CrossMark

Level of initial training moderates the effects of distributing practice over multiple days with expanding, contracting, and uniform schedules: Evidence for study-phase retrieval

Thomas C. Toppino[1] · Heather-Anne Phelan[1] · Emilie Gerbier[2]

Abstract

Inconsistent results have been obtained in experiments comparing the effects on retention of *expanding*, *contracting*, and *uniform* practice schedules, in which the spacing between successive practice sessions progressively increases, progressively decreases, or remains constant, respectively. In the present study, we experimentally assessed an apparent trend in the literature for expanding schedules to be more advantageous than other schedules following a low level of training during the initial learning session, but not following a high level of initial training. College students studied pseudoword–word pairs in multiple practice sessions distributed over a 13-day period according to expanding, contracting, and uniform schedules. During their initial learning session, participants received either low-level training (two study trials) or high-level training (one study trial and then five rounds of practice testing with corrective feedback). All participants were treated identically in the subsequent practice sessions. A final cued-recall test after a two-week retention interval revealed an expanding-schedule superiority following low-level initial training but not following high-level initial training. These results are interpreted in terms of a study-phase-retrieval mechanism and help explain the mixed results obtained in the prior literature.

The hallmark of learning through practice is repetition. Retention of information generally improves with greater amounts of practice, and the extent of the improvement depends on how the practice is distributed, or spaced, over time (e.g., Ebbinghaus, 1885/1964). If two occurrences of the same information occur in immediate succession (i.e., massed practice), repetition often has little or no beneficial effect. As the lag or spacing interval between occurrences increases, performance generally improves markedly, although the spacing interval can eventually become too long (Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006).

The majority of the research on distributed practice has focused on the effect of varying the spacing interval between two occurrences of the same information. (See Toppino & Gerbier, 2014, for a recent review.) There is no doubt that this line of research has contributed greatly to our knowledge and understanding of distributed-practice effects. However, learning in the real world rarely involves only two practice opportunities, prompting some researchers to study the effects of distributed practice involving multiple (i.e., three or more) practice trials or practice sessions (e.g., Landauer & Bjork, 1978; Tsai, 1927). Much of this research has examined the effect of schedules that differ in how the lags between successive practice opportunities change during learning when the number of presentations and the total amount of spacing are held constant. Three kinds of schedule—*expanding*, *contracting*, and *uniform*—have been investigated most frequently. In *expanding* schedules, the spacing intervals are relatively short initially and then become progressively longer with practice. In *contracting* schedules, the spacing intervals are relatively long initially and then become progressively shorter with practice. In *uniform* schedules, the spacing intervals remain constant throughout practice. The present article is primarily concerned with the comparative effects of these schedules on a final test of retention.

A consensus explanation for distributed-practice effects has long eluded researchers, but the majority of accounts involve one or more of three proposed theoretical mechanisms:

✉ Thomas C. Toppino
thomas.toppino@villanova.edu

[1] Department of Psychological & Brain Sciences, Villanova University, 800 Lancaster Avenue, Villanova, PA 19085, USA

[2] Université Côte d'Azur, CNRS, BCL, Nice, France

study-phase retrieval, *encoding variability*, or *deficient processing* (Toppino & Gerbier, 2014). Most of the theoretical work has taken place in the context of research that has varied the spacing between two presentations of to-be-remembered events, but the proposed mechanisms also should be relevant for the effects of schedules involving three or more presentations.

According to study-phase-retrieval theories, a repetition is effective in improving retention to the extent that it triggers retrieval of a prior occurrence (e.g., Thios & D'Agostino, 1976). As the spacing interval increases between successive practice sessions, successful study-phase retrieval becomes more difficult and requires processing that has a greater beneficial effect on subsequent retention (e.g., Benjamin & Tullis, 2010; Toppino & Gerbier, 2014). With respect to schedules involving three or more presentations of the to-be-remembered information, the operation of a study-phase-retrieval mechanism implies that there should be conditions in which expanding schedules will lead to superior final-test performance, as compared to other schedules. Using relatively short spacing intervals in the beginning of learning, when memory traces are weak and fragile, should work to increase the likelihood of study-phase retrieval and avoid the retrieval failures that would occur with longer lags. Theoretically, this allows the weak trace to be strengthened so that it can bridge a longer interval preceding the next study opportunity. By gradually increasing the spacing interval over successive practice trials, an expanding schedule presumably maintains the likelihood of successful study-phase retrieval, allowing memory to benefit from progressively longer lags.

Encoding-variability theories (e.g., Bower, 1972; Glenberg, 1979) are descendants of stimulus-sampling theory (Estes, 1955a, 1955b). As such, the stimulus elements, including contextual cues, are learned in an all-or-none manner, and memory performance improves with repeated practice as a result of learning more stimulus elements rather than of strengthening the degree to which the stimulus elements are learned. As applied to distributed practice, Bower (1972) and Glenberg (1979) emphasized the role of contextual cues, which are assumed to guide encoding and may be represented in the encoded information. Subsequent retrieval success is assumed to be greater to the extent that there is overlap between the contextual cues present during encoding and those that are available in the retrieval environment. Finally, contextual cues are assumed to fluctuate with the passage of time and events (i.e., *contextual drift*; Bower, 1972), so that increasing the spacing between presentations of an item increases the likelihood that it will be encoded differently (e.g., with different contextual information) on each occurrence. This, in turn, improves memory performance because it increases the chances that some aspect of the encoded information will match information in the retrieval environment. The exact predictions, however, depend on the retention interval

(Glenberg, 1979). When the retention interval is long relative to the spacing intervals, the contextual cues at retrieval are assumed to be randomly related to the cues that were in effect during encoding. With respect to schedules involving three or more presentations of repeated information, this leads to a prediction of no difference between expanding and contracting schedules, because these schedules typically use the same spacing intervals presented in opposite orders. The predicted relative performance with a uniform schedule is indeterminate.[1] In contrast, if the retention interval is short relative to the spacing intervals, the contextual cues present during the final test may overlap greatly for the information that was encoded most recently, because there has been little opportunity for the contextual cues to fluctuate. In these cases, performance will be determined primarily by the overlap between the cues in the retrieval environment and those in effect during the most recent study opportunities. Consequently, a contracting schedule should have an advantage over other schedules, because the last *two* (or more) study opportunities will have occurred near the final test, thereby creating more chances for overlap with the retrieval cues.

Finally, from the perspective of deficient-processing theories, transient processes (e.g., temporary memory activation) cause learners to be unable or unwilling to process a repetition fully if it occurs too soon after its previous presentation, thereby leading to poorer retention (e.g., Russo, Parkin, Taylor, & Wilks, 1998; Shaughnessy, Zimmerman, & Underwood, 1972). Massed repetition may lead to virtually no benefit for memory, but increased spacing leads to more processing and, consequently, to improved memory performance. Due to the transient nature of the underlying process, a deficient-processing mechanism should influence the effect of schedules primarily when they involve very short spacing intervals, on the order of seconds. In this case, an advantage for uniform schedules might be expected, because the short spacing intervals of expanding and contracting schedules would be more likely to result in severe deficient processing, thereby impairing retention.

Although research on schedules of distributed practice involving more than two study opportunities per item has a long history (e.g., Tsai, 1927), it did not capture the interest of the research community until the publication of an article by Landauer and Bjork (1978). In that research, participants

---

[1] Küpper-Tetzel, Kapler, and Wiseheart (2014) contended that a uniform schedule should yield superior final-test performance after a long retention interval, because the encoding opportunities in a uniform schedule are spread evenly through the acquisition phase of an experiment. However, it is not clear that the mixture of short and long spacing that occurs in both expanding and contracting schedules necessarily produces less overall encoding variability than the sequence of medium spacing intervals that are encountered in uniform schedules. This and the fact that experimenters usually equate the total amounts of spacing for all schedules suggest that performance with a uniform schedule may not be expected to differ from that of the other two standard schedules.

initially studied a set of word pairs (e.g., first and last names in Exp. 1) before completing three cued-recall practice tests (without corrective feedback), in which only the first name of the pair was presented and participants tried to supply the corresponding surname. The practice tests were presented according to an expanding, contracting, or uniform schedule, with spacing intervals ranging from 0 to 90 s in length. The total amount of spacing was equated across schedules. The results indicated that expanding schedules produced superior final-test performance. Landauer and Bjork explained this finding in a manner reminiscent of study-phase retrieval. That is, short initial spacing intervals presumably increased the likelihood of successful practice-test retrieval when memories were weak at the beginning of learning. However, subsequent research has produced conflicting results: Expanding schedules have sometimes yielded superior retention (e.g., Cull, Shaughnessy, & Zechmeister, 1996, Exps. 1–4; Maddox, Balota, Coane, & Duchek, 2011, Exp. 2; Storm, Bjork, & Storm, 2010, Exps. 2 and 3), sometimes yielded no retention advantage (e.g., Balota, Duchek, Sergent-Marshall, & Roediger, 2006; Carpenter & DeLosh, 2005; Cull, 2000, Exps. 1 and 2; Karpicke & Bauernschmidt, 2011; Karpicke & Roediger, 2010; Maddox et al., 2011, Exp. 1), and sometimes yielded retention that was inferior to that produced by a uniform schedule, particularly when there was a relatively long delay before the final test (e.g., Karpicke & Roediger, 2007; Logan & Balota, 2008).

These mixed results may be attributable partially to the paradigm that was used, in which acquisition was completed in a single experimental session that entailed a series of practice tests without feedback, separated by short spacing intervals. As a result of the short lags, the probability of successful retrieval on the practice tests often differed little as a function of schedule, thereby eliminating the purported benefit of expanding schedules unless, for example, forgetting was induced by interpolating an interfering task between practice tests (Storm et al., 2010). In addition, when practice-test retrieval was successful irrespective of schedule, the shortest spacing intervals used in the expanding schedules may have impaired retention on the final test, perhaps due to the involvement of a deficient-processing mechanism. Reduced retrieval demands and the detrimental effects of very short spacing intervals also may have contributed to why schedule effects are typically absent when acquisition occurs in a single session and the repeated practice involves either repeated study or practice tests followed by feedback (e.g., Carpenter & DeLosh, 2005, Exp. 2; Cull, 2000, Exps. 1 and 2; Cull et al., 1996, Exp. 5; Landauer & Bjork, 1978, Exp. 2).

An alternative paradigm for studying the effects of schedules of distributed practice uses spacing and retention intervals measured in days and weeks and either re-presents information for practice or provides practice tests followed by corrective feedback. This approach avoids many of the limitations associated with using a series of short lags to test schedule effects involving three or more practice opportunities, and it may have greater implications for application due to the use of educationally relevant temporal intervals. Unfortunately, once again, conflicting results have been obtained.

In one recent study, Gerbier, Toppino, and Koenig (2015) asked French participants to study pseudoword–word pairs once on each of three study trials, which were distributed according to expanding, contracting, and uniform schedules. The numbers of days separating the first and second and the second and third study sessions were 1 and 11, 11 and 1, and 6 and 6, for the expanding, contracting, and uniform schedules, respectively. The final test of cued recall was administered after retention intervals of 2, 6, or 13 days. Consistent with the operation of a study-phase-retrieval mechanism, the results indicated an overall expanding-schedule superiority, with a trend for the advantage of an expanding schedule to increase with the length of the retention interval. Expanding-schedule superiority also has been obtained by Gerbier and Koenig (2012, Exp. 1) and by Tsai (1927). Gerbier and Koenig (2012, Exp. 2) also obtained expanding-schedule superiority, although the advantage was only reliable in comparison to a contracting schedule.[2] Other experiments have failed to obtain expanding-schedule superiority, usually finding no significant difference among schedules (e.g., Cull, 2000, Exps. 3 and 4; Kang, Lindsey, Mozer, & Pashler, 2014), although one study (Küpper-Tetzel et al., 2014) found that, depending on the retention interval, a contracting schedule could be better or worse than expanding and uniform schedules, which did not differ from each other. This unusual pattern of results cannot be explained by any existing theory. Table 1 of the article by Gerbier et al. (2015) presents a detailed summary of the literature involving long spacing intervals and repeated study or practice tests with feedback.

A close examination of the multisession literature suggests a potential resolution of the empirical findings, which is related to the extent of practice on the first practice session and is consistent with the operation of a study-phase retrieval mechanism (Gerbier et al., 2015). When the experiments yielded expanding-schedule superiority, participants' first study session involved a low level of practice in which the to-be-remembered items were presented for study only once or a few times (Gerbier & Koenig, 2012; Gerbier et al., 2015; Tsai, 1927). From a study-phase-retrieval perspective, a low level of initial learning would require a relatively short initial spacing interval in order to ensure study-phase retrieval upon the next study opportunity, thereby allowing the memory trace to be strengthened so that it could benefit from the progressively

---

[2] The participants in this experiment were required to make a recognition decision during repeated practice in an effort to directly assess study-phase retrieval. This procedure may have reduced the advantage of an expanding schedule by inducing a higher level of study-phase retrieval in all schedules than would have occurred spontaneously (Gerbier et al., 2015).

longer lags to follow. When experiments did *not* result in expanding-schedule superiority, participants usually engaged in a high level of practice during the first study session, which included multiple practice trials involving cued-recall tests followed by corrective feedback (Cull, 2000, Exps. 3 and 4; Kang et al., 2014; Küpper-Tetzel et al., 2014). This form of practice is known to be especially effective, because retrieval practice enhances retention more than restudying (e.g., Roediger & Karpicke, 2006) and also potentiates or enhances learning from subsequent study opportunities, such as those provided by feedback (e.g., Arnold & McDermott, 2013). In terms of theory, a high level of initial learning would be more likely to enable a high degree of study-phase retrieval regardless of the initial spacing interval, thereby reducing or eliminating the differences among schedules.[3]

Although the proposed empirical and theoretical resolution of the conflicting findings in the literature may be intriguing, it is not convincing. The experiments differed in multiple ways in addition to the level of initial practice, raising the possibility that other, unidentified variables were critical in accounting for the mixed results, and leaving open the possibility of an alternative theoretical account. In the present experiment, we tested the proposed resolution directly. Participants studied pseudoword–word pairs in multiple practice sessions distributed over 13 days according to expanding, contracting, and uniform schedules. Two experimental groups (*low training* vs. *high training*) received different levels of practice during the first practice session but were treated identically in all other respects. We expected the overall final cued recall after a two-week retention interval to be lower in the low- than in the high-training condition. However, for the reasons described previously, the study-phase-retrieval hypothesis predicts that expanding-schedule superiority should be obtained in the low- but not in the high-training condition.

Predictions based on the operation of an encoding-variability mechanism (Glenberg, 1979) depend on the retention interval, which we did not vary in this experiment. If our two-week retention interval is construed to be long relative to the spacing intervals, encoding variability could accommodate a finding of no difference among schedules or, possibly, of uniform-schedule superiority (Küpper-Tetzel et al., 2014), although we previously questioned the latter prediction. If our retention interval is construed to be short relative to the spacing intervals, encoding variability could accommodate a finding of contracting-schedule superiority. However, regardless of the retention interval, it is not clear how an encoding-

variability mechanism could accommodate expanding-schedule superiority by participants in either initial training condition.

Method

Design

The experiment had a 2 × 3 mixed factorial design involving initial training level (low vs. high), which was varied between participants, and practice schedule (expanding vs. contracting vs. uniform), which was manipulated within participants. There were three practice sessions involved in each schedule, with the spacing intervals between the first and second sessions and between the second and third sessions, respectively, being 1 and 11 days for the expanding schedule, 11 and 1 days for the contracting schedule, and 6 and 6 days for the uniform schedule, resulting in five practice sessions distributed over 13 days, in which the words from each schedule appeared in only three (see Fig. 1).

Participants

A total of 142 college students volunteered to participate in six experimental sessions distributed over a 27-day period. Those who satisfactorily completed the experiment received class credit and the opportunity to enter a cash lottery. However, 44 participants failed to do so, as a result of missing an experimental session ($n = 34$), noncompliance during unsupervised sessions ($n = 3$), or technological issues ($n = 7$). This resulted in a final sample of 98 participants, 49 of whom were assigned to each of the two initial-training-level groups.[4]

Previous experiments with a similar methodology (e.g., Gerbier & Koenig, 2012) had used a sample size in the mid-30s for a within-participants design. We sought to increase the number of participants per group by about one-third to ensure sufficient power. Subsequently, a priori power analyses using G*Power 3 (Faul, Erdfelder, Lang, & Buchner, 2007) indicated that we could have safely used fewer participants and still achieved a true power of .90 for effect sizes only half as large as those obtained in similar previous studies (Gerbier & Koenig, 2012; Gerbier et al., 2015).

Materials

The stimulus materials consisted of 72 word pairs, each containing one pronounceable pseudoword (e.g., "proome") and one concrete English noun (e.g., "jacket"). All pseudowords

---

[3] This analysis of the literature excludes an experiment by Clark (1928) that did not obtain a difference due to schedules. The method of study during the first practice session of that experiment was not controlled, making it impossible to categorize her experiment according to the extent or type of initial practice.

[4] Of the 44 participants who failed to satisfactorily complete the experiment, 23 and 21 had been assigned to the high-training and low-training conditions, respectively.

Fig. 1 Participation calendar. The letters E, U, and C refer to the schedule assignments, representing the sets of eight-item repeating lists assigned to the expanding-, uniform-, and contracting-schedule conditions, respectively. The letters f_1, f_2, and f_3 represent the three sets of 16 once-presented filler stimuli

came from the ARC Nonword Database (Rastle, Harrington, & Coltheart, 2002), and all English nouns came from the MRC Psycholinguistic Database, Version 2 (Wilson, 1988) or from Van Overschelde, Rawson, and Dunlosky's (2004) revision of Battig and Montague's (1969) category norms.

In each of the five practice sessions, participants studied a set of 24 pairs. To manipulate practice schedule within participants, we adapted the method used by Gerbier et al. (2015). The 24 pairs were assigned to three repeating sublists containing eight pairs each, while the remaining 48 pairs served as nonrepeating filler items (f_1, f_2, and f_3 in Fig. 1) that were used to maintain a consistent list length across sessions. For a given participant, each repeating sublist was assigned to a different practice-schedule condition. As is shown in Fig. 1, the sublist assigned to the expanding schedule was part of the 24-pair study set on Days 1, 2, and 13; the sublist assigned to the contracting schedule was part of the study set on Days 1, 12, and 13; and the sublist assigned to the uniform schedule was part of the study set on Days 1, 7, and 13. There were six possible ways to assign the three repeating sublists to the three practice schedules. Each combination was assigned to a different randomly determined subgroup of participants.

Each of the 24 repeating nouns (targets) belonged to a different semantic category, whereas the nonrepeating nouns were selected in groups of three from an additional 16 semantic categories. The nouns were distributed such that no semantic category was represented more than once per practice session. There were no forward or backward associations between items within or across the repeating and filler lists, on the basis of the University of South Florida Free Association Norms (Nelson, McEvoy, & Schreiber, 2004).

Procedure

Participants completed all sessions online using the Qualtrics platform. Up to three participants at a time (assigned to the same initial-training condition) completed Session 1 (i.e., Day 1) on computers in a Villanova University laboratory, supervised by one of the authors (H.-A.P.). All participants were instructed to carefully study the pairs and to try to memorize them for a later recall test. However, they were instructed not to write down any of the words or mentally repeat them when they were not on the screen. Following the instructions, all participants viewed a randomly intermixed arrangement of all 24 repeating pseudoword–word pairs, each presented once for 8 s and separated by 1.5-s blank screens. Then, the participants in the low-training condition simply re-viewed the 24 pairs in a new random order, whereas the participants in the high-training condition completed five consecutive rounds of cued-recall practice testing on these items, each presented in an independent random order. During each round, the pseudoword member of each pair appeared alone for 8 s while participants attempted to type the corresponding "English translation." After 8 s, the entire pseudoword–word pair was presented for 3 s, followed by a 1.5-s blank screen. At the start of each testing round, participants again were reminded to use the 3-s feedback that followed every practice test as an additional restudy opportunity.

Participants completed Sessions 2–5 and the final test (Session 6) remotely using their own devices. The day before each practice session, participants received an e-mail reminder of the date and the 8-h window within which to complete the session. At the start of that 8-h window, participants received another e-mail containing a temporary hyperlink (unique to each participant and set to expire after 8 h), which granted individual, one-time access to the appropriate session.

The procedures of Sessions 2–5 resembled those of the first trial of Session 1. At the outset of each session, participants again were instructed to carefully study the pairs and to try to memorize them for a later recall test. They also were reminded not to write down any of the words or mentally repeat them when they were not on the screen. During each session, all participants viewed a total of 24 pseudoword–word pairs once, for 8 s each, separated by 1.5-s blank screens. As is illustrated in Fig. 1, eight of the pairs studied during Session 1 were randomly intermixed with 16 once-presented filler pairs in Sessions 2–4. The eight-item sublists assigned to the expanding, uniform, and contracting schedules were included in the sets of items studied on Session 2 (Day 2), Session 3 (Day 7), and Session 4 (Day 12), respectively. In Session 5 (Day 13), participants viewed all 24 repeating pairs without fillers.

As a check on participants' engagement during Sessions 1–5, simple text-entry questions (e.g., *Please type the following*

numbers in the box below: 93847620) were randomly intermixed in each session. Participants were informed that failure to answer any of these questions within 20 s would result in their dismissal from the study.

After a 14-day retention interval (Day 27), participants completed a final cued-recall test on all 24 repeating pseudoword–word pairs. The final test mimicked the practice tests from the high-training Session 1, except that the pseudoword cues appeared alone for up to 20 s and there were no attention-checking questions. To help detect the possibility that final recall was contaminated by written records of the items that had been presented during Sessions 2–5, the final test ended with six of the once-presented items, for which unaided recall was unlikely. None of the participants recalled any of these items.

Results

We first analyzed the performance of participants in the high-training condition on the five training trials involving cued recall with feedback that took place in Session 1. The percentage of correct recall on each of the five test trials was submitted to a 5 (successive test trials) × 3 (set of items assigned to each schedule) analysis of variance (ANOVA), with repeated measures on both factors. The results indicated that training performance improved markedly over successive test trials, $F(4, 192) = 304.27$, $MSE = 271.883$, $p < .001$, $\eta_p^2 = .89$. The mean percentages correct on Test Trials 1–5, respectively, were 19.98, 40.14, 62.67, 77.38, and 85.54. Neither the main effect of item set nor the interaction approached significance, $F(2, 96) = 0.70$, $MSE = 328.54$, and $F(8, 384) = 1.25$, $MSE = 121.57$, respectively, indicating that there were no differences in difficulty among the sets of items assigned to the three practice-schedule conditions.

The dependent measure of primary interest was the percentage of correct cued recall on the final test, administered in Session 6. The data were submitted to a 2 (initial-training level) × 3 (practice schedule) ANOVA with repeated measures on the second factor.

The results shown in Fig. 2 reveal a significant main effect of initial training, $F(1, 96) = 126.31$, $MSE = 979.57$, $p < .001$, $\eta_p^2 = .57$, such that high training produced better overall performance than low training. There was also a significant main effect of practice schedule, $F(2, 192) = 5.34$, $MSE = 196.47$, $p = .006$, $\eta_p^2 = .05$ (Ms = 45.28, 41.39, and 38.78, for the expanding, uniform, and contracting schedules, respectively), but both main effects were qualified by a significant Practice Schedule × Initial Training interaction, $F(2, 192) = 8.40$, $MSE = 196.47$, $p < .001$, $\eta_p^2 = .08$.

The data were also examined by conducting a Bayesian ANOVA (bANOVA) in order to estimate Bayes factors (BF; see Morey, Rouder, & Jamil, 2015; Wagenmakers, 2007). The data were fitted and then compared in different statistical models that included the two factors and their interaction, following the recommendations of Rouder, Morey, Verhagen, Swagman, and Wagenmakers (2017). This analysis indicated that the best model was the one that included both main effects and their interaction. A BF of 9.38×10^{17} to 1 indicated very strong evidence in favor of the existence of the main effect of initial training, as did the BF of 235 to 1 in favor of the existence of the main effect of practice schedule. A BF of 78 to 1 indicated strong evidence in favor of the Initial Training × Practice Schedule interaction.

To probe the interaction, the effect of practice schedule was analyzed separately at each level of initial training. Although practice schedule had no reliable effect on high-training participants' cued recall, $F(2, 96) = 1.07$, $MSE = 208.56$, $p = .35$, $\eta^2 = .02$ (BF = 10.9 to 1 in favor of the absence of a schedule effect), the effect of schedule was significant for low-training
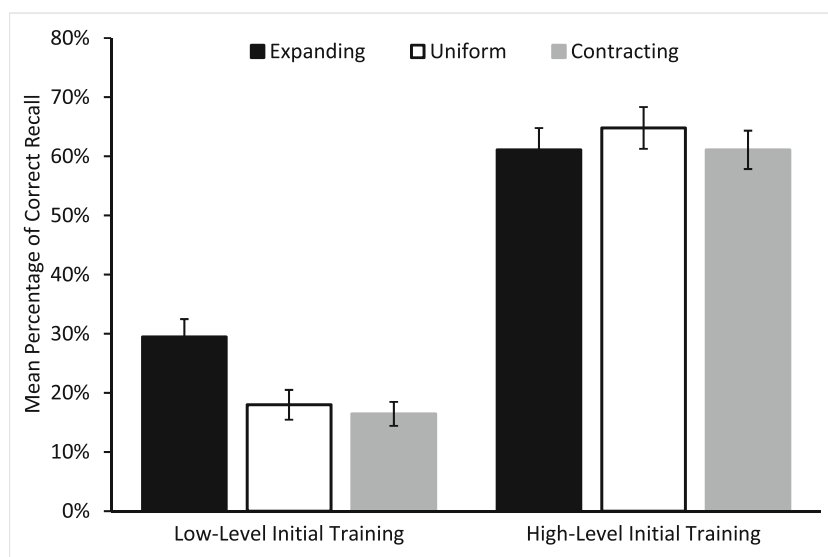

Fig. 2 Final cued recall as a function of the level of initial training and type of practice schedule. Error bars indicate ±1 *SEM*

participants, $F(2, 96) = 13.44$, $MSE = 184.38$, $p < .001$, $\eta^2 = .22$ (BF = 44.3 to 1 in favor of the schedule effect). The data for the low-training participants were further probed using two-tailed paired-samples t tests (with Bonferroni adjustments). These analyses revealed that practice with an expanding schedule led to better final recall performance than did practice with a uniform schedule, $t(48) = 4.03$, $p < .001$, $d = 0.591$, (BF = 123 to 1 in favor of a difference) or with a contracting schedule, $t(48) = 4.74$, $p < .001$, $d = 0.725$, (BF = 1,039 to 1 in favor of a difference). However, the uniform and contracting schedules did not differ significantly in terms of their effects on final recall, $t(48) = 0.58$, $p = .56$ (BF = 5.49 to 1 in favor of an absence of difference).

In terms of the design of the experiment, the sublists of repeating items were crossed with the practice-schedule conditions, producing six unique ways in which the three 8-item repeated sublists could be combined with the three schedule conditions. Each combination was assigned to a different randomly determined subgroup of participants within each initial-training condition. The numbers of participants in the final sample who were assigned to each of the resulting 12 subgroups (6 ways of combining lists and schedules × 2 initial-training conditions) were almost evenly distributed, with seven to nine participants per subgroup. However, as a precaution, a second set of analyses was performed after the data from zero to two participants were randomly eliminated from each subgroup, to create a sample with a uniform membership of seven participants per subgroup. The results of these additional analyses were virtually identical to the results obtained with the full dataset.

Discussion

Our primary result was that an expanding schedule led to better final cued recall than did a uniform or contracting schedule when participants engaged in a low level of training during the first of three practice sessions separated by one or more days. The difference among schedules was eliminated when the initial practice session involved a high level of training. This corresponds to a trend in the literature noted by Gerbier et al. (2015), in which expanding-schedule superiority has been obtained following a low level of initial training (e.g., Gerbier & Koenig, 2012; Gerbier et al., 2015; Tsai, 1927), but not following a high level of initial training (e.g., Cull, 2000; Kang et al., 2014; Küpper-Tetzel et al., 2014). Our results provide experimental evidence that the level of initial training is a causal factor influencing the relative efficacy of expanding, uniform, and contracting schedules, and not a spurious consequence of multiple methodological differences among experiments.

Our results support study-phase retrieval (Benjamin & Tullis, 2010; Thios & D'Agostino, 1976; Toppino &

Gerbier, 2014) as a theoretical mechanism contributing to distributed-practice effects. Of the three most prominent mechanisms that have been proposed to underlie distributed-practice effects, study-phase retrieval is the only one that seems able to account for our results. A critical assumption of study-phase retrieval is that, for a spaced repetition to be effective, it must prompt retrieval of a prior occurrence of the repeated information. When the initial level of learning is low, the resulting memory trace will be weak and subject to rapid forgetting. Hence, a relatively short spacing interval such as that provided by an expanding schedule may be necessary for study-phase retrieval to be successful during the next study opportunity, whereas the longer initial spacing intervals of uniform and contracting schedules may be too long. Given the successful study-phase retrieval enabled by an expanding schedule, the initially weak trace may be strengthened sufficiently to bridge the longer and more beneficial spacing intervals to come. In contrast, when participants engage in extensive training initially, the resulting memory representation may be strong enough to survive the initial spacing interval, regardless of the schedule being used. This would minimize performance differences among schedules in part because of the standard practice used in our experiment, in which schedules were equated with respect to the total amount of spacing between practice sessions.

The study-phase-retrieval explanation of our findings is similar to the one offered by Landauer and Bjork (1978) and by others (e.g., Storm et al., 2010) for expanding-schedule superiority when this result has been obtained in the single-session paradigm involving a series of practice tests without feedback. This theory assumes that the short initial lags of expanding schedules facilitate performance by increasing the likelihood of retrieval on the initial practice tests, without which a learning opportunity would be lost. Although this process is analogous to the study-phase-retrieval account of our results, there is an important difference: Failure to retrieve on a practice test when no corrective feedback is provided prevents learning, because the target information is literally not available for practice. It is as though the practice trial had simply been omitted. In contrast, if study-phase retrieval fails when an item is presented for repeated study, as in our experiment, the target information *is* available for practice, although the theory purports that the beneficial effect for the learner will be limited.

The other prominent theoretical mechanisms commonly proposed to underlie distributed-practice effects cannot account for our finding. Our use of spacing intervals measured in days seems to have precluded substantial involvement of a deficient-processing mechanism, which is thought to be based on transitory processes with an effect that is typically measured in seconds. An encoding-variability mechanism should have been operable under the conditions of our experiment but seems unable to explain the results. Our expanding and

contracting schedules used the same spacing intervals presented in opposite orders and, thus, should have produced similar degrees of variable encoding by the end of acquisition. As we described in the introduction, an encoding-variability mechanism (Glenberg, 1979) could predict findings ranging from schedule equivalence to contracting-schedule superiority, depending on whether our two-week retention interval is construed to be relatively short or relatively long. However, we see no obvious way for existing accounts of an encoding-variability mechanism to predict expanding-schedule superiority in either of our initial-training conditions.

Our results specifically support the assumption of study-phase-retrieval theory that study-phase retrieval must be successful for repeated, spaced practice to be effective in improving memory. Although this assumption may seem self-evident, it is not included in all theories. For example, encoding-variability theory (e.g., Bower, 1972; Glenberg, 1979) predicts that repetition will be most effective in fostering future retrieval when each presentation is encoded entirely independently (e.g., no overlapping cues), even though independent encoding would preclude study-phase retrieval according to the theory. (We will return to this point later.)

Our results did not address why greater spacing produces better memory when study-phase retrieval is successful. Presumably, the act of study-phase retrieval somehow strengthens the memory. Benjamin and Tullis (2010) offered a straightforward hypothesis, in this regard. They proposed that study-phase retrieval (or "reminding") is more difficult after a longer spacing interval and that more difficult retrieval potentiates memory to a greater degree. Toppino and Gerbier (2014) proposed an abstraction process to explain why the beneficial effect of successful study-phase retrieval becomes greater with increasing spacing. Retrieval is assumed to depend on cues in the retrieval environment that overlap with the cues originally encoded as part of the memory trace, and on the strength of the associations between those cues and the to-be-retrieved information. Study-phase retrieval is proposed to differentially strengthen cue–target associations for the specific overlapping cues that were responsible for the successful retrieval. As the spacing between repetitions increases, contextual drift, or the fluctuation of contextual cues over time, causes a decrease in the number of overlapping cues, while there is a simultaneous increase in the proportion of overlapping cues that are slow-changing and durable. Therefore, as spacing increases, the strengthening effect of retrieval becomes more focused or concentrated on a smaller set of cues that are more likely to include durable cues that have the greatest chance of being available to mediate retrieval on a later memory test.

Another set of theories to be considered is hybrid theories that combine a study-phase-retrieval mechanism with an encoding-variability mechanism (e.g., Greene, 1989; Mozer, Pashler, Cepeda, Lindsey, & Vul, 2009; Raaijmakers, 2003).

These theories propose that study-phase retrieval is necessary for repeated practice to be effective but that, when study-phase retrieval is successful, the beneficial effect of spacing on memory is due to encoding variability. By invoking the necessity for study-phase retrieval, this class of theories can predict expanding-schedule superiority in at least some circumstances. For example, simulations conducted by Lindsey, Mozer, Cepeda, and Pashler (2009) indicated that their multiscale context model predicts expanding-schedule superiority with relatively long retention intervals. However, combining encoding variability and study-phase retrieval in a single theory raises questions about the degree to which these mechanisms are compatible in other ways. As we noted earlier, encoding-variability theories predict maximum performance when encodings are completely independent (i.e., no overlapping cues), but independent encoding precludes study-phase retrieval. Thus, the combination of study-phase retrieval and encoding variability seems to predict that memory for repeated information can approach, but not attain, the level that would be expected on the basis of independent encodings (i.e., the independence baseline). This poses a problem for the standard encoding-variability mechanism, because superadditive performance, which exceeds the independence baseline, occurs frequently in both free- and cued-recall tests (Begg & Green, 1988; Benjamin & Tullis, 2010). In contrast, the hybrid theories proposed by Mozer et al. (2009) and by Raaijmakers (2003) do not appear to include a true encoding-variability mechanism, because they abandon the critical all-or-none learning assumption of encoding variability. This change opens the possibility that such hybrid theories may be able to account for superadditive performance, although, to the best of our knowledge, this remains to be demonstrated.

Regardless of how one explains the improvement in memory performance with increased spacing, all theories that incorporate a study-phase retrieval mechanism postulate that increasing the spacing between practice sessions will have two opposing effects. As the spacing interval gets longer, the probability of successful study-phase retrieval declines but, if it is successful, the benefits of repeated practice are greater. This implies that the most beneficial level of spacing will be the longest possible lag that still allows successful study-phase retrieval. It also implies that optimal spacing will depend on a memory's susceptibility to forgetting. Theoretically, the optimal lag will be longer to the extent that the memory trace is stronger or otherwise is retained better, and that this optimum will become progressively longer as learning advances, consistent with an expanding schedule. More generally, in learning any unit of information, the effectiveness of any given spacing interval will vary with the degree to which the information is vulnerable to forgetting over that time period.

Results consistent with this view were obtained by Storm et al. (2010) when they compared the effectiveness of schedules in conditions in which all training occurred in a single

experimental session and involved retrieval practice without feedback. They found expanding-schedule superiority when they induced forgetting by filling the spacing interval with highly interfering material, but not when they filled the interval with relatively noninterfering material. Without introduced interference, their lags apparently were not long enough to produce the forgetting necessary for expanding-schedule superiority to be observed. Although we used a very different methodology (multiple practice sessions involving restudying rather than retrieval practice), our results converge with those of Storm et al. We used the level of initial training to vary susceptibility to forgetting and obtained expanding-schedule superiority when a low level of initial training made forgetting likely, but not when a high level of training created more durable memory traces. In the latter case, no initial spacing interval was long enough to produce sufficient forgetting to yield expanding schedule superiority.

Finally, as a cautionary note, the critical role of retention and forgetting suggests that the effectiveness of a given set of lags may vary as a function of a variety of factors that affect the susceptibility of the to-be-remembered information to forgetting over the spacing interval. For example, in addition to the degree of initial learning and the interfering potential of interpolated material, the effectiveness of a given set of lags may be influenced by the amount of information to be learned, the nature of the to-be-learned material, and even characteristics of the individual learners. Thus, it would be a mistake to assume that a particular set of lags would produce the same effect in all circumstances.

References

Arnold, K. M., & McDermott, K. B. (2013). Test-potentiated learning: Distinguishing between direct and indirect effects of tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 940–945. https://doi.org/10.1037/a0029199

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. https://doi.org/10.1037/0882-7974.21.1.19

Battig, W. F., & Montague, W. E. (1969). Category norms for verbal items in 56 categories: A replication and extension of the Connecticut norms. *Journal of Experimental Psychology*, *80*(3, Pt. 2), 1–46. https://doi.org/10.1037/h0027577

Begg, I., & Green, C. (1988). Repetition and trace interaction: Superadditivity. *Memory & Cognition*, *16*, 232–242.

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

Bower, G. H. (1972). Stimulus-sampling theory of encoding variability. In A. W. Melton & E. Martin (Eds.), Coding processes in human memory (pp. 85–123). Washington, DC: Winston.

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

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

Clark, B. E. (1928). The effect upon retention of varying lengths of study periods and rest intervals in distributed learning time. *Journal of Educational Psychology*, *19*, 552–559. https://doi.org/10.1037/h0070774

Cull, W. L. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology*, *14*, 215–235. https://doi.org/10.1002/(SICI)1099-0720(200005/06)14:3<215::AID-ACP640>3.0.CO;2-1

Cull, W. L., Shaughnessy, J. J., & Zechmeister, E. B. (1996). Expanding understanding of the expanding-pattern-of-retrieval mnemonic: Toward confidence in applicability. *Journal of Experimental Psychology: Applied*, *2*, 365–378. https://doi.org/10.1037/1076-898X.2.4.365

Ebbinghaus, H. E. (1964). *Memory: A contribution to experimental psychology* (H. A. Ruger, C. E. Bussenius, & E. R. Hilgard, Trans.). New York, NY: Dover. (Original work published in 1885)

Estes, W. K. (1955a). Statistical theory of distributional phenomena in learning. *Psychological Review*, *62*, 369–377. https://doi.org/10.1037/h0046888

Estes, W. K. (1955b). Statistical theory of spontaneous recovery and regression. *Psychological Review*, *62*, 145–154. https://doi.org/10.1037/h0048509

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191. https://doi.org/10.3758/BF03193146

Gerbier, E., & Koenig, O. (2012). Influence of multiple-day temporal distribution of repetitions on memory: A comparison of uniform, expanding, and contracting schedules. *Quarterly Journal of Experimental Psychology*, *65*, 514–525. https://doi.org/10.1080/17470218.2011.600806

Gerbier, E., Toppino, T. C., & Koenig, O. (2015). Optimising retention through multiple study opportunities over days: The benefit of an expanding schedule of repetitions. *Memory*, *23*, 943–954. https://doi.org/10.1080/09658211.2014.944916

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

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

Kang, S. H. K., Lindsey, R. V., Mozer, M. C., & Pashler, H. (2014). Retrieval practice over the long term: Should spacing be expanding or equal-interval? *Psychonomic Bulletin & Review*, *21*, 1544–1550. https://doi.org/10.3758/s13423-014-0636-z

Karpicke, J. D., & Bauernschmidt, A. (2011). Spaced retrieval: Absolute spacing enhances learning regardless of relative spacing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 1250–1257. https://doi.org/10.1037/a0023436

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

Karpicke, J. D., & Roediger, H. L., III. (2010). Is expanding retrieval a superior method for learning text materials? *Memory & Cognition*, *38*, 116–124. https://doi.org/10.3758/MC.38.1.116

Küpper-Tetzel, C. E., Kapler, I. V., & Wiseheart, M. (2014). Contracting, equal, and expanding learning schedules: The optimal distribution of learning sessions depends on retention interval. *Memory & Cognition*, *42*, 729–741. https://doi.org/10.3758/s13421-014-0394-1

Landauer, T. K., & Bjork, R. A. (1978). Optimum rehearsal patterns and name learning. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory* (pp. 625–632). London: Academic Press.

Lindsey, R., Mozer, M., Cepeda, N. J., & Pashler, H. (2009). Optimizing memory retention with cognitive models. In A. Howes, D. Peebles, & R. Cooper (Eds.), *Proceedings of the Ninth International Conference on Cognitive Modeling—ICCM2009, Manchester, UK.* Retrieved from www.yorku.ca/ncepeda/publications/LMCP2009. html

Logan, J. M., & Balota, D. A. (2008). Expanded vs. equal interval spaced retrieval practice: Exploring different schedules of spacing and retention interval in younger and older adults. *Aging, Neuropsychology, and Cognition*, *15*, 257–280. https://doi.org/10. 1080/13825580701322171

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. https://doi.org/10.1037/a0022942

Morey, R. D., Rouder, J. N., & Jamil, T. (2015). BayesFactor: Computation of Bayes factors for common designs (R package version 0.9.12-2). Retrieved from https://cran.r-project.org/web/packages/BayesFactor

Mozer, M. C., Pashler, H., Cepeda, N., Lindsey, R., & Vul, E. (2009). Predicting the optimal spacing of study: A multiscale context model of memory. In Y. Bengio, D. Schuurmans, J. Lafferty, C.K.I. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems* (Vol. 22, pp. 1321–1329). La Jolla, CA: NIPS Foundation.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*, 402–407. https://doi.org/10.3758/BF03195588

Raaijmakers, J. G. (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

Rastle, K., Harrington, J., & Coltheart, M. (2002). 358,534 nonwords: The ARC Nonword Database. *Quarterly Journal of Experimental Psychology*, *55A*, 1339–1362. https://doi.org/10.1080/02724980244000099

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

Rouder, J. N., Morey, R. D., Verhagen, J., Swagman, A. R., & Wagenmakers, E.-J. (2017). Bayesian analysis of factorial designs. *Psychological Methods*, *22*, 304–321. https://doi.org/10.1037/met0000057

Russo, R., Parkin, A. J., Taylor, S. R., & Wilks, J. (1998). Revising current two-process accounts of spacing effects in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 161–172. https://doi.org/10.1037/0278-7393.24.1.161

Shaughnessy, J. J., Zimmerman, J., & Underwood, B. J. (1972). Further evidence on the MP–DP effect in free-recall learning. *Journal of Verbal Learning and Verbal Behavior*, *11*, 1–12. https://doi.org/10.1016/S0022-5371(72)80053-7

Storm, B. C., Bjork, R. A., & Storm, J. C. (2010). Optimizing retrieval as a learning event: When and why expanding retrieval practice enhances long-term retention. *Memory & Cognition*, *38*, 244–253. https://doi.org/10.3758/MC.38.2.244

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.

Toppino, T. C., & Gerbier, E. (2014). About practice: Repetition, spacing, and abstraction. In B. H. Ross (Ed.), *The psychology of learning and motivation* (Vol. 60, pp. 113–189). Burlington: Academic Press. https://doi.org/10.1016/B978-0-12-800090-8.00004-4

Tsai, L. (1927). The relation of retention to the distribution of relearning. *Journal of Experimental Psychology*, *10*, 30–39. https://doi.org/10.1037/h0071614

Van Overschelde, J. P., Rawson, K. A., & Dunlosky, J. (2004). Category norms: An updated and expanded version of the Battig and Montague (1969) norms. *Journal of Memory and Language*, *50*, 289–335. https://doi.org/10.1016/j.jml.2003.10.003

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review*, *14*, 779-804. https://doi.org/10.3758/BF03194105

Wilson, M. (1988). MRC Psycholinguistic Database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers*, *20*, 6–11. https://doi.org/10.3758/BF03202594