliers (response latencies < 200 ms) and trials without a participant response were removed from further analysis. Additionally, response latencies for correct trials were subsequently standardized within participants to account for interindividual differences in processing speed (see Faust, Balota, Spieler, & Ferraro, 1999). Specifically, an overall mean and standard deviation were calculated for each participant, and raw response latencies were converted to z scores by taking each raw response latency and subtracting the participant’s mean response latency and then dividing by the participant’s standard deviation. Outliers defined as standardized response latencies exceeding three standard deviations from the mean were removed from further analysis. Due to the skew of reaction time distributions, disproportionately more positive z scores were removed compared to negative z scores. Mean standardized response latency for correct repetition detection judgments is presented in the bottom panel of Table 1 as a function of presentation and lag. There are

¹ Given the novelty of the current paradigm, the predicted effect size in acquisition phase response latency between Lags 1 and 5 was unclear. Nonetheless previous studies indicate that longer lags are often associated with longer retrieval times (e.g., Karpicke & Roediger, 2007; Pyc & Rawson, 2009). Moreover, a reversed lag effect in acquisition accuracy is generally large in size when using a spaced retrieval paradigm (i.e., retrieval after shorter lags is easier than longer lags; see Balota, Duchek & Logan, 2007 for a review). Additionally, Karpicke and Roediger (2007) utilized a Lag 1 vs. Lag 5 comparison with 48 participants and reported a difference in response latencies with $p_{rep} = .90$. Our method utilized an encoding duration (4.5 s) nearly half of the duration (8 s) used by Karpicke and Roediger, which in turn should disproportionately impact continuous repetition detection judgments for Lag 5 items. Although we would expect the current sample size to provide sufficient statistical power for detecting a true effect, we report mediation analyses as a means of addressing concerns with the moderate sample size and analysis of response latencies (e.g., Fritz, Taylor, & MacKinnon, 2012; Kenny & Judd, 2013; MacKinnon, Lockwood, & Williams, 2004; Preacher, & Hayes, 2004).

Table 1 Mean (SEM) proportion of correct repetition detection judgments (top panel), raw response latencies in milliseconds (middle panel), and standardized response latencies (bottom panel) as a function of Lag and Presentation in Experiment 1

	Presentation 1	Presentation 2
Lag 1	.89 (.03)	.91 (.02)
Lag 5	.87 (.03)	.91 (.02)
	$M = .88 (.02)$	$M = .91 (.02)$
Lag 1	1230 (57)	1089 (40)
Lag 5	1214 (52)	1118 (37)
	$M = 1222 (53)$	$M = 1103 (37)$
Lag 1	.02 (.03)	-.22 (.03)
Lag 5	-.01 (.03)	-.15 (.03)
	$M = .01 (.02)$	$M = -.19 (.02)$

three observations to note. First, repetition detection response latency was quicker on an item's second presentation than on its first. Second, repetition detection response latency was similar across lag conditions. Third, the difference in response latencies across presentations was comparable across lag conditions. These three observations were supported by the results of a 2 (Presentation) $\times$ 2 (Lag) within-participants ANOVA. Results revealed a main effect of presentation, $F(1, 43) = 21.44, p < .001, \eta^2_p = .336$, which reflected faster responding to an item's second presentation ($M = -.19$) than its first presentation ($M = .01$). There was no effect of lag, ($F(1, 43) = .96, p > .33$). However, the Presentation $\times$ Lag interaction was marginally significant, $F(1, 43) = 3.74, p = .060, \eta^2_p = .08$. Although the difference between lag conditions was larger on the second presentation ($M_{\text{diff}} = .07$) than on the first presentation ($M_{\text{diff}} = .03$), follow-up t tests failed to reveal any significant differences ($ps > .05$), suggesting comparable difficulty across conditions on both presentations. Importantly, analysis of unstandardized (i.e., raw) response latencies yielded the same pattern of results such that there was a main effect of presentation, $F(1, 43) = 15.23, p < .001, \eta^2_p = .26$, no effect of lag, $p > .67$, and no interaction, $p > .17$. A follow-up t test failed to yield a significant difference in response latency across lag conditions on the second presentation, $p > .20$.

Final recognition accuracy Final test accuracy was examined three ways. First, we analyzed overall performance. Based on the predictions discussed in the *Introduction*, the lack of difference in repetition detection difficulty observed in acquisition accuracy and response latency suggests that there should be a null lag effect in final test performance (d') for items that were successfully identified as repeated items on their second presentation during the acquisition phase. Second, the relationship between acquisition phase response latency and final test hit rate was examined across all items and for each lag condition separately to further assess the desirable difficulty

account. Third, we utilized a mediation analysis to examine the possible role of repetition detection difficulty (i.e., response latency) in modulating the effect of lag on final test performance.

d' Final test hit rates for Lag 1 and Lag 5 ($M = .84, SEM = .02$ and $M = .88, SEM = .02$, respectively) and the false alarm rate ($M = .16; SEM = .02$) were used to calculate the signal detection measure d' . Overall accuracy (d') was submitted to a paired t test, which revealed a significant effect of lag, $t(43) = 3.15, p = .003, d = .47$, reflecting the typical benefit in memory for items separated by a long ($M = 2.48$) versus a short lag ($M = 2.26$). Given that these two lag conditions produced comparable levels of repetition detection difficulty during the acquisition phase, one would not have predicted a lag effect in final test performance according to assumptions of the desirable difficulty account (Bjork, 1994).

Response latency correlations To further examine the effect of repetition detection difficulty on final test performance, point-biserial correlations were calculated for each participant across lag conditions and within each lag condition. Specifically, standardized response latency on an item's second presentation when correctly recognized as a repetition was correlated with final test recognition accuracy (i.e., success vs. failure). When collapsed across lag conditions, the mean point-biserial correlation (M correlation = $-.06$) across participants² differed significantly from zero, $t(42) = 2.13, p = .040, d = .32$. Although not significant, this same pattern was observed for Lag 1 items ($M = -.05, p = .19$) and Lag 5 items ($M = -.06, p = .11$). These results suggest that response latency (i.e., retrieval difficulty) played a minimal role in modulating final test performance in the current paradigm. Interestingly, when there was a significant relationship, it functioned in the opposite direction predicted by the desirable difficulty account such that more easily retrieved items were better remembered than more difficult-to-retrieve items.

Mediation analysis The current and subsequent mediation analyses were conducted using the MEMORE macro in SPSS (Montoya & Hayes, 2017) in which the direct effect of lag on final recognition accuracy was examined, with first and second presentation standardized response latency for repetition detection judgments serving as potential mediators. Consistent with the previous analysis of d' , results from the mediation analysis revealed a direct effect of lag on final memory performance, $t(39) = 2.46, p = .019, d = .39$. More important for the question of interest was the estimated bootstrap confidence intervals (based on 5,000 bootstrap samples) for each of the two potential response latency mediators. The

² One participant was removed from the analysis due to a lack of variation in hit rate.

95% confidence interval for response latency on the second presentation was $-.013$ to $.016$, and the 95% confidence interval for response latency on the first presentation was $-.009$ to $.003$. The inclusion of the null value in each confidence interval suggests that neither variable significantly mediated the relationship between lag and final test performance.

Discussion

Analysis of acquisition phase repetition detection (i.e., reminding) accuracy and response latency suggested that Lags 1 and 5 yielded similarly difficult reminding conditions. Nonetheless, a significant lag effect was obtained in final test recognition performance (i.e., d'). Converging evidence from point-biserial correlation and mediation analyses further substantiate a limited role for desirable difficulty and reminding in accounting for the lag effect observed in the current paradigm.

Although the current results provide evidence inconsistent with the reminding and retrieval difficulty account of the lag effect (Benjamin & Tullis, 2010), the modulatory effect of reminding retrieval difficulty may only be observed with a more demanding retrieval test. Specifically, past research has indicated that the lag effect often increases in magnitude with more difficult final test assessments (e.g., free recall; see Maddox, 2016). In turn, increasing final test difficulty may reveal a role for the desirable difficulty account in explaining the lag effect; thus, Experiment 2 utilized the same acquisition paradigm but examined memory performance with a more difficult final free recall test.

Experiment 2

Method

Participants Forty-eight Rhodes College undergraduates received partial course credit for their participation.

Materials, design, and procedure The materials, design, and procedure were the same as those used in Experiment 1 with one exception. The final test format was free recall rather than recognition. For the free recall test, participants were asked to type as many words from the previously studied list as possible. Responses were scored as *correct* if they were correctly spelled or if they had no more than two spelling or transposition errors.

Results

Acquisition phase repetition detection accuracy The proportion of correct repetition detection judgments is presented in the top panel of Table 2 as a function of presentation and lag. There are three observations to note. First, repetition detection

Table 2 Mean (SEM) proportion of correct repetition detection judgments (top panel), raw response latencies in milliseconds (middle panel), and standardized response latencies (bottom panel) as a function of Lag and Presentation in Experiment 2

	Presentation 1	Presentation 2
Lag 1	.88 (.02)	.94 (.02)
Lag 5	.89 (.02)	.94 (.01)
	$M = .88$ (.02)	$M = .94$ (.01)
Lag 1	1293 (59)	1073 (38)
Lag 5	1252 (48)	1111 (39)
	$M = 1272$ (52)	$M = 1092$ (37)
Lag 1	.19 (.03)	-.21 (.03)
Lag 5	.13 (.03)	-.14 (.03)
	$M = .16$ (.02)	$M = -.18$ (.02)

accuracy was higher on an item's second presentation compared to its first. Second, repetition detection was similar across lag conditions. Third, the difference in repetition detection as a function of presentation was similar across lag conditions. These observations were supported by a 2 (Presentation) $\times$ 2 (Lag) within-participants ANOVA, which revealed a main effect of presentation, $F(1, 47) = 17.06$, $p < .001$, $\eta^2_p = .27$. This effect reflected increased accuracy on an item's second presentation ($M = .94$) relative to the first presentation ($M = .88$). Critically, the analysis failed to yield a significant effect of lag, $p > .55$, or a significant Presentation $\times$ Lag interaction, $p > .999$. These results replicate the results obtained in Experiment 1, suggesting that the lag conditions were comparably difficult in terms of the repetition detection event on an item's second presentation.

Acquisition phase repetition detection response latency

Mean response latency for correct repetition detection judgments is presented in the middle panel of Table 2 as a function of presentation and lag. Mean standardized response latency for correct repetition detection judgments is presented in the bottom panel of Table 2 as a function of presentation and lag. Three observations are noteworthy. First, mean response latency was faster on an item's second versus first presentation. Second, mean response latency was similar across lag conditions for each presentation. Third, there was a larger change in response latency from the first to the second presentation in the Lag 1 condition compared to the Lag 5 condition. These observations were supported by results of a 2 (Presentation) $\times$ 2 (Lag) within-participants ANOVA, which revealed a main effect of presentation, $F(1, 47) = 63.83$, $p < .001$, $\eta^2_p = .58$, reflecting slower response latency on an item's first versus second presentation ($M_s = .17$ vs. $-.20$). Although the analysis failed to reveal a significant effect of lag, $p > .90$, the Lag $\times$ Presentation interaction was significant, $F(1, 47) = 6.51$, $p = .014$, $\eta^2_p = .12$. Follow-up analysis revealed greater speeding

in response time from the first to second presentation in the Lag 1 condition ($M_{\text{diff}} = .41$) compared to the Lag 5 condition ($M_{\text{diff}} = .27$), but there were no significant differences in standardized response latency across lag conditions at either presentation, $p_s > .05$. Similar results were observed in the analysis of raw response latency.

Final recall accuracy Final test accuracy was again examined in three ways. First, Analysis of final test free recall performance revealed significantly better memory for Lag 5 ($M = .24$) versus Lag 1 items ($M = .18$), $t(47) = 4.40$, $p < .001$, $d = .64$, which is inconsistent with predictions based on a strict test of the desirable difficulty account given that difficulty during acquisition was statistically equivalent across lag conditions.

Response latency correlations Point-biserial correlations between acquisition retrieval latency on an item's second presentation and free recall were again calculated for each participant. When collapsed across lag conditions, the mean point-biserial correlation (M correlation = $-.03$) across participants did not differ significantly from zero, $t(47) = 1.50$, $p = .141$. This same pattern was observed for Lag 1 items (M correlation = $-.014$; $t(47) = 0.58$, $p > .55$) and Lag 5 items (M correlation = $-.032$; $t(47) = 1.06$, $p > .25$). These results replicate the recognition test data from Experiment 1 in suggesting that response latency (i.e., retrieval difficulty) plays a minimal role in modulating memory performance in this paradigm.

Mediation analysis Results from the mediation analysis revealed a direct effect of lag on final memory performance, $t(43) = 4.22$, $p < .001$, $d = .64$. More important for the question of interest was the estimated bootstrap confidence intervals (based on 5,000 bootstrap samples) for each of the two potential response latency mediators. The 95% confidence interval for response latency on the second presentation was $-.006$ to $.012$, and the 95% confidence interval for response latency on the first presentation was $-.008$ to $.007$. The inclusion of the null value in each confidence interval suggests that neither variable significantly mediated the relationship between lag and final test performance.

Discussion

The free recall results of Experiment 2 replicated the recognition results from Experiment 1. Analysis of acquisition phase repetition detection judgments indicated that Lag 1 and Lag 5 conditions yielded similar levels of reminding difficulty. In turn, statistically equivalent reminding difficulty should produce similar levels of final test performance across lag conditions. This was not observed, and instead, a benefit was obtained in free recall performance for Lag 5 items over Lag 1 items despite repetition detection difficulty being constant across conditions. Moreover, this conclusion was

substantiated by the null point-biserial correlations observed between repetition detection judgment response latency and final test performance for each individual participant, as well as the results of the mediation analysis.

Given the lack of support for the reminding difficulty account in the first two experiments, Experiment 3 utilized a massed condition and two lag conditions that were more distinct in nature. Moreover, elements of the encoding context were manipulated in an attempt to yield differences in repetition detection difficulty. In these ways Experiment 3 was designed to replicate and extend findings from Experiments 1 and 2.

Experiment 3

Evidence from the first two experiments suggested a limited role for desirable difficulty in accounting for the lag effects observed across recognition and recall performance in the current paradigm. To further substantiate these findings, the current experiment incorporated three notable changes. First, a massed repetition (i.e., Lag 0) condition was included in the study phase. Second, different lags were used to provide evidence of generalizability of findings from Experiments 1 and 2. Third, encoding conditions were explicitly manipulated or held constant across repetitions of items in an attempt to produce differences in repetition detection difficulty.

Method

Participants and design Forty-two Rhodes College undergraduate students received partial course credit or US\$10 monetary remuneration for their participation. A 2 (Encoding Variability: Constant, Variable) $\times$ 3 (Lag: 0, 3, 10) within-participants design was utilized. All items were presented visually in white font on a black background. Items were simultaneously presented aurally and the voice in which items were presented depended on the encoding condition to which they were assigned. Items assigned to the *constant encoding* condition were spoken in the same gender voice across repetitions, whereas items assigned to the *variable encoding* condition were spoken in different gender voices across repetitions. Participants were aware that they would hear items in both male and female voices. Repetitions of critical items occurred immediately (Lag 0) or were separated by three (Lag 3) or ten (Lag 10) intervening items.

Materials A continuous recognition task was used for the acquisition phase of the memory task. Sixty words were selected as critical stimuli and were divided into six sets that were counterbalanced across lag and encoding variability conditions. Stimuli were statistically equated on word length and

frequency (Balota et al., 2007; $ps > .40$). An additional set of 19 words served as once-presented filler items and possessed the same lexical characteristics as the critical items, $ps > .10$. Filler items were included to ensure that average serial position across lag and encoding variability conditions was equated for both first and second presentations ($ps > .45$). In total, the acquisition phase included 143 trials consisting of 120 trials for the critical conditions, 19 once-presented filler trials, and four trials that were equally split between primacy and recency buffer items. Finally, 60 additional items were selected as recognition test lures from the same database. Lures shared similar lexical characteristics to the target items previously studied during acquisition ($ps > .30$).

Procedure Similar to Experiments 1 and 2, participants first completed an acquisition phase in which they were instructed to learn a series of words presented visually on a computer screen (white font on a black background) and aurally via headphones. Words were spoken in either a male or female voice. For each item, they were asked to make a repetition detection judgment (i.e., “Has this word been presented before?”). Participants were informed that some items would be presented once and other items would be presented twice and that all “yes” or “no” responses should be made as quickly as possible. Items were presented for 5 s each, and participants were told to use any time remaining following each repetition detection judgment to study the item for an upcoming memory test.

Following the acquisition phase participants completed a brief demographics questionnaire. They were then asked to freely recall as many studied items as possible for 5 min. After completion of the recall test, participants took a recognition test in which the 60 critical items and 60 lure items were presented randomly for a yes/no recognition judgment.

Results

Acquisition phase repetition detection accuracy The proportion of correct repetition detection judgments is presented in the top panel of Table 3 as a function of presentation and lag. There are three observations to note. First, repetition detection accuracy was similar across all lag and encoding variability conditions. Second, repetition detection judgments were more accurate on second presentation trials than on first presentation trials. Third, the difference between first and second presentations was comparable across lag and encoding variability conditions. These observations were supported by the results of a 2 (Presentation) $\times$ 2 (Encoding Variability) $\times$ 3 (Lag) repeated measures ANOVA, which revealed main effects of lag, $F(2, 82) = 4.09$, $p = .020$, $\eta^2_p = .09$, and presentation, $F(1, 41) = 5.55$, $p = .023$, $\eta^2_p = .12$. Follow-up comparisons across lag conditions revealed a single significant difference between Lag 0 ($M = .86$) and Lag 3 ($M = .89$) performance, $p = .010$.

Table 3 Mean (SEM) proportion of correct repetition detection judgments (top panel), raw response latencies in milliseconds (middle panel), and standardized response latencies (bottom panel) as a function of Lag, Encoding Condition, and Presentation in Experiment 3

	Constant encoding		Variable encoding	
	Presentation 1	Presentation 2	Presentation 1	Presentation 2
Lag 0	.83 (.03)	.87 (.03)	.85 (.03)	.89 (.03)
Lag 3	.88 (.03)	.91 (.03)	.85 (.03)	.91 (.02)
Lag 10	.84 (.03)	.90 (.03)	.86 (.03)	.89 (.03)
	$M = .85 (.03)$	$M = .89 (.03)$	$M = .85 (.03)$	$M = .90 (.03)$
Lag 0	922 (57)	742 (39)	924 (53)	797 (50)
Lag 3	992 (64)	819 (45)	977 (55)	853 (42)
Lag 10	915 (51)	973 (48)	897 (56)	836 (36)
	$M = 943 (54)$	$M = 811 (41)$	$M = 932 (52)$	$M = 829 (40)$
Lag 0	.02 (.05)	-.31 (.06)	.02 (.05)	-.35 (.05)
Lag 3	.07 (.05)	-.29 (.04)	.11 (.05)	-.25 (.04)
Lag 10	.04 (.07)	-.16 (.04)	.02 (.05)	-.26 (.05)
	$M = .04 (.03)$	$M = -.25 (.03)$	$M = .05 (.03)$	$M = -.29 (.02)$

Similar to previous experiments, the main effect of presentation revealed significantly better performance on second presentation trials ($M = .89$) compared to first presentation trials ($M = .85$). Although there were no interactions with presentation, a follow-up 2 (Encoding Variability) $\times$ 3 (Lag) ANOVA was conducted for performance on the second presentation to confirm there were no differences in repetition detection accuracy. Results failed to reveal any significant effects, $ps > .15$.

Acquisition phase repetition detection response latency³

Mean response latency for correct repetition detection judgments is presented in the middle panel of Table 3 as a function of presentation, lag, and encoding variability condition. The corresponding mean standardized response latencies are presented in the bottom portion of Table 3. Results of a 2 $\times$ 2 $\times$ 3 repeated measures ANOVA of standardized response latency revealed a single main effect of presentation, $F(1, 41) = 59.06$, $p < .001$, $\eta^2_p = .59$, which reflected faster standardized response latency on second presentation ($M = -.27$) compared to first presentation ($M = .05$). Again, a follow-up 2 (Encoding Variability) $\times$ 3 (Lag) ANOVA was conducted for performance on the second presentation to confirm there were no differences in repetition detection response latency. Results failed to reveal any significant effects, $ps > .10$.

³ In contrast with the previous experiments, analysis of raw response latency yielded a pattern of results that diverged from analysis of standardized response latency. Specifically, analysis of raw response latency revealed a main effect of presentation, $F(1, 41) = 25.16$, $p < .001$, $\eta^2_p = .38$. Additionally, the main effect of lag was significant, $F(2, 82) = 8.08$, $p = .001$, $\eta^2_p = .17$, and further qualified by a lag $\times$ presentation interaction, $F(2, 82) = 7.21$, $p = .001$, $\eta^2_p = .15$.

Final free recall Final test performance is presented in Fig. 1 as a function of encoding variability, condition, and lag. There are two observations to note. First, performance generally increased with increases in lag. Second, the benefit of Lag 10 was disproportionately larger in the variable encoding condition relative to the constant encoding condition.

Results from a 2 (Encoding Variability) $\times$ 3 (Lag) repeated measures ANOVA revealed a main effect of lag, $F(2, 82) = 10.33$, $p < .001$, $\eta^2_p = .20$, a marginal effect of encoding variability, $F(1, 41) = 3.95$, $p = .053$, $\eta^2_p = .09$, and a significant Lag $\times$ Encoding Variability interaction, $F(2, 82) = 4.43$, $p = .015$, $\eta^2_p = .10$. Separate analysis of the constant encoding condition revealed a marginal effect of lag, $F(2, 82) = 2.511$, $p = .087$, $\eta^2_p = .06$, whereas analysis of the variable encoding condition revealed a significant effect of lag, $F(2, 82) = 13.40$, $p < .001$, $\eta^2_p = .25$. Bonferroni corrected comparisons revealed significantly higher performance in the Lag 10 condition relative to Lag 0 and Lag 3 conditions ($ps < .01$).

Recognition Although recognition test performance may be influenced by the preceding free recall test, the extent to which performance is consistent with the pattern observed in the previous experiments may be informative. Proportion of hits (i.e., identifying a studied item as studied) was submitted to a 2 (Encoding Variability) $\times$ 3 (Lag) repeated measures ANOVA. Results revealed a main effect of lag, $F(2, 82) = 7.37$, $p = .001$, $\eta^2_p = .15$, which reflected a spacing effect for both Lag 3 and Lag 10 conditions ($ps < .05$) but no difference between performance in the two lag conditions ($p > .85$).

Response latency correlations Point-biserial correlations between acquisition retrieval latency on an item's second presentation and free recall were calculated for each participant. When collapsed across lag conditions, the mean point-biserial correlation (M correlation = $-.03$) across participants did not differ significantly from zero, $t(41) = 1.54$, $p = .132$. This same pattern was observed across all conditions when analyzed separately as a function of lag and encoding variability (rs ranging from $-.05$ to 0 ; $ps > .40$). A similar pattern was observed when analyzing the point-biserial correlations for recognition performance collapsed across all conditions (M correlation = $-.01$; $t(42) = .36$, $p = .723$), individually for each lag condition (rs ranging from $-.06$ to $.04$; $ps > .55$), and for each a combination of conditions, (rs ranging from $-.10$ to $.03$, $ps > .15$).

Mediation analysis Provided that the MEMORE macro only allows for a comparison of variables with two levels and the extension of this macro is currently under development, multiple mediation analyses were conducted to compare each combination of lag conditions within the variable encoding condition given the significant interaction observed in overall final free recall performance. Critically, the direct effect of lag on final test performance was significant in all three

Fig. 1 Mean (error bars are ± 1 SEM) proportion free recall as a function of Lag and Encoding Variability condition in Experiment 3

comparisons, $ts > 2.20$, and $ps < .05$. Similar to the previous mediation analyses, the more important result for the question of interest was the estimated bootstrap confidence intervals (based on 5,000 bootstrap samples) for each of the two potential response latency mediators across comparisons of the three lag conditions. All of the 95% confidence intervals for standardized response latency on the first and second presentations included the null value, which suggests that neither variable significantly mediated the relationship between lag and final test performance. Specifically, the 95% confidence intervals for the potential mediation effects were calculated for the Lag 0 vs. Lag 10, Lag 0 vs. Lag 3, and Lag 3 vs. Lag 10 comparisons. The confidence intervals for the potential mediation effect of standardized response latency on the item's second presentation were: $-.018$ to $.017$, $-.006$ to $.030$, and $-.020$ to $.022$, respectively. The confidence intervals for the potential mediation effect of standardized response latency on the item's first presentation were: $-.017$ to $.011$, $-.007$ to $.014$, and $-.016$ to $.017$, respectively. Again, with respect to recognition test performance in the current experiment, one must be concerned that the recognition test always followed the free recall test. However, it is important to note that similar to the free recall results, all of the 95% confidence intervals for standardized response latency on the first and second presentations included the null value.

Discussion

Experiment 3 extended on the first two experiments with the inclusion of a massed condition, more distinct lag conditions (Lag 3 vs. Lag 10), and explicitly manipulated encoding variability conditions. Analysis of final free recall performance was inconsistent with the desirable difficulty account when making predictions based on acquisition phase accuracy and standardized response latency. This was also true for recognition performance. However, it is important to note that the recognition test came after the cued recall test, which may limit any strong conclusions about the relationship between

desirable difficulty and recognition test performance in this experiment. These conclusions were further substantiated by point-biserial correlation analyses and the results from mediation analyses.

General discussion

The aim of the current study was to provide a strong test of the desirable difficulty mechanism (Bjork, 1994) and its contributions to the reminding process as a means of accounting for the lag and spacing effects. We extended on past examinations of desirable difficulty by operationally defining difficulty and assessing its relationship with long-term memory performance. Specifically, we utilized repetition detection accuracy and standardized response latency as proxies for reminding difficulty across different lag conditions to test whether more difficult reminding events (as indicated by longer response latencies) led to enhanced memory performance. There were three critical and consistent findings across experiments to be considered in evaluating predictions. First, short and long lag conditions (1 vs. 5 and 0 vs. 3 vs. 10 items, respectively) produced equally difficult reminding on an item's second presentation as indicated by statistically equivalent accuracy and mean response latency. Despite equally difficult reminding across lag conditions, a significant lag effect was observed in final test performance (recognition and free recall). Second, analysis of response latency as a continuous variable suggested that difficult (but successful) repetition detection events actually led to *worse* final test performance than easier repetition detection events (Experiment 1) or had no effect on final test performance (Experiments 2 and 3). Third, mediation analyses across all three experiments revealed a direct effect of lag on final test performance but failed to reveal mediation effects of response latency on either presentation. We first consider the extent to which the reminding account emphasizing desirable difficulty (Benjamin & Tullis, 2010) can accommodate these consistent findings before considering the implications of the current results for other mechanistic explanations of the lag effect.

Reminding and desirable difficulty

The desirable difficulty account (Bjork, 1994) rests on the assumption that more (versus less) difficult successful retrieval during the acquisition phase of material will yield superior memory. When lag conditions produce differentially difficult reminding events, one would predict a lag effect in final test performance, but when lag conditions produce equally difficult reminding events, no lag effect should be observed in final test performance. As previously noted, the original desirable difficulty framework did not specify the dimensions of difficulty that would be desirable and lead to changes in

encoding strategies, and it did not specify ways of operationally defining and manipulating difficulty during learning. Thus, to the extent that reminding accuracy and response latency are sufficient proxies for desirable difficulty, the consistent finding of a lag effect in final test recognition and free recall performance despite equivalent reminding across lag conditions during the acquisition phase in the current paradigm is contrary to predictions based on the desirable difficulty account. Moreover, point-biserial correlation analyses between repetition detection response latency and final test performance from Experiment 1 suggested that items with successful, *easy* repetition detection were more likely to be retrieved on a final memory test than items with successful, *difficult* repetition detection. Of course, this finding must be taken with caution given the small magnitude of the correlation and considering that this correlation was not significant in the second or third experiments. It is also important to note that despite all items having been stored sufficiently on their initial learning event, evidenced by correctly recognizing items on their second presentation, our item-level analyses may still confound item difficulty with condition-level difficulty. That is, it is possible that longer response latencies are the result of multiple distinct processes (e.g., desirably difficult and successful reminding, successful reminding based on guessing rather than retrieval). Nonetheless, the point-biserial correlation analyses converge on the broader pattern of data in suggesting that the difficulty of repetition detection did not strongly modulate the lag effect across the three experiments described herein that relied on operationally defining difficulty as a combination of accuracy and response latency. Although the current results are inconsistent with the desirable difficulty account of spacing and lag effects, they underscore the need for explicitly defining, measuring, and manipulating difficulty during the learning process to more directly evaluate this framework.

Encoding variability and study-phase retrieval dual mechanism

An alternative to combining reminding with desirable difficulty is to combine encoding variability and study-phase retrieval mechanisms (Greene, 1989; Raaijmakers, 2003). The increasing portion of the inverted U-shape function relating lag and long-term memory performance can be explained by increased encoding variability (Glenberg, 1979). As lag increases it is assumed that encoding is more variable due to changes in context (e.g., mental, physical, and experimental contexts). In contrast, when the target item is separated by a short lag during acquisition, it is more likely to occur adjacent to the same item(s), which in turn would produce fewer retrieval routes during later retrieval. The decreasing portion of the function is attributed to failed study-phase retrieval (Thios & D'Agostino, 1976) that functions in a way

similar to the reminding account previously outlined. Specifically, study-phase retrieval (i.e., reminding) will eventually fail once the lag between repetitions of an item is too long, and as a result, the benefit of repetition will not be obtained in final test performance. Although there are concerns about this mechanism's ability to accommodate all of the consistent findings in the spacing and lag effect literatures (e.g., the finding that increasing the lag between repeated study events leads to super-additive performance relative to predicted performance for remembering one of two different items separated by the same lag; see Benjamin & Tullis, 2010), these concerns are in relation to the assumptions outlined in original versions of this mechanism (e.g., Melton, 1970) and can be accommodated by subsequent instantiations of the encoding variability account proposed to account for the empirical findings driving contemporary research (e.g., Glenberg, 1979; see Maddox, 2016 for a review). Thus, it is important to consider the extent to which the current findings can be accommodated by the encoding variability and study-phase retrieval mechanism.

To examine the influence of encoding variability on memory performance, many studies have relied on the intuitive assumption that items repeated after longer lags are more likely to be encoded in differing ways than items repeated after shorter lags given variability in list context. To substantiate this conclusion, we compared the number of unique items that appeared on either side of target items for each lag condition. Results revealed that Lag 5 items appeared adjacent to significantly more unique words on average than Lag 1 items, $t(58) = 6.22$, $p < .001$, which suggests that encoding context varied to a greater extent in the Lag 5 condition than the Lag 1 condition. In turn, the lag effect in final test performance, despite a lack of difference in repetition detection difficulty, is consistent with the encoding variability mechanism.

Experiment 3 explicitly manipulated encoding context in an attempt to produce varying degrees of difficulty in repetition detection between conditions (context change vs. no change). Analysis of final cued recall performance in Experiment 3 revealed a unique benefit for encoding variability in the long lag condition compared to the massed and short lag conditions. Although one may expect intentionally incorporating encoding variability to have its greatest influence in conditions that otherwise lack sufficient variability as a result of the list context (e.g., lag 0 and lag 3), one interpretation of the current pattern of results is that incorporating variability in ways that are not salient or critical for task completion may limit their effectiveness in inducing encoding variability, and this may be particularly true when (a) repetitions occur after a short lag and (b) the way material was initially encoded is still accessible. For example, the gender of the speaker was not important for completing the current task successfully. Thus, relatively little attention may have been directed to this

dimension of the stimulus. The combination of variable gender voices and list context may have been more successful in inducing variable encoding than short lag conditions in which the stimulus was likely to be interpreted in the same way across repetitions. Of course, this interpretation is speculative, and future studies may wish to replicate the current findings and capitalize on past paradigms in which experimental and physical contexts were explicitly manipulated in a range of ways to further evaluate this mechanistic account of the lag effect (e.g., Appleton-Knapp, Bjork, & Wickens, 2005; Hintzman et al., 1975; Slamecka & Barlow, 1979; Verhoeven, Rikers, & Schmidt, 2004, 2005).

Potential directions for future research may involve extending the range of lags beyond those used in the current study. Although we modified the lag comparisons in Experiment 3, future studies may wish to use more widely varying lag conditions (e.g., Pyc & Rawson, 2009). Additionally, incorporating repeated restudy opportunities may more clearly show that multiple mechanisms contribute to the continued strengthening of an item in memory (cf., Maddox & Balota, 2015). Finally, future studies may extend on the current approach or explore alternative ways of operationally defining and explicitly testing the role of desirable difficulty in producing the lag effect. The original instantiation of the desirable difficulty account (Bjork, 1994) provided a mechanism to explain how memory is strengthened through differences in acquisition conditions but left the mechanism and ways of measuring and manipulating it underspecified. In the current study we examined acquisition accuracy and addressed issues of inter-individual variability and the distribution shape of acquisition response latencies through standardizing response latency at the individual level and removing outliers (both before and after standardization). Neither measure served as an adequate proxy of difficulty or predictor of long-term memory in the current paradigm. Thus, identifying successful ways of measuring and manipulating forms of desirable difficulty provides substantial fodder for future studies.

We evaluated both accuracy and response latencies in the current paradigm because we believe that together they provide a more sensitive measure of reminding difficulty than accuracy alone. To extend on this approach, future studies may wish to obtain measures of confidence or remember/know/guess judgments as a means of examining the contributions of different bases of retrieval in the reminding process. Although past studies suggest that recollection-based versus familiarity-driven judgments yield different response latencies (e.g., Wiseman & Ishai, 2008; Woodruff, Hayama, & Rugg, 2006) and that response latency often differs as a function of confidence (e.g., Kahana, 2012; see also Merkow, Burke, & Kahana, 2015), one may expect that recollection-based reminding would yield different memorial consequences than familiarity-based reminding. Indeed, an examination of recollection- versus familiarity-based reminding may be

useful in identifying meaningful dimensions of encoding variability and their contributions to long-term memory. Moreover, examining differences in remember, know, and guess-based reminding may help address concerns related to item-specific effects within our item-level analyses discussed above. Thus, the current study's approach to explicitly defining and testing desirable difficulty's contributions to the lag effect is an initial step in discriminating confounded mechanisms proposed to underlie the benefits of spaced repetitions.

Conclusion

In summary, the current paper utilized repetition detection response latency as a proxy for repetition detection difficulty to assess the contributions of repetition detection difficulty to the lag effect. With reminding difficulty operationally defined in this way, results from the current study could not be accommodated by the reminding account that relies on desirable difficulty (e.g., Benjamin & Tullis, 2010) to explain the lag effect. Instead, the consistent finding of a lag effect in final recognition and recall performance despite equivalent repetition detection difficulty during the acquisition phase can be accommodated by accounts incorporating a reminding process with encoding variability (e.g., Greene, 1989; Raaijmakers, 2003).

Clearly, more research is needed to better understand the role of encoding variability in producing the memory benefit of spacing, the extent to which changes in environment are detected and preserved when reminding occurs, and how such variability across repetitions of material may serve to modify retrieval cues. In addition, future studies should further examine response latency as well as other proxies for reminding difficulty to more fully and explicitly evaluate the desirable difficulty account. Moreover, it may be the case that multiple mechanisms contribute in different capacities to the lag effect depending upon specific lags used within a given paradigm. Nonetheless, our results suggest that change in context across repetitions (as indicated by our follow-up analysis of unique items surrounding Lag 1 and Lag 5 items in the list structure for Experiments 1 and 2) and the explicit manipulation of encoding context in Experiment 3 may be a critical consideration for enhancing memory.

Author Note Portions of this work were supported by NIA Training Grant AG00030 awarded to David Balota.

References

- Appleton-Knapp, S., Bjork, R. A., & Wickens, T. D. (2005). Examining the spacing effect in advertising: Encoding variability, retrieval processes and their interaction. *Journal of Consumer Research*, 32, 266–276. <https://doi.org/10.1086/432236>
- 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 extant literature. To appear in Nairne, J.S. (Ed.), *The foundations of remembering: Essays in honor of Henry L. Roediger III*. Chapter 6, pp. 83–105.
- Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. I., Kessler, B., Loftis, B., ... Treiman, R. (2007b). The English Lexicon Project. *Behavior Research Methods*, 39, 445–459.
- Benjamin, A. S., & Tullis, J. G. (2010). What makes distributed practice effective? *Cognitive Psychology*, 61, 228–247. <https://doi.org/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. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). Cambridge: MIT Press.
- 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. <https://doi.org/10.1037/0033-2909.132.3.354>
- Crowder, R. G. (1976). *Principles of learning and memory*. Hillsdale: Erlbaum.
- Delaney, P. F., Verkoeijen, P. P. J. L., & Spigler, A. (2010). Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. *Psychology of Learning and Motivation*, 53, 63–147. [https://doi.org/10.1016/S0079-7421\(10\)53003-2](https://doi.org/10.1016/S0079-7421(10)53003-2)
- Dempster, F. N. (1996). Distributing and managing the conditions of encoding and practice. In E. L. Bjork & R. A. Bjork (Eds.), *Memory* (pp. 317–344). San Diego: Academic Press. <https://doi.org/10.1016/B978-012102570-0/50011-2>
- Ebbinghaus, H. (1885). *Über das Gedächtnis*. New York: Dover.
- 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.
- Fritz, M. S., Taylor, A. B., & MacKinnon, D. P. (2012). Explanation of two anomalous results in statistical mediation analysis. *Multivariate Behavioral Research*, 47, 61–87. <https://doi.org/10.1080/00273171.2012.640596>
- Glenberg, A. M. (1979). Component-levels theory of the effects of spacing of repetitions on recall and recognition. *Memory & Cognition*, 7, 95–112. <https://doi.org/10.3758/BF03197590>
- Glover, J. A. (1989). The “testing” phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology*, 81, 392–399.
- 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. <https://doi.org/10.1037/0278-7393.15.3.371>
- Greeno, J. G. (1967). Paired-associate learning with short term retention: Mathematical analysis and data regarding identification of parameters. *Journal of Mathematical Psychology*, 4, 430–472. [https://doi.org/10.1016/0022-2496\(67\)90033-8](https://doi.org/10.1016/0022-2496(67)90033-8)
- Hintzman, D. L. (2004). Judgment of frequency versus recognition confidence: Repetition and recursive reminding. *Memory & Cognition*, 32, 336–350. <https://doi.org/10.3758/BF03196863>
- Hintzman, D. L. (2010). How does repetition affect memory? Evidence from judgments of recency. *Memory & Cognition*, 38, 102–115. <https://doi.org/10.3758/MC.38.1.102>
- Hintzman, D. L., Summers, J. J., & Block, R. A. (1975). Spacing judgments as an index of study-phase retrieval. *Journal of Experimental Psychology: Human Learning and Memory*, 1, 31–40. <https://doi.org/10.1037/0278-7393.1.1.31>
- Jacoby, L. L. (1974). The role of mental contiguity in memory: Registration and retrieval effects. *Journal of Verbal Learning and Verbal Behavior*, 13, 483–496. [https://doi.org/10.1016/S0022-5371\(74\)80001-0](https://doi.org/10.1016/S0022-5371(74)80001-0)

- Jacoby, L. L., & Wahlheim, C. N. (2013). On the importance of looking back: The role of recursive reminders in recency judgments and cued recall. *Memory & Cognition*, 41, 625–637. <https://doi.org/10.3758/s13421-013-0298-5>
- Jacoby, L. L., Wahlheim, C. N., & Yonelinas, A. P. (2013). The role of detection and recollection of change in list discrimination. *Memory & Cognition*, 41, 638–649. <https://doi.org/10.3758/s13421-013-0313-x>
- Kahana, M. J. (2012). *Foundations of human memory*. New York: Oxford University Press.
- Karpicke, J. D., & Roediger III, H. L. (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. <https://doi.org/10.1037/0278-7393.33.4.704>
- Kenny, D. A., & Judd, C. M. (2013). Power anomalies in testing mediation. *Psychological Science*, 25, 334–339. <https://doi.org/10.1177/0956767613502676>
- Logan, J. M., & Balota, D. A. (2008). Expanded vs equal spaced retrieval practice in healthy young and older adults. *Aging, Neuropsychology, and Cognition*, 15, 257–280. <https://doi.org/10.1080/13825580701322171>
- MacKinnon, D. P., Lockwood, C. M., & Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research*, 39, 99–128. https://doi.org/10.1207/s15327906mbr3901_4
- Maddox, G. B. (2016). Understanding the underlying mechanism of the spacing effect in verbal learning: A case for encoding variability and study phase retrieval. *Journal of Cognitive Psychology*. <https://doi.org/10.1080/20445911.2016.1181637>
- Maddox, G. B., & Balota, D. A. (2015). Retrieval practice and spacing effects in and young and older adults: An examination of the benefits of desirable difficulty. *Memory & Cognition*, 45(5), 760–774. <https://doi.org/10.3758/s13421-014-0499-6>
- Madigan, S. A. (1969). Intraserial repetition and coding processes in free recall. *Journal of Verbal Learning and Verbal Behavior*, 8, 828–835. [https://doi.org/10.1016/S0022-5371\(69\)80050-2](https://doi.org/10.1016/S0022-5371(69)80050-2)
- Melton, A. W. (1970). The situation with respect to the spacing of repetitions and memory. *Journal of Verbal Learning and Verbal Behavior*, 9, 596–606. [https://doi.org/10.1016/S0022-5371\(70\)80107-4](https://doi.org/10.1016/S0022-5371(70)80107-4)
- Merkow, M. B., Burke, J. F., & Kahana, M. J. (2015). The human hippocampus contributes to both the recollection and familiarity components of recognition memory. *Proceedings of the National Academy of Sciences*, 112, 14378–14383. <https://doi.org/10.1073/pnas.1513145112>
- Montoya, A. K., & Hayes, A. F. (2017). Two condition within-participant statistical mediation analysis: A path-analytic framework. *Psychological Methods*, 22(1), 6–27.
- Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, & Computers*, 36, 717–731.
- 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. <https://doi.org/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. [https://doi.org/10.1016/S0364-0213\(03\)00007-7](https://doi.org/10.1016/S0364-0213(03)00007-7)
- Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, 1, 181–210. <https://doi.org/10.1111/j.1745-6916.2006.00012.x>
- Rundus, D. (1971). Analysis of rehearsal processes in free recall. *Journal of Experimental Psychology*, 89, 63–77.
- Slamecka, N. J., & Barlow, W. (1979). The role of semantic and surface features in word repetition effects. *Journal of Verbal Learning and Verbal Behavior*, 18, 617–627. [https://doi.org/10.1016/S0022-5371\(79\)90344-X](https://doi.org/10.1016/S0022-5371(79)90344-X)
- Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory*, 4, 592–604. <https://doi.org/10.1037/0278-7393.4.6.592>
- Soderstrom, N. C., Kerr, T. K., & Bjork, R. A. (2016). The critical importance of retrieval – and spacing – for learning. *Psychological Science*, 27, 223–230. <https://doi.org/10.1177/0956797615617778>
- 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. [https://doi.org/10.1016/0022-5371\(76\)90047-5](https://doi.org/10.1016/0022-5371(76)90047-5)
- Tullis, J. G., Benjamin, A. S., & Ross, B. H. (2014). The reminding effect: Presentation of associates enhances memory for related words in a list. *Journal of Experimental Psychology: General*. <https://doi.org/10.1037/a0036036>
- Verkoeijen, P. P. J. L., Rikers, R. M. J. P., & Schmidt, H. G. (2004). Detrimental influence of contextual change on spacing effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30, 796–800. <https://doi.org/10.1037/0278-7393.30.4.796>
- Verkoeijen, P. P. J. L., Rikers, R. M. J. P., & Schmidt, H. G. (2005). Limitations to the spacing effect: Demonstration of an inverted Ushaped relationship between interrepetition spacing and free recall. *Experimental Psychology*, 52, 257–263. <https://doi.org/10.1027/1618-3169.52.4.257>
- Wahlheim, C. N., Maddox, G. B., & Jacoby, L. L. (2014). The role of reminding in the effects of spaced repetitions on cued recall: Sufficient but not necessary. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40, 94–105. <https://doi.org/10.1037/a0034055>
- Wiseman, W., & Ishai, A. (2008). Recollection- and familiarity-based decisions reflect memory strength. *Frontiers in Systems Neuroscience*, 2, 1–9. <https://doi.org/10.3389/neuro.06.001.2008>
- Woodruff, C. C., Hayama, H. R., & Rugg, M. D. (2006). Electrophysiological dissociation of the neural correlates of recollection and familiarity. *Brain Research*, 1100(1), 125–135. <https://doi.org/10.1016/j.brainres.2006.05.019>