Title
Additional acquisition sessions monotonically benefit retention and relearning

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 43(43)

ISSN
1069-7977

Authors
Fiechter, Joshua
Sense, Florian
Collins, Michael Gordon
et al.

Publication Date
2021

Peer reviewed

Additional acquisition sessions monotonically benefit retention and relearning

Joshua L. Fiechter
josh.fiechter@gmail.com
Ball Aerospace & Technologies, Wright-Patterson Air Force Base, Dayton, OH, USA

Florian Sense
floriansense@gmail.com
InfiniteTactics, LLC, Beavercreek, OH, USA
Department of Experimental Psychology & Behavioral and Cognitive Neuroscience
University of Groningen, Groningen, The Netherlands

Michael Collins
collins.283@wright.edu
ORISE at Air Force Research Laboratory, Wright-Patterson Air Force Base, Dayton, OH, USA

Michael Krusmark
michael.krusmark.ctr@us.af.mil
L3Harris Technologies, Wright-Patterson Air Force Base, Dayton, OH, USA

Tiffany S. Jastrzembski
tiffany.jastrzembski@us.af.mil
Air Force Research Laboratory, Wright-Patterson Air Force Base, Dayton, OH, USA

Abstract

Spacing is a highly effective encoding strategy that has been shown to benefit memory in a variety of domains. Recent work has emphasized the evaluation of spaced practice under conditions that more closely reflect daily life. This work found that spacing repetitions over days, weeks, or months is effective over retention intervals as long as one year. One aspect of spaced study that has received less attention, however, is the relationship between the number of acquisition sessions and final retention. That is, if a student preparing for an exam plans to allocate 10 hours to preparing for that exam, is there an optimal, or perhaps minimal, number of study sessions that they should engage in to best leverage the benefits of spacing? In the present experiment we had participants complete 16 practice tests of Japanese-English pairs (e.g., *boushi* – hat). These practice tests were either all completed in one session, or distributed across two, three, or four sessions. These sessions were spaced either 1, 7, 30, or 90 days apart. Participants completed four test trials following a retention interval of 90 days, 180 days, or some variable length. Our results suggest that the number of acquisition sessions monotonically enhanced first-trial test performance as well as relearning, though evidence for enhanced relearning between one and two sessions was ambiguous. Unexpectedly, these monotonic trends were stable across practice lags and retention intervals. These findings suggest that, in addition to the temporal lag between practice episodes, the number of sessions over which one elects to distribute those episodes also has ramifications for long-term retention, and that each additional session yields meaningful benefits.

Keywords: spacing effects; relearning; practice schedules; learning; memory

Introduction

Spacing one's practice of information is a powerful technique for enhancing retention. This *spacing effect* was first discovered in the late nineteenth century (Ebbinghaus, 1885) and has since been replicated with various stimuli, groups of learners, and learning environments (see Carpenter, Cepeda, Rohrer, Kang, & Pashler, 2012, for a review). For example, spacing has been shown to benefit simple stimuli such as single words (e.g. Madigan, 1969) and word pairs (e.g., Glenberg, 1976); and has been shown to benefit more complex stimuli such as lectures (Glover & Corkill, 1987), math problems (e.g., Rohrer & Taylor, 2006), and second-language vocabulary (e,g,, Bahrick, 1979). Spacing has also been shown to be effective for learners of all ages, ranging from infants (Cornell, 1980) to older adults (Balota, Duchek, & Paulin, 1989).

Ebbinghaus (1885) provided the first empirical demonstration of the spacing effect. He noted that distributing practice of nonsense syllables and poetry stanzas over a period of days led to faster relearning of those materials than did a greater number of exposures over a single session (this effect of spacing on relearning was recently emphasized by Walsh et al., 2018, and is also addressed in the present study). Subsequent work quickly established that spacing was effective for a variety of materials and for learners of various ages (see Ruch, 1928, for a review of these early studies).

Two major empirical findings emerged in the latter half of the twentieth century. First, Melton (1967) found that increasing lags (time between repetitions within a session) monotonically enhanced retention (i.e., the *lag effect*). Second, Glenberg (1976) qualified the lag effect by noting a *lag-by-retention-interval interaction*. That is, lags are monotonically beneficial only to a degree and will impede retention if they are disproportionately longer than an upcoming retention interval.

Soon after these seminal discoveries, researchers began evaluating lags and retention intervals at larger scales with an eye towards evaluating the ecological validity of spacing effects (e.g., Bahrick, 1979). More recent work has been focused on evaluating the benefits of spacing at timescales that more directly correspond to those found in educational practice, whereby the effects of spacing, practice lags, and retention intervals have been replicated at intervals spanning days and weeks (e.g., Gerbier, Toppino, & Koenig, 2015) to months and years (e.g., Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008).

The spacing-effect literature has established the significance of lag length and retention interval (among other phenomena) for distributed practice, but one issue that has received little attention is the number of sessions[1] over which a fixed amount of practice should be distributed. We addressed this point in the present experiment by manipulating the distribution of practice such that it occurred over one, two, three, or four sessions. Our participants were given 16 practice opportunities, with each session consisting of multiple, spaced practice episodes. We were interested in evaluating the benefits of the number of acquisition sessions over and above the benefits already afforded by within-session spacing.

In considering the possible outcomes of manipulating number of practice sessions, we appealed to accounts of spacing effects that have emphasized a role of forgetting and subsequent relearning (e.g., Benjamin & Tullis, 2010; Bjork & Bjork, 1992; Pavlik & Anderson, 2005). These accounts posit that the benefits of spacing are derived primarily from remembering past stimuli following forgetting of the stimuli over the spacing lag, and more difficult relearning corresponds to better long-term retention. Lags can become costly, however, if memory for a stimulus is attenuated to the degree that learners will not recall it upon a subsequent exposure (Benjamin & Tullis, 2010). These dual effects of forgetting (i.e., potentiating relearning while also risking unsuccessful recall) allows these theories to straightforwardly account for fundamental phenomena in the spacing literature (namely, the lag-by-retention-interval interaction and superadditivity [Begg & Green, 1988]).

The notion that too little or too much forgetting between practice episodes may hinder memory raises the possibility that (a) there may be some minimum number of sessions that a learner should complete, or (b) that the benefits of more spacing might quickly asymptote or even be nonmonotonic. First, regarding a possible minimum number of sessions, it might be the case that learners do not experience enough forgetting between sessions. This is of particular concern when completing more practice trials in fewer sessions. Thus, relearning on an ensuing session will be too slight to benefit long-term retention.

Second, regarding limitations on the benefits of more sessions, it might be the case that distributing practice over an excessively large number of sessions results in too few exposures during any single session, leading to poor encoding of information to an extent that makes it difficult to overcome forgetting between sessions (that is, forgetting may become costly). In this scenario, each new acquisition session is akin to learning information for the first time.

Finally, it could also be the case that more sessions are monotonically better if the amount of inter-session forgetting grows increasingly optimal as the number of within-session repetitions decreases. This scenario seems plausible in light of findings suggesting that spaced practice tests are beneficial by virtue of spacing alone, and not retrieval success on those tests (Pashler, Zarow, & Triplett, 2003; see also Kornell, Klein, & Rawson, 2015). Such a scenario places primary importance on the spacing between first repetitions of each session. In sum, the relationship between the number of acquisition sessions and final retention has no obvious a priori relationship. Here, we seek to clarify the nature of this relationship.

We also evaluated whether presentations distributed over four sessions might be more beneficial if the number of those presentations is gradually decreased. This question was motivated by past studies demonstrating that a gradual expansion of time between practice tests is beneficial for retention (e.g., Landauer & Bjork, 1978). We analogously scaffolded the difficulty of practice tests in this experiment by manipulating the number of within-session practice tests, rather than the time between sessions. If more practice tests early on in acquisition render items more robust to inter-session forgetting, then this decreasing condition should be more beneficial than the uniform four-session practice schedule.

Method

Participants

Participants were 241 nursing students who performed this task as part of a study on effective cardiopulmonary resuscitation (CPR) skills refresher training (see Oermann, Krusmark, Kardong-Edgren, Jastrzembski, & Gluck, in press).

[1] Note that we use the term "session" for convenience, and acknowledge that it is the larger spacing lag between sessions that drives the effects of sessions.

Table 1: Number of participants in each practice lag by retention interval condition (lags and intervals are in days).

	Retention interval			
Lag	90	180	Variable	Total
1	17	18	29	64
7	21	20	27	68
30	21	18	22	61
90	18	9	21	48
Total	77	65	99	**241**

Stimuli

Our stimuli included 104 Japanese-English word pairs (e.g., *boushi* – hat) taken from previously published spacing effect research (Pavlik & Anderson, 2005). A random subset of 27 pairs (3 pairs per schedule) was sampled for each participant and randomly assigned to each condition. The English targets were selected from the MRC Psycholinguistic database to have similar familiarity and imageability ratings.

Design

We used a 5 (within-subject manipulation of practice schedule: one to four sessions and a decreasing condition in which we tapered the number of within-session practice trials across 4 sessions.) by 4 (between-subject random assignment to training lag between sessions: 1, 7, 30, or 90 days) by 3 (between-subject random assignment to retention interval: 90 days, 180 days, or variable [range of 7 to 180 days]) mixed design (see Table 1 and 2 for details).

Table 2: Number of trials per item in each practice condition and experiment phase. (See Table 1 for the distribution of participants across possible lags between acquisition sessions, and possible retention intervals between "Acq. 4" and "Test.")

	Condition				
Phase	Decrease	4 ses.	3 ses.	2 ses.	1 ses.
Acq. 1	6	4	--	--	--
Acq. 2	5	4	6	--	--
Acq. 3	3	4	5	8	--
Acq. 4	2	4	5	8	16
Test	4	4	4	4	4

Procedure

Participants completed four acquisition sessions. The first session consisted of items from the four-session condition, the second consisted of items from the four- and three-session conditions, and so on (see Table 2). Because the number of critical items varied between sessions, filler items were included to ensure that the same number of items (i.e., 18) was practiced on each day. Furthermore, three filler items were practiced for three trials each at the beginning of every session to offset primacy effects. Three items each were randomly assigned to each practice condition. Over the course of the entire study, each item received an initial study trial followed by a total number of 16 practice trials. Table 2 presents the distribution of these trials over sessions for each of our practice schedules. Items were presented according to a scheduling algorithm that ensured an equivalent lag length between within-session repetitions of each item.

Participants saw the complete pair during the initial study trial and only the Japanese cue during subsequent practice trials. These trials proceeded identically except for the different prompts (see Figure 1). The Japanese cue (or pair) was presented for up to five seconds and participants were to provide the English target. The cue (or pair) was removed after five seconds, at which time participants had two additional seconds to provide a response if they had not already done so. Seven seconds after stimulus onset, or upon submission of their response, participants were shown the complete pair for two seconds and explicitly told that their response was either correct or incorrect. A half-second interstimulus interval separated the trials.

Participants were randomly assigned to an inter-session lag of either 1, 7, 30, or 90 days. After completing the final acquisition session, participants completed a test following a retention interval of either 90 or 180 days, or some varying interval (range of 7 to 180 days). This varying interval was determined by a learner-adaptive schedule that prescribed personalized retention intervals based on performance from a different task (a CPR-training task; see Oermann et al., in press) that participants completed in tandem with the word-learning task. As such, personalized learning intervals were prescribed independent of verbal-learning performance on this paired-associate learning task. Participants completed four test trials of each critical item.

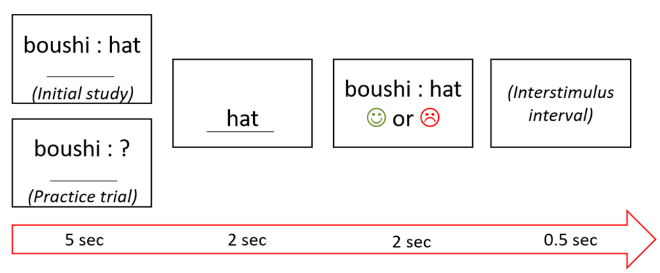

Figure 1: Time course of initial-study and practice trials.

Results

We relied on Bayesian mixed-effects models for all reported analyses. Additionally, all our models used a robit linking function for binary data (Liu, 2004), which has been shown to be more robust to influential observations than logit or probit links. Model estimation was carried out in Stan (Stan Development Team, 2020) via the "brms" package (Bürkner, 2018) in R statistical software. We placed a Cauchy prior (location = 0, scale = 1) on all population-level coefficients; we used default priors for all other parameters for all reported analyses. Additionally, each model had practice schedule (in the case of decreased versus uniform presentation) or number of sessions as the lone population-level predictor, and we included participants, stimuli, practice lag, and retention interval as grouping variables[2] (the latter two were entered as a single, crossed variable so that effects were estimated for each cell of our design; this policy allowed us to test for a lag-by-retention interval interaction) with group-level effects estimated for each. We will report estimated effects of practice lag and retention interval for all analyses included here.

We relied on Bayesian hypothesis testing for evaluating our population-level regression coefficients. Specifically, we calculated Bayes factors via Savage-Dickey ratios (Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010), which are the ratios of zero-point-densities of the prior and posterior distributions, and we report Bayes factors in terms of support for the alternative hypothesis (BF_{10}). Following recommendations by Jeffreys (1961), we considered evidence to be convincing if either the null or alternative hypothesis was supported by at least a factor of 3 ($BF_{10} \leq 0.33$ or $BF_{10} \geq 3$, respectively).

First test trial

Decreasing versus uniform trials First-trial test performance is plotted in Figure 2. We first tested for differences between the decreasing and uniform four-session schedule. Our mixed-effects robit-regression model yielded evidence against a difference between the two conditions ($\beta = 0.04$, $SD = 0.10$; $BF_{10} = 0.16$), suggesting that a decreasing schedule did not enhance initial test performance relative to a uniform schedule. Furthermore, all group-level estimates for crossed levels of practice lag and retention interval yielded a 95% credible interval[3] that included zero, and the estimated standard deviation for these trends supported the null ($\beta_{SD} = 0.16$, $SD = 0.11$; $BF_{10} = 0.09$), suggesting that this null effect was consistent across those variables.

Number of sessions To evaluate whether performance monotonically improved as a function of number of sessions, we analyzed the data using a Bayesian mixed-effects monotonic regression (Bürkner & Charpentier, 2020). Monotonic regression, as implemented here, yields normalized, simplex estimates (that is, all estimates sum to one) of differences between levels of an ordinal variable, as well as a scale parameter indicating the size and direction of the monotonic trend. Because differences between levels are estimated individually, monotonic regression can more accurately estimate nonlinear monotonic trends relative to linear regression.

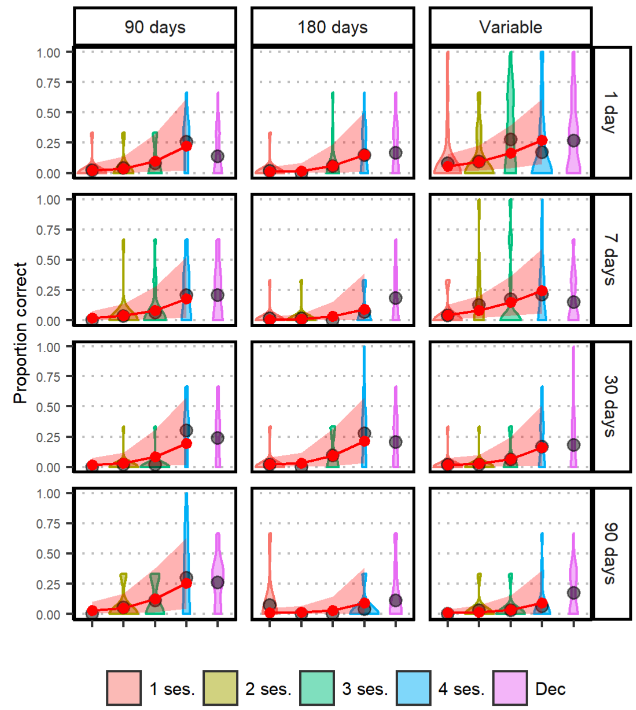

Figure 2: Average first-trial test accuracy as a function of practice lag (rows) and retention interval (columns). Contours of the colored regions indicate the distribution of participants' average accuracy. Black dots indicate observed overall means. Red dots indicate averaged fits from our monotonic robit-regression model (the decreasing condition was not included in that model). The red shaded regions indicate averaged 95% credible intervals.

Averaged model fits are indicated by red points in Figure 2. We obtained strong evidence for an overall positive monotonic trend ($\beta = 0.43$, $SD = 0.06$; $BF_{10} = 3.28 \times 10^{16}$). The simplex estimate for the difference between one and two sessions was 0.23 ($SD = 0.10$; $BF_{10} = 3.40$), between two and three sessions was 0.40 ($SD = 0.10$; $BF_{10} = 1.40 \times 10^{13}$) and between three and four sessions was 0.37 ($SD = 0.07$; $BF_{10} =$

[2] The appropriate treatment of variables as population-level predictors (usually with corresponding group-level effects) or grouping variables in hierarchical models is much-discussed with little consensus (see Gelman, 2005). Our choice to treat our between-subject manipulations as grouping variables, rather than population-level predictors, arose from two considerations: 1) constraints of our study, owing to a primary task (i.e., CPR training) that required a unique population of participants, resulted in limited sample sizes for each cell of our experimental design; in such cases the estimated effects of variables are usually more generalizable when treating them as grouping variables (Gelman, 2005), and 2) a model comparison (conducted via leave-one-out cross-validation; Vehtari, Gelman, & Gabry, 2017) for first-trial test performance showed a very strong preference for a model with lag and retention interval treated as grouping variables rather than population-level predictors. Critically, we did *not* elect to treat these variables as grouping variables because we were not interested in them; we accordingly still report estimates for these variables throughout the manuscript.

[3] We placed a prior on the standard deviation of group-level trends and not the magnitude of those trends. We were therefore unable to calculate a Bayes factor for group-level estimates.

8.09×10^{16}). Thus, we found convincing evidence for differences between each successive number of sessions. All group-level estimates for crossed levels of practice lag and retention interval yielded 95% credible intervals outside of zero, and the estimated standard deviation for these trends supported the null ($\beta_{SD} = 0.03$, $SD = 0.02$; $BF_{10} = 0.01$). These results suggest that the monotonic trend of number of sessions was roughly uniform across those conditions.

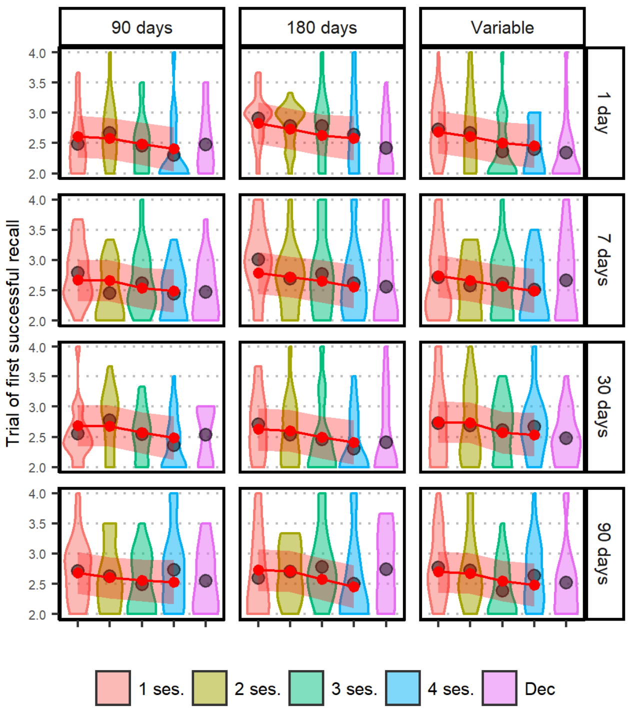

Figure 3: Average trial of first successful recall as a function of schedule, practice lag (rows), and retention interval (columns). See Figure 2 for an explanation of plot elements.

Relearning over all test trials

Decreasing versus uniform trials We next analyzed relearning at test by looking only at those items that were recalled incorrectly on the first test trial (see Walsh et al., 2018). Eighty-nine percent of items were initially recalled incorrectly. Overall test performance for these initial errors was 49%, 71%, and 81% in the respective final three trials. We operationalized relearning as the trial of first successful recall. Some items were never recalled and so we also evaluated the probability of no successful retrieval. Our first analysis evaluated whether these two outcomes varied between the decreasing and uniform, four-session condition. We fit a multivariate mixed-effects model in which we simultaneously estimated effects for both of our response variables (we assumed a *t* distribution for trial of first recall to ensure robust estimates). We found evidence suggesting that trial of first recall was approximately the same in both conditions ($\beta = 0.01$, $SD = 0.05$; $BF_{10} = 0.11$), as was the probability of never recalling an item ($\beta = 0.06$, $SD = 0.14$; $BF_{10} = 0.07$). Thus, decreasing the number of practice trials did not affect relearning relative to a uniform number of practice trials. Furthermore, all group-level estimates for crossed levels of practice lag and retention interval yielded 95% CIs that included zero, and the estimated standard deviation for these trends supported the null for both trial of first recall ($\beta_{SD} = 0.05$, $SD = 0.04$; $BF_{10} = 0.05$) and probability of never recalling ($\beta_{SD} = 0.11$, $SD = 0.09$; $BF_{10} = 0.09$), suggesting that this null effect was consistent across those variables.

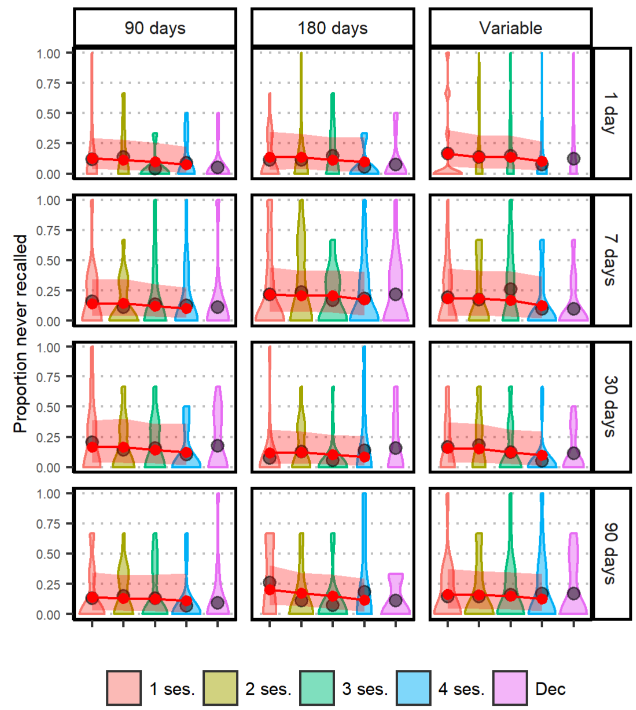

Figure 4: Proportion of items never recalled at test as a function of schedule, practice lag (rows) and retention interval (columns). See Figure 2 for an explanation of plot elements.

Number of sessions Finally, we evaluated whether relearning was enhanced as a function of the number of acquisition sessions. To do so, we fit a model much like the one reported in the preceding section but estimated the effect of the number of sessions via monotonic regression. Averaged fits from this model are presented in Figures 3 and 4. We found strong evidence for an overall negative monotonic trend on trial of first recall ($\beta = -0.08$, $SD = 0.01$; $BF_{10} = 3.26 \times 10^{26}$). The simplex estimate for the difference between one and two sessions was 0.23 ($SD = 0.12$; $BF_{10} = 1.38$), between two and three sessions was 0.45 ($SD = 0.16$; $BF_{10} = 23.09$) and between three and four sessions was 0.33 ($SD = 0.14$; $BF_{10} = 5.04$). Thus, all differences supported the alternative, but only the latter two yielded convincing evidence. All group-level trends for crossed levels of practice lag and retention interval yielded 95% credible intervals outside of zero, and the estimated standard deviation of these trends supported the null ($\beta_{SD} = 0.01$, $SD = 0.01$; $BF_{10} =$

0.005), suggesting that this monotonic effect of acquisition sessions was stable across those conditions.

We also found evidence for an overall negative monotonic trend on proportion of items never recalled (β = -0.11, *SD* = 0.04; BF_{10} = 8.20). The simplex estimate for the difference between one and two sessions was 0.20 (*SD* = 0.13; BF_{10} = 0.91), between two and three sessions was 0.22 (*SD* = 0.17; BF_{10} = 0.69) and between three and four sessions was 0.57 (*SD* = 0.19; BF_{10} = 38.98). Thus, only the difference between three and four sessions yielded convincing evidence. Most group-level trends for crossed levels of practice lag and retention interval yielded 95% credible intervals outside of zero, except for the 90-by-90 and 7-by-180 conditions. That said, the estimated standard deviation of these trends supported the null (β_{SD} = 0.03, *SD* = 0.03; BF_{10} = 0.01), suggesting that this monotonic effect of acquisition sessions was relatively stable across those conditions, with the caveat that two conditions did not yield a reliable trend.

General Discussion

Our findings suggest that more acquisition sessions monotonically improve several aspects of retention, and that these effects are generally applicable to long, ecologically relevant timescales. First, initial levels of retention were higher with more sessions. Second, items that were initially retrieved incorrectly at test were relearned faster if they were acquired across more sessions. Finally, among those same items, the probability of never recalling an item over four trials monotonically decreased with more sessions. Recall that, in our design, more sessions came at the expense of fewer spaced practice opportunities within each individual session (see Table 2), raising the possibility that more sessions could be costly. Clearly, however, the benefit of more sessions consistently outstripped the benefit of more within-session practice trials.

We also found that decreasing the number of trials over four sessions relative to a uniform schedule had little impact on retention. However, analyses not presented here suggest that the decreasing schedule also had no meaningful impact on practice accuracy relative to the uniform schedule. To the degree that we would expect a decreasing schedule to yield benefits by facilitating practice success, we cannot make any strong claims about the relative benefits of a decreasing versus uniform schedule.

The monotonic effect of the number of sessions was extremely stable across manipulations of practice lag and retention interval. One possible reason for this invariance is lack of experimental power, as discussed in footnote 1. Another possibility is that participants were able to substantially relearn material during their multi-repetition practice sessions, and relearning has been shown to attenuate the effects of spacing (Rawson & Dunlosky, 2013). Finally, it might also be the case that inter-session spacing entails critical boundary values, and that longer intervals are not any more beneficial once they have reached, for example, 24 hours. More work is clearly needed to more precisely delineate the relationship between the number of practice sessions and other important variables pertaining to distributed practice.

Our findings suggest that practicing in smaller doses over more sessions is preferable to larger doses in fewer sessions (see also Rawson & Dunlosky, 2011). We have demonstrated here the value of distributing practice among sessions, a consequential dimension of distributed practice that has until now received scant attention in the spacing literature.

References

Bahrick, H. P. (1979). Maintenance of knowledge: Questions about memory we forgot to ask. *Journal of Experimental Psychology: General*, *108*, 296–308.

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

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.

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.) *From learning processes to cognitive processes: Essays in Honor of William K. Estes* (Vol. 2, pp. 35–67). Hillsdale, NJ: Erlbaum.

Bürkner, P. C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*, 395–411.

Bürkner, P. C., & Charpentier, E. (2020). Modelling monotonic effects of ordinal predictors in Bayesian regression models. *British Journal of Mathematical and Statistical Psychology*, *73*, 420–451.

Carpenter, S. K., Cepeda, N. J., Rohrer, D., Kang, S. H. K., & Pashler, H. (2012). Using spacing to enhance diverse forms of learning: Review of recent research and implications for instruction. *Educational Psychology Review*, *24*, 369–378.

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: A temporal ridgeline of optimal retention. *Psychological Science*, *19*, 1095–1102.

Cornell, E. H. (1980). Distributed study facilitates infants' delayed recognition memory. *Memory & Cognition*, *8*, 539–542.

Ebbinghaus, H. (1885/1964). *Memory: A Contribution to Experimental Psychology* (H. A. Ruger, C. E. Bussenius, & E. R. Hilgard, Trans.). New York: Dover Publications.

Gelman, A. (2005). Analysis of variance—why it is more important than ever. *The Annals of Statistics*, *33*, 1–53.

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

Glenberg, A. M. (1976). Monotonic and nonmonotonic lag effects in paired-associate and recognition memory paradigms. *Journal of Verbal Learning and Verbal Behavior*, *15*, 1–16.

Glover, J. A., & Corkill, A. J. (1987). Influence of paraphrased repetitions on the spacing effect. *Journal of Educational Psychology*, *79*, 198–199.

Jeffreys, H. (1961). Theory of probability (3rd). Oxford: Oxford University Press.

Kornell, N., Klein, P. J., & Rawson, K. A. (2015). Retrieval attempts enhance learning, but retrieval success (versus failure) does not matter. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*, 283–294.

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.

Liu, C. (2004). Robit regression: a simple robust alternative to logistic and probit regression. In A. Gelman & X-L. Meng (Eds.) *Applied Bayesian modelling and causal inference from incomplete-data perspectives* (pp. 227–238). New York: Wiley.

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

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

Oermann, M. H., Krusmark, M. A., Kardong-Edgren, S., Jastrzembski, T. S., & Gluck, K. A. (in press). Personalized training schedules for retention and sustainment of CPR skills. *Simulation in Healthcare.*

Pashler, H., Zarow, G., & Triplett, B. (2003). Is temporal spacing of tests helpful even when it inflates error rates? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 1051–1057.

Pavlik, P. I., & Anderson, J. R. (2005). Practice and forgetting effect on vocabulary memory: An activation-based model of the spacing effect. *Cognitive Science*, *29*, 559–586.

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

Rawson, K. A., & Dunlosky, J. (2013). Relearning attenuates the costs and benefits of spacing. *Journal of Experimental Psychology: General*, *142*, 1113-1129.

Rohrer, D., & Taylor, K. (2006). The effects of overlearning and distributed practice on the retention of mathematics knowledge. *Applied Cognitive Psychology*, *20*, 1209–1224.

Ruch, T. C. (1928). Factors influencing the relative economy of massed and distributed practice in learning. *Psychological Review*, *35*, 19–45.

Stan Development Team (2020). Stan Modeling Language Users Guide and Reference Manual, Version 2.26. https://mc-stan.org

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*, 1413–1432.

Wagenmakers, E-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage-Dickey method. *Cognitive Psychology*, *60*, 158–189.

Walsh, M. M., Gluck, K. A., Gunzelmann, G., Jastrzembski, T., Krusmark, M., Myung, J. I., … Zhou, R. Mechanisms underlying the spacing effect in learning: A comparison of three computational models. *Journal of Experimental Psychology: General*, *147*, 1325–1348.