Mem Cogn (2016) 44:897–909
DOI 10.3758/s13421-016-0606-y

Effects of successive relearning on recall: Does relearning override the effects of initial learning criterion?

Kalif E. Vaughn[1] · **John Dunlosky**[2] · **Katherine A. Rawson**[2]

Published online: 30 March 2016

Abstract Retrieval practice improves memory for many kinds of materials, and numerous factors moderate the benefits of retrieval practice, including the amount of successful retrieval practice (referred to as the *learning criterion*). In general, the benefits of retrieval practice are greater with more than with less successful retrieval practice; however, learning items to a higher (vs. lower) criterion requires more time and effort. If students plan on relearning material in a subsequent study session, does the benefit of learning to a higher criterion during an initial session persist? In Session 1, participants studied and successfully recalled Swahili–English word pairs one, two, three, four, five, six, or seven times. In subsequent sessions, all of the pairs were relearned to a criterion of one correct recall at one-week intervals across four or five successive relearning sessions. Experiments 1 and 2 revealed that the substantial benefits of learning to a higher initial criterion during the first session do not persist across relearning sessions. This *relearning-override effect* was also demonstrated in Experiment 2 after a one-month retention interval. The implications of relearning-override effects are important for theory and for education. For theories of test-enhanced learning, they support the predictions of one theory and appear inconsistent with the predictions of another. For education, if relearning is to occur, using extra time to learn to a higher initial learning criterion is not efficient. Instead, students should devote their time to subsequent spaced relearning sessions, which produce substantial gains in recall performance.

Keywords Memory · Successive relearning · Recall

For more than a century, research has demonstrated that testing improves subsequent memory (for reviews, see Carpenter, 2012; Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Karpicke, 2012; Roediger & Butler, 2011), and that the benefits of testing are influenced by numerous factors (Roediger & Karpicke, 2006). Of interest for the present purposes, the benefits of testing (a) increase with the number of successful retrievals during practice (i.e., the number of times an item is tested *and* successfully recalled during initial learning, referred to as the *initial learning criterion*; see Pyc & Rawson, 2009), and (b) are pronounced when retrieval success is achieved across distributed relearning sessions (referred to as *successive relearning*; see, e.g., Bahrick, Bahrick, Bahrick, & Bahrick, 1993). Although the effects of the initial learning criterion and successive relearning are robust and produce substantial gains in memory, only two studies (described below) have examined how the benefits of initial learning criterion and successive relearning interact.

This interaction is the focus of the present research, and to motivate our approach, we provide details about the initial learning criterion and about successive relearning. We then discuss the theoretical and applied implications for exploring the degree to which they interact. First, consider how the initial learning criterion influences subsequent memory performance (e.g., Karpicke & Roediger, 2007; Nelson, Leonesio, Shimamura, Landwehr, & Narens, 1982; Pyc & Rawson, 2009). Pyc and Rawson (2009) had participants engage in retrieval practice with restudy for Swahili–English word pairs (e.g., *mashua–boat*) until the target words had been correctly

✉ Kalif E. Vaughn
kevaughn2@gmail.com

[1] Department of Psychology, Williams College, Williamstown, MA 01267, USA

[2] Kent State University, Kent, OH, USA

recalled anywhere from one to ten times. Performance was assessed on a subsequent cued-recall test (e.g., *mashua*–???) after either a 25-min or a one-week retention interval. Of greatest interest for the present purposes, final test performance increased as a function of higher versus lower initial learning criterion (i.e., performance increased as the number of successful recalls during practice increased, although with diminishing returns as the initial learning criterion increased; see also Vaughn & Rawson, 2011).

Second, consider the benefits of *successive relearning*. A classic example of successive relearning comes from Bahrick, Bahrick, Bahrick, and Bahrick (1993), who had participants learn foreign-language word pairs to criterion across multiple relearning sessions (either 13 or 26 sessions). The relearning sessions were spaced 14, 28, or 56 days apart. Each relearning session began with a cued-recall test (e.g., *casa*–???) and ended when all of the items had been correctly recalled one time (with feedback provided for incorrect responses). Retention was assessed after a 1-, 2-, 3-, 4-, or 5-year delay. Successive relearning yielded impressive levels of final cued-recall performance, considering the lengths of the various retention intervals. For instance, when relearning sessions were spaced 56 days apart, participants successfully recalled approximately 60 % of the foreign-language word pairs after a 5-year delay.

Given the potency of both a higher initial learning criterion and successive relearning, an interesting question emerges: Does the initial learning criterion matter if successive relearning is to follow? Two empirical hypotheses that provide different answers to this question are of particular interest here. The *superadditive hypothesis* is that the benefits of a higher initial learning criterion will increase across relearning sessions (i.e., performance will increasingly favor items learned to a higher vs. a lower initial learning criterion across relearning sessions). By contrast, the *subadditive hypothesis* is that the benefits of a higher versus lower initial learning criterion will be attenuated across relearning sessions. That is, this hypothesis states that subsequent relearning will override—in part or entirely—the benefits of a higher initial learning criterion, which we refer to as a *relearning-override effect*. Evaluating whether the effects of the initial learning criterion and successive relearning are superadditive or subadditive is important for both application and for theory. For application, establishing the nature of the interaction between these two factors allows for stronger prescriptions for students seeking to maximize the efficiency (i.e., the time investment) and durability (i.e., long-term retention) of their learning. Successive relearning is analogous to students revisiting a stack of flashcards throughout the semester to study for a final exam. If students plan to relearn information periodically, does the level of initial learning matter? The learning paradigm used in these experiments provides an empirical examination of this applied question by manipulating the initial learning criterion (i.e., the number of times an item was successfully retrieved during practice), as well as by implementing spaced relearning sessions (which is parallel to students relearning their academic material across several sessions). Of interest, students practicing with flashcards have reported using them on more than one day (see Wissman, Rawson, & Pyc, 2012), suggesting that the results from our study have practical implications in terms of recommendations for students attempting to optimize their practice schedules.

In addition to the applied aspects of the present research, investigating the potential interaction between the initial learning criterion and relearning has theoretical implications. To motivate our theoretical discussion, we have centered our discussion on two recent theoretical accounts: the *retrieval effort hypothesis* and the *two-stage framework*. The retrieval effort hypothesis predicts a superadditive effect, given an assumption about retrieval effort that we evaluated in the present experiments. By contrast, the two-stage framework predicts subadditivity.

According to the retrieval effort hypothesis (REH; Pyc & Rawson, 2009), successful retrieval practice benefits memory most when successful retrieval is more rather than less effortful. The key to the REH's prediction of superadditivity concerns the retrieval effort involved during each relearning session. During the relearning session that follows initial learning, by design, all items will be successfully recalled once. However, a reasonable assumption (and one we tested here) is that the difficulty of successful retrieval will depend on when an item is successfully recalled during the relearning session. Successfully recalling an item on the initial retrieval attempt during relearning (after a one-week delay) would presumably be more effortful than recalling an item on some other retrieval attempt later in the relearning session (e.g., after a delay of only a few minutes after restudying the item). Importantly, items learned to a higher (vs. lower) criterion during initial learning are more likely to be recalled on the initial retrieval attempt (as was established in prior research on criterion learning effects). In contrast, items learned to a lower initial criterion are less likely to be recalled on the initial retrieval attempt. Thus, successful retrieval of lower-criterion items will more likely occur on later trials in the relearning session. As such, the REH predicts that the memory gains during relearning will favor the items learned to a higher (vs. lower) initial criterion, because proportionally more higher-criterion items will be recalled successfully with greater effort (on the initial retrieval attempt after one week) versus with less effort (on a later retrieval attempt after a few minutes).

REH's prediction of superadditivity rests on the subtle but important distinction between the likelihood of successful retrieval versus the effort involved in successful retrieval during relearning. To reiterate, during relearning, all items will be practiced until they are correctly recalled once. Thus, the *likelihood* of successful retrieval during a relearning session is the same for items across all initial learning criteria. The only difference concerns *when* during the relearning session an item is successfully retrieved (on the first attempt vs. during a later attempt). We assume (and empirically confirmed in Exps. 1 and 2) that successful retrieval is more versus less effortful on the first versus on later trials. Given that this

assumption is met, REH's prediction of superadditivity should follow.

In contrast, the *two-stage framework* (Kornell, Klein, & Rawson, 2015) predicts subadditivity. According to the two-stage framework, the retrieval process comprises two stages: (1) the retrieval attempt itself, and (2) the postprocessing of the correct answer. Retrieval enhances learning, provided that both stages occur. If the retrieval attempt is successful, then both stages have occurred, and memory is benefited. If the retrieval attempt is unsuccessful but correct-answer feedback is provided, then both stages have occurred and memory is benefited (of interest, all retrieval failures were followed by immediate feedback in the present experiments). Importantly, the two-stage framework states that all retrieval attempts benefit learning equally, regardless of whether Stage 2 processing of the correct answer is afforded by retrieval success or by feedback (for evidence in support of this claim, see Kornell et al., 2015). Items learned to a higher (vs. a lower) criterion during initial learning are more likely to be recalled on the initial retrieval attempt during the subsequent relearning session (as was established in prior research on criterion learning effects). As a result, learners will necessarily engage in a greater number of retrieval attempts for lower-criterion items to reach the required criterion of one correct recall during relearning. Therefore, the benefit of relearning sessions will be greater for the lower- than for the higher-criterion items. As such, any initial benefit favoring the higher-criterion items will increasingly diminish across subsequent relearning sessions.

However, the extent to which successive relearning might trump the initial learning criterion is largely an open question, because only two prior studies have examined the interactive benefits of criterion learning and successive relearning (Rawson & Dunlosky, 2011, 2013), but these studies were not designed to evaluate these predictions, and hence provide a limited and arguably unfair test of them. For instance, Rawson and Dunlosky (2011, Exp. 1) had participants study and recall eight key-term definitions one, two, three, or four times during initial practice, and then all items were relearned to a criterion of one correct recall after a two-day delay. Participants completed a final cued-recall test approximately six weeks later. Most important, a relearning-override effect was observed: The benefit of learning to a higher initial criterion was present at the outset of the relearning session, but not on the final test (after relearning had taken place). Although this observed relearning-override effect suggests that successive relearning attenuates the benefits of a higher versus lower initial learning criterion, it is important to note that this research involved key-term definitions. Key-term definitions cannot be used for evaluating the theoretical accounts above, for several reasons. First, a key component of the REH is that retrieval effort should be positively related to retrieval gains (i.e., more effort results in greater gains), and thus the predictions of these theories only apply under conditions in which retrieval effort differs. Retrieval effort can be measured via first key-press latencies for correct responses, with longer latencies reflecting more effortful retrieval attempts, which results in greater gains according to the REH. Key-term definitions do not afford latency measures that are readily interpretable, given that the definitions are recalled in idea units (suggesting that one portion of a correct response may be recalled considerably faster than other portions, rendering the assessment invalid as a measure of overall retrieval effort). In contrast, first key-press latencies for paired associates (which were used in the present experiments) afford valid estimates of retrieval effort, given that only one word is recalled, allowing for more controlled evaluations of theoretical accounts involving retrieval effort. Second, given that the recall of key-term definitions includes multiple idea units, many retrieval attempts involve partially correct responses (i.e., some but not all of the target idea units are recalled) prior to attaining criterion. Accordingly, the assigned nominal learning criterion for a given key-term definition typically is not the same as the functional learning criterion achieved for many of the idea units within the definition (e.g., the entire definition for a given term may be correctly recalled only once, but one or more of the idea units within the definition will have been correctly recalled multiple times on previous retrieval attempts). In contrast, paired-associate recall is typically all-or-none (i.e., the response is either correct or incorrect), resulting in more precise manipulations of the initial learning criterion. Third, in the prior research, the recall of key-term definitions could not be machine-scored during learning, and thus the determination of when a recall response was completely correct (and thus decisions about when that item should be dropped from further practice) was based on learners' judgments about the quality of their responses. Participants' judgments are sometimes inaccurate, which adds noise to the actual criterion achieved across items (Dunlosky & Rawson, 2012). In contrast, paired associates are easily machine-scored, and therefore eliminate the need for participants' judgments to score the accuracy of a response. Finally, the key-term definitions used in prior research were adapted from real-world sources (e.g., general psychology textbooks), and participants reported that 14 %–58 % of these key-term definitions had been presented in their general psychology class (across experiments in Rawson & Dunlosky, 2011, 2013). This additional reexposure may have further compromised the manipulation of learning criterion within and across relearning sessions.

In summary, evidence based on key-term definitions poses interpretive difficulties with respect to evaluating theoretical predictions for how successive relearning and the initial learning criterion interact. To minimize these limitations, we used paired-associate materials that support more precise manipulations of learning criterion and degree of relearning. Paired-associate recall also allows for a more valid measurement of retrieval effort, which is critical for evaluating the REH. Moreover, in contrast to the vast literature on testing effects, which includes hundreds of articles (and in which even spinoff effects have received numerous replications and extensions), successive relearning and possible relearning-override effects

(which are relevant both for theory and for education) have been explored and reported in only two prior articles. Thus, the available empirical evidence is minimal, and it remains an open issue whether criterion learning effects and relearning will be subadditive or superadditive.

Finally, it is important to note that the evidence presented within this article will not disprove any particular theory. For instance, if a particular theory cannot account for our evidence as predicted, we are not suggesting that the theory is necessarily incorrect, because any theory could potentially be modified to account for relearning-override effects. In contrast, we present the aforementioned theories in order to motivate our approach, as well as to promote further discussion and research on relearning-override effects and the prevailing theories that could be modified to account for them. We will return to these issues in the General Discussion, wherein we discuss these and other theories and their particular strengths and limitations for explaining the present evidence.

Experiment 1

Method

Participants and design Forty-one Kent State University students participated for course credit. Initial learning criterion (one, two, three, four, five, six, or seven correct recalls during Session 1) and number of relearning sessions (zero, one, two, or three relearning sessions) were within-participants manipulations.

Materials Participants learned seven lists of ten Swahili–English word pairs (for a total of 70 pairs). The lists had similar levels of difficulty, with the proportions of first-trial recall following an initial study phase ranging from .18 to .20 across lists (based on norms reported by Nelson & Dunlosky, 1994). The assignment of each list of ten Swahili–English word pairs to each initial learning criterion was approximately counterbalanced across participants.

Procedure During Session 1, participants learned two blocks of 35 Swahili–English word pairs via initial study, followed by test–restudy practice until each item reached criterion (i.e., until each item had been recalled its preassigned number of times). Each block of 35 word pairs contained five pairs from each learning criterion (one, two, three, four, five, six, and seven), intermixed in random order. During the initial study trials, the Swahili cue and English target were presented on the screen for 10 s. After all 35 word pairs had been studied one time, the test phase began. During test trials, the Swahili cue was presented alone, and participants had up to 8 s to retrieve and type the corresponding English translation (i.e., if they finished typing their response before the 8 s had elapsed, they could press a button to submit it). If the retrieval attempt was successful, the computer added one correct recall to the running criterion count for that particular item. If the retrieval attempt was not successful, the Swahili cue and English target were presented for restudy for 4 s. The word was then placed at the end of the current block to be retested later. Once an item had reached criterion, it was dropped and received no further practice. Once all items had reached criterion for the first block, the study–test procedure was repeated for the second block of 35 words. Session 1 ended when all items had been learned to their preassigned learning criterion or 90 min had elapsed.

Sessions 2, 3, 4, and 5 were relearning sessions. All sessions were spaced one week apart. During each relearning session, participants relearned all items to a criterion of one correct recall using test–restudy practice. Words were tested and relearned in two blocks of 35 word pairs. Importantly, and in contrast to Session 1, each relearning session began with a test phase without the opportunity to first study the word pairs. Without an initial study phase, the first test trial for each item in each relearning session served as an interim retention test (i.e., affording comparisons of one-week retention after no, one, two, or three prior relearning sessions). As in the test trials during Session 1, the Swahili cue was presented for 8 s, and the participant was instructed to type in the corresponding English target. If the response was correct, the item was dropped from further practice until the next relearning session. If the response was incorrect, participants restudied both the Swahili cue and the English target for 4 s, after which the word was placed at the end of the current block to be retested later. Once all items in the first block had been relearned to one correct recall, the procedure was repeated for the second block of 35 words. Each relearning session ended when the participant had recalled each item once or when 30 min had expired.

Results

Regarding the following analyses, two of the participants did not learn all of the items to their preassigned criterion before time had expired. We did not exclude these participants from further analysis, however, given that they reached 95 % of the assigned learning criterion in Session 1. Nine other participants reached criterion but were subsequently absent for one or more relearning sessions; however, we did not exclude their data for the sessions they completed.[1]

[1] The results of Experiment 1 were nearly identical even when we excluded the aforementioned 11 participants. Of interest, all main effects and interactions persisted concerning recall performance (both across initial learning criteria and relearning sessions). Furthermore, the pattern of first key-press latencies was consistent (i.e., the latencies were always longer for items recalled correctly on the first trial versus some other trial, across all criterion levels and across all relearning sessions). To enhance statistical power, we decided to leave these 11 participants in the analyses for Experiment 1.

To revisit, the predictions of the REH (i.e., superadditive effects of initial criterion and relearning) only hold if retrieval effort differs for items successfully recalled on the first trial during a relearning session versus items successfully recalled on a later trial during relearning. To confirm that this condition was met, we first report first key-press latencies as an indicator of retrieval effort. First key-press latencies pertain to the time that elapsed between the presentation of the cue word on a test trial and when the first key was pressed during the typing of a correct response. First key-press latencies reflect in part the amount of time that participants have spent attempting to retrieve the answer, with the plausible assumption being that the more difficult and effortful a particular item is to retrieve, the more time participants will need in order to retrieve (so as to begin typing) the correct response. First key-press latencies (and other closely related reaction time measures) have been used previously as a proxy for retrieval effort (including in the original article proposing the REH; see, e.g., Buckner, Koutstaal, Schacter, Wagner, & Rosen, 1998; Pyc & Rawson, 2009; van den Broek, Segers, Takashima, & Verhoeven, 2014; Wixted & Rohrer, 1993). Although retrieval effort may be influenced by a variety of factors (e.g., differences in practice lag, retention interval, or the type of prior practice), the predictions stemming from the REH are the same regardless of *why* retrieval effort varied. First key-press latencies simply measure these differences in retrieval effort, which is necessary to investigate the predictions from the REH. Figure 1 highlights that mean first key-press latencies were longer when items were correctly recalled on the first trial rather than on a later trial during relearning sessions (the outcomes in Fig. 1 are collapsed across initial learning criteria; as we report in Appendix Table 2, this pattern also held for items at each initial learning criterion). One concern might be that first key-press latencies were influenced by practice effects more than by retrieval effort (i.e., the response latencies changed as participants became more acclimated to the experimental tasks). Fortunately, we could examine this possibility, given that the experimental protocol was repeated across two blocks including 35 items each (for details, see the Method section). If practice effects influenced key-press latencies, latencies would be faster in the second than in the first block, and the same patterns of effects might not be evident in both blocks. Thus, we analyzed the first key-press latencies as a function of position in the relearning session (i.e., in either the first block of 35 items or the second block of 35 items). For each session, linear regression analyses were conducted with trial type (i.e., first vs. later trial within a block) and practice block (i.e., first vs. second) entered as predictors of the first key-press latencies on correct trials. Across all relearning sessions, trial type significantly predicted first key-press latencies (all ps < .024). In contrast, practice block was never significant (all ps > .713). To summarize, first key-press latencies were longer when items were correctly recalled on the first trial versus a later trial within a block, regardless of the block in which an item was relearned. In contrast, practice effects had a negligible influence on first key-press latencies, with no significant differences in latencies as a function of practice block.

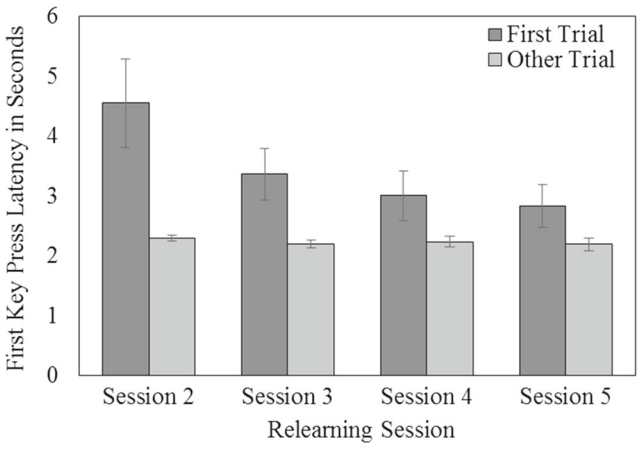

Fig. 1 Mean first key-press latencies, in seconds, for items correctly recalled either on the first trial of a relearning session or on a later trial (after restudy) during the relearning sessions in Experiment 1. Error bars report standard errors of the means

Concerning the other necessary condition for the predictions of the REH to hold, we obtained the expected initial criterion learning effects (i.e., the proportion of items successfully recalled on the first trial of Session 2 was greater for higher- than for lower-criterion items, as we discuss below). Thus, higher-criterion items on average were more difficult to retrieve successfully during relearning, and hence the REH predicts a superadditive effect of learning criterion and relearning under these conditions. In contrast, the two-stage framework predicts subadditive effects between initial learning criterion and relearning under these conditions, because the lower-criterion items would necessarily receive more concluded retrieval attempts during relearning.

Cued-recall performance on the first test trial of each relearning session is reported in Fig. 2. For ease of exposition, inferential statistics are reported in Table 1. A 7 (Initial Learning Criterion) × 4 (Number of Prior Relearning Sessions) repeated measures mixed-factor analysis of variance (ANOVA) revealed main effects of initial learning criterion and relearning session. Most importantly, the interaction was also significant, and inspection of the pattern of outcomes shown in Fig. 2 makes clear that this interaction reflected subadditive effects of initial learning criterion and relearning (the achieved power to detect an interaction was .98). Prior to relearning, initial learning criterion had pronounced effects (a prerequisite for testing the predictions of the two-stage

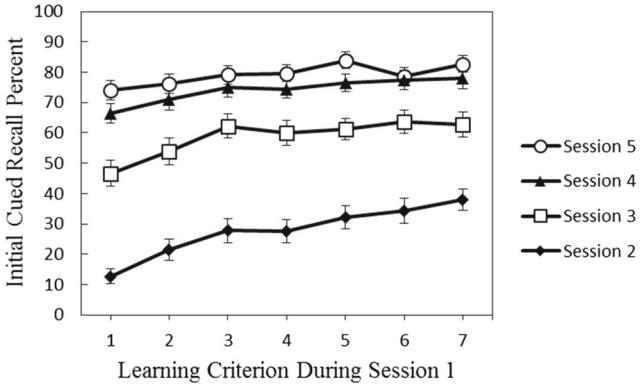

Fig. 2 Mean cued-recall performance on the first test trial at the start of each relearning session in Experiment 1. Error bars report standard errors of the means

framework; η_p^2 = .34 for recall at the outset of Session 2). However, the benefit of a higher initial learning criterion was substantially attenuated after relearning (η_p^2s = .17, .11, and .11 for recall in Sessions 3–5). This relearning-override effect is inconsistent with the predictions of the REH, whereas it is consistent with the predictions of the two-stage framework.

The number of relearning sessions had a marked effect on retention (η_p^2 = .57 across Sessions 2–5). To understand the differential gains following successive relearning versus initial learning criterion, notice that performance one week after an initial learning criterion of four correct recalls is approximately 28 % (see the outcome for Session 2 in Fig. 2). However, performance one week after four correct recalls that were instead spaced across successive relearning sessions was approximately 74 % (see the outcome for criterion 1 items at the beginning of Session 5 in Fig. 2). These results demonstrate the power of successive relearning across multiple learning sessions.

Experiment 2

Experiment 2 was identical to Experiment 1 in terms of its materials, methods, and procedures, except for one key difference: We added a one-month retention interval following Session 5. The key motivation for extending the final retention interval following Session 5 was to replicate the patterns of Experiment 1 with a longer final retention interval. Numerous scholars have emphasized the importance of replication (e.g., Cumming, 2008; Francis, 2012; Maner, 2014; Pashler & Harris, 2012), particularly for establishing whether a new phenomenon is robust. Thus, Experiment 2 was primarily designed to replicate and extend the primary outcomes of Experiment 1. One interesting possibility is that the criterion learning effects might reemerge after a long delay, suggesting a faster rate of forgetting for the pairs learned to a lower versus a higher initial learning criterion. Some prior research has suggested that the rate of forgetting is independent of the level of initial learning (e.g., Kornell, Bjork, & Garcia, 2011; Slamecka & Katsaiti, 1988; Slamecka & McElree, 1983); however, other researchers have disputed that point (e.g., R. A. Bjork & Bjork, 1992; Loftus, 1985). Recent evidence also suggests that testing can slow forgetting (e.g., Congleton & Rajaram, 2012; Wheeler, Ewers, & Buonanno, 2003). If testing does slow forgetting, then the rates of forgetting might differ, given differences in the initial learning criterion. Thus, it is an open-ended issue whether initial learning

Table 1 Inferential statistics for recall performance on the first trial in Session *N* in Experiments 1 and 2

	Main Effect of Learning Criterion					Main Effect of Session					Interaction		
	df	*MSE*	*F*	*p*	η_p^2	*df*	*MSE*	*F*	*p*	η_p^2	*F*	*p*	η_p^2
Experiment 1													
Sessions 2–5	6,840	1.46	28.48	<.001	.17	3,140	22.81	62.49	<.001	.57	1.97	.009	.04
Session 2	6,240	1.47	20.36	<.001	.34								
Session 3	6,210	1.96	6.99	<.001	.17								
Session 4	6,198	1.35	4.23	<.001	.11								
Session 5	6,192	1.03	3.92	.001	.11								
Experiment 2													
Sessions 2–5	6,504	1.55	15.12	<.001	.15	3,84	11.65	89.11	<.001	.76	1.90	.014	.06
Sessions 2–6	6,630	1.57	16.59	<.001	.14	4,105	12.39	64.68	<.001	.71	1.51	.056	.05
Session 2	6,126	1.48	8.92	<.001	.30								
Session 3	6,126	1.91	7.56	<.001	.27								
Session 4	6,126	1.67	<1.3										
Session 5	6,126	1.15	2.19	.048	.09								
Session 6	6,126	1.62	1.93	.080	.08								

criterion effects would again become manifest after a longer retention interval.

Method

Participants and design Twenty-six Kent State University students participated for course credit. As in Experiment 1, initial learning criterion (one, two, three, four, five, six, or seven correct recalls during Session 1) and number of relearning sessions (four relearning sessions spaced one week apart) were within-participants manipulations. Extending beyond Experiment 1, we added a fifth relearning session (Session 6). Session 6 was the same as the previous relearning sessions, but occurred one month after Session 5. The materials and procedure were otherwise identical to those of Experiment 1.

Results

Regarding the following analyses, five participants did not learn all of the items to their preassigned criterion before time had expired. Of these five participants, three were excluded because they were not close to reaching their assigned learning criteria for Session 1 (these participants only completed an average of 55 % of their assigned learning criteria). The fourth participant was also excluded for noncompliance and missing several relearning sessions. The fifth participant was included in the subsequent analyses because this participant reached an acceptable level of initial performance (88 % of his or her assigned learning criteria) and also returned for all relearning sessions.

As is shown in Fig. 3, first key-press latencies were longer for items successfully recalled on the first trial of each relearning session versus on a subsequent trial (for values as a function of initial learning criterion, see Appendix Table 3). For each session, linear regression analyses were conducted with trial type (i.e., first vs. later trial within a block) and practice block (i.e., first vs. second) entered as predictors of first key-press latencies on correct trials. Across all relearning sessions, trial type significantly predicted first key-press latencies (all $ps < .025$). In contrast, practice block was never a significant predictor (all $ps > .25$). To summarize, first key-press latencies were always longer when items were correctly recalled on the first trial rather than on a later trial, regardless of the block in which the item was relearned. And as in Experiment 1, practice effects had a negligible influence on response latencies, with no significant differences in latencies as a function of practice block.

Most importantly, the primary outcomes for Experiment 2 are reported in Fig. 4 and are consistent with the results of Experiment 1. Replicating the cued-recall outcomes of Experiment 1 (which included only four relearning sessions), a 7 (Initial Learning Criterion) × 4 (Number of Prior Relearning Sessions) mixed-factor repeated measures ANOVA revealed main effects of initial learning criterion, relearning session, and a significant interaction (the achieved power to detect an interaction was .97; see Table 1). When including all five relearning sessions, a 7 (Initial Learning Criterion) × 5 (Number of Prior Relearning Sessions) mixed-factor repeated measures ANOVA revealed main effects of initial learning criterion and relearning session, as well as an interaction that approached significance (the achieved power to detect an interaction was .97; see Table 1). Although the interaction only approached significance when analyzing Sessions 2–6, the effects were clearly not superadditive, and the overall pattern was consistent with Experiment 1: The effects of relearning trumped the benefits of initial learning criterion (i.e., the results

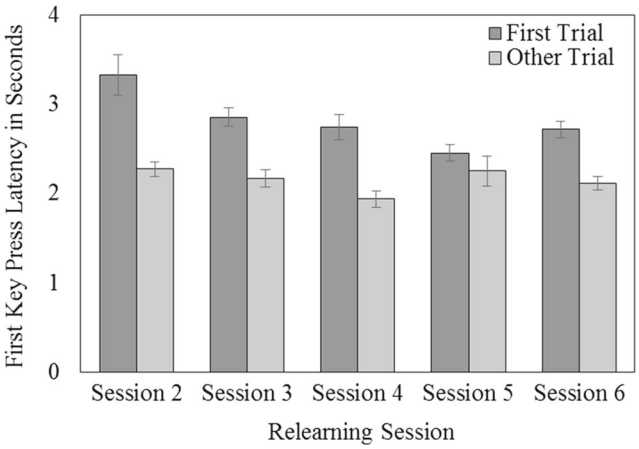

Fig. 3 Mean first key-press latencies, in seconds, for items correctly recalled either on the first trial of a relearning session or on a later trial (after restudy) during the relearning sessions in Experiment 2. Error bars report standard errors of the means

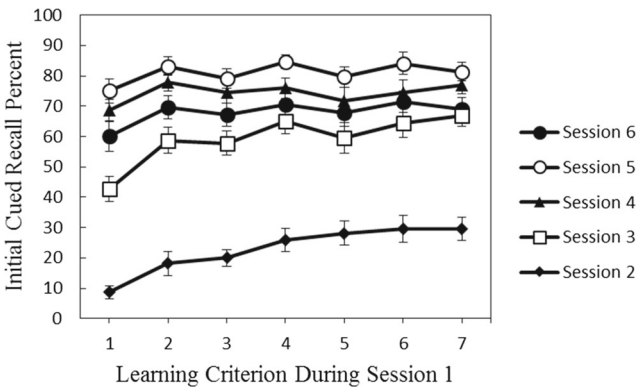

Fig. 4 Mean cued-recall performance on the first test trial at the start of each relearning session in Experiment 2. Error bars report standard errors of the means

supported the subadditive hypothesis). Prior to relearning, the initial learning criterion had pronounced effects (η_p^2 = .30 for recall at the outset of Session 2). However, the benefit of a higher initial learning criterion was substantially attenuated after relearning (η_p^2s = .27, .06, and .09 for recall in Sessions 3–5), again demonstrating a substantial relearning-override effect.

In Experiment 2, we also investigated the extent to which initial learning criterion effects may have reemerged after a one-month retention interval. Visual inspection of Fig. 4 shows that the cued-recall performance following a one-month delay (i.e., at the outset of Session 6) was approximately the same regardless of the initial learning criterion (η_p^2 = .08; the achieved power to detect this effect was .98). Inferential statistics confirmed that the main effect of learning criterion in Session 6 was not significant (although the main effect did approach significance; see Table 1). Finally, the drops in performance from Session 5 to Session 6 were similar, regardless of initial learning criterion, suggesting that the rates of forgetting seemed to be equivalent, regardless of the degree of prior learning (see, e.g., Kornell et al., 2011; Slamecka & Katsaiti, 1988).

The number of relearning sessions had a marked effect on retention (η_p^2 = .71 across Sessions 2–5). To understand the differential gains following successive relearning versus initial learning criterion, notice that performance one week after an initial learning criterion of three correct recalls was approximately 20 % (see the outcome for Session 2 in Fig. 4). However, performance one week after three correct recalls that were instead spaced across successive relearning sessions was approximately 69 % (see the outcome for criterion 1 items at the beginning of Session 4 in Fig. 4). Once again, these results demonstrate the power of successive relearning across multiple learning sessions.

General discussion

In two experiments, we investigated the nature of the interaction between two powerful memory modifiers: initial learning criterion and successive relearning. The results demonstrated substantial relearning-override effects: The benefits of learning to a higher versus a lower initial learning criterion were strong prior to relearning, but then were substantially attenuated across subsequent relearning sessions. Therefore, learning to a higher versus a lower initial learning criterion is inefficient if relearning is to follow. In contrast, successive relearning produced large memory gains across all relearning sessions.

Theoretical implications

These results appear inconsistent with the predictions from the REH (Pyc & Rawson, 2009). Both experiments established conditions under which the predictions of the REH would hold: (a) Retrieval effort was greater for items successfully retrieved on the first trial versus on a later trial during relearning, and (b) higher-criterion items (vs. lower-criterion ones) were more likely to be successfully recalled on the effortful first trial (particularly during Session 2). Under these conditions, the REH predicts that relearning will yield greater incremental gains in memory strength for higher- than for lower-criterion items; in contrast, both experiments demonstrated subadditive effects of initial learning criterion and relearning.

In contrast, the pattern of subadditivity is mostly consistent with the predictions of the two-stage framework (Kornell et al., 2015). The two-stage framework states that all retrieval attempts benefit memory (as long as correct-answer feedback is provided following retrieval failures). During relearning, the lower-criterion items were less likely to be recalled on the initial test trial, necessarily affording them a greater number of retrieval attempts than the higher-criterion items. Therefore, the two-stage framework posits that the lower-criterion items would benefit more during relearning, thereby attenuating the benefits of learning to a higher initial criterion with each subsequent relearning session. Our results are consistent with these predictions and provide support for the two-stage framework during relearning. However, note that the two-stage framework cannot completely accommodate the full set of outcomes reported here, given that the higher-criterion items still enjoyed a greater number of total test trials (concluded with either success or feedback) than the lower-criterion items (i.e., when combining across all learning sessions). The two-stage framework does not currently address factors that might affect the quality of a specific test trial (e.g., whether it is spaced or massed), as it was not originally intended to address such issues. Therefore, although the two-stage framework provides a useful way to think about the benefits of retrieval during relearning, applying its global predictions to a successive-relearning paradigm would require further specification of the framework. Likewise, any other general learning theory that assumes the total number of trials to be the determining factor for final performance after a delay would also need further specification to account for the present results (presumably by accounting for factors that influence the degree of learning from a specific retrieval attempt).

How do other theories of testing effects fare with respect to providing potential explanations for these relearning-override effects? According to the *elaborative-retrieval hypothesis* (ERH; see Carpenter, 2009; Carpenter & Delosh, 2006), retrieval practice benefits memory due to the activation of cue-related semantic information during memory search for the target. If the target is successfully retrieved, the activated semantic information is encoded along with the cue–target pair, to provide additional retrieval routes to the target information. For instance, after studying the word pair *eggs–breakfast*, a subsequent test trial (e.g., *eggs*–???) may activate additional related information (e.g., *bacon*, *cereal*, *sausage*). Importantly, after correctly retrieving the target word (e.g., *breakfast*), the additionally activated information would be retained in the cue–target association (e.g., *eggs–bacon–breakfast*). During a subsequent test trial (e.g., *eggs*–???), the additional information also becomes activated (e.g., *bacon*), which helps facilitate recall of the target (e.g., *breakfast*). As currently formulated, ERH does not make specific claims about the potential effects of relearning beyond initially successful retrieval attempts. However, one reasonable extrapolation of this account would predict superadditive effects. In brief, the degree of activation of elaborative information likely depends on how extensive the memory search for the target is. For example, Rawson, Vaughn, and Carpenter (2015) recently reported that elaborative retrieval is enhanced when retrieval occurs after long versus short lags. By comparison, a more extended search of memory would likely be required for items retrieved on the first trial of a relearning session (i.e., after a long lag) than on a later trial (i.e., after a shorter lag). Given these differential levels of semantic activation during relearning (and subsequently different levels of memory gains), ERH would also predict that the effects of initial learning criterion would be enhanced with relearning. Thus, ERH will require further specification, and likely modification, to account for the present outcomes.

In contrast, other theories of testing effects may fare better with respect to providing an explanation for the present outcomes. According to the *mediator shift hypothesis* (MSH; Pyc & Rawson, 2012), failures during retrieval practice promote the encoding of more effective mediators during subsequent restudy opportunities, which in turn increases the likelihood of subsequent retrieval success. One possible explanation for the relearning-override effects observed here follows from this account. On the first trial of a relearning session, retrieval failure is greater for lower- than for higher-criterion items. Thus, the encoding of effective mediators during restudy in relearning sessions would be more likely for lower-criterion items, which in turn would increase the likelihood of retrieval success in the next relearning session.

The *bifurcated-distribution model* (BDM; see Kornell, Bjork, & Garcia, 2011) may also provide an explanation for why the effects of initial learning criterion and successive relearning are subadditive. In brief, BDM is a memory strength model that assumes no gain in memory strength from retrieval failure (assuming that no feedback is provided), an intermediate gain in memory strength from a study trial, and a large gain in memory strength from successful retrieval. As applied to the present paradigm, all items were successfully retrieved once during each relearning session, and thus all items would receive a large gain in memory strength. In addition, however, more lower-criterion (vs. higher-criterion) items would also receive gains in memory strength from the restudy opportunities that followed the initial retrieval failures. These additional increments in memory strength would presumably attenuate the recall gap between lower- and higher-criterion items in subsequent learning sessions.

Thus far, we have discussed our results through the lens of various memory-based theories of testing effects. Although these theories are not mutually exclusive, it is important to note that each theory emphasizes a different set of causal mechanisms. For instance, retrieval effort is the lynchpin of REH but is absent from the two-stage framework, ERH, MSH, and BDM. Similarly, elaboration is critical to ERH and MSH, but is not directly addressed by the two-stage framework, REH, and BDM. Although there may be some overlap between these theories of testing effects, we have chosen to view and discuss each theory independently for two key reasons: (1) This allows for greater ease of exposition and discussion, and (2) we view these theories as being more distinct than similar, since they posit nonidentical sets of causal mechanisms to explain testing effects. To revisit, even though some theories seem unable to account for our present findings of subadditivity, we are not suggesting that these theories are incorrect. Nevertheless, they will need to be modified to account for the relearning-override effects. For instance, in other experimental contexts, retrieval effort may play a pivotal role in the size of subsequent testing effects. However, within the present experimental conditions, the REH could not account for the interaction between initial learning criterion and successive relearning. Of course, successive relearning is a powerful learning strategy, and factors that have previously produced benefits in learning (e.g., retrieval effort, elaboration, test potentiation effects, etc.) may all be overshadowed by the benefits of successive relearning.

We mentioned above that first key-press latencies provide an acceptable proxy for retrieval effort, but are not a process-pure measure of retrieval effort. Given that numerous theories appeal to retrieval effort for their predictions, further research investigating theoretical predictions based on retrieval effort (e.g., the REH, in particular) should explore other measures of retrieval effort (e.g., measuring pupil dilation during the learning phase), which could provide converging evidence.

Finally, a major goal of the present research has been to highlight the importance of investigating learning in contexts that are representative of real-world learning goals; that is, to investigate learning when students use effective strategies (testing with spaced practice) to obtain mastery levels of performance. Thus, although one-session experiments can provide insight into what techniques may boost learning, rarely does a single session of learning lead to levels of performance that could support passable levels of real-world performance. In the present case, even after recalling pairs seven times correctly during an initial study session, participants retained less than 40 % of these items on the subsequent test. One inherent property of investigating mastery learning over multiple sessions, however, is that differences in item difficulty can influence the outcomes. For instance, during the second session, items that are not correctly recalled on the first trial will need to be relearned, and it seems likely that the item difficulty for this subset of pairs would be greater for those that had originally been learned to a higher criterion (than for a lower one) during the initial learning session. For instance, to provide an indicator of item difficulty, we computed the number of trials needed to reach the first correct recall during Session 1 for each item (and for each participant). Appendix Table 5 reports the mean item difficulties (i.e., mean trials to the first correct recall in Session 1) for items that were correctly recalled on the first trial of Session 2 versus items that were correctly recalled on a later trial in Session 2. Overall, the item difficulty was lower for those items that were correctly recalled on the first trial of Session 2 versus those that were recalled on a later trial (i.e., those that needed to be relearned). For the items that needed to be relearned in Session 2, the item difficulty tended to be lower in the lower-criterion than in the higher-criterion conditions (e.g., in Exp. 1, the criterion 1 items that needed to be relearned had only required 3.06 trials in Session 1, whereas the criterion 7 items that needed to be relearned had required 3.89 trials in Session 1). Thus, the greater benefit of relearning for lower-criterion versus higher-criterion items (i.e., the attenuated effect of initial learning criterion observed on the first trial of Session 3) may partly reflect that the relearned items were easier in the lower-criterion than in the higher-criterion conditions. Even so, item difficulty effects are unlikely to completely account for the relearning-override effects observed here, given that the item difficulty differences between relearned items in the lower- versus the higher-criterion conditions were modest at best. Furthermore, item difficulty differed minimally for the medium- versus higher-level criterion conditions, but an attenuated effect of initial learning criterion was still observed on the first trial of Session 3.

Of course, to the degree that item effects do contribute, the present data do not provide the most stringent tests of the prevailing theories, and this should be considered when evaluating the theories' relative merits. As importantly, these observations also highlight another limitation in the field: The theories were typically designed to explain the outcomes from single-session experiments, and hence by design they may be limited in their ability to explain real-word learning in which people's goals are to master content across multiple sessions. These representative contexts could also have been influenced by item effects, and if so, the theories will need to be adapted if they are meant to account for mastery learning in real-world learning contexts.

Practical implications

In addition to their theoretical implications, these results have applications for student learning. Provided that students engage in relearning sessions, practicing to a higher initial criterion is an inefficient strategy. Obtaining a higher initial learning criterion in Session 1 incurred more costs in terms of time spent studying (see Appendix Table 4). Although this additional cost was partially recouped via faster relearning in Session 2, the overall cost was not entirely recouped, and thus resulted in a net loss in terms of efficiency. For example, in Experiment 2, criterion 7 items required approximately 15 trials per item across Sessions 1–5 to achieve 69 % recall at the outset of Session 6, whereas criterion 2 items only required ten trials to achieve 70 % recall at the outset of Session 6. Students seeking to maximize retention should avoid spending time and effort on learning to a high initial criterion, but should instead devote those resources to subsequent relearning sessions. For example, in Experiment 2, learners spent 9.1 trials, on average, to achieve criterion 7 during initial learning, which yielded 29 % recall one week later. By contrast, learners only spent 8.5 trials to achieve criterion 1 during initial learning and then relearn those items three times, which yielded 72 % recall one week later.

In the present set of experiments, successive relearning trumped the benefits of a higher versus lower initial learning criterion. Future research will be needed to see whether successive relearning can trump other factors shown to enhance memory, such as a longer versus shorter practice lag (e.g., Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006). Of interest, successive relearning does not require a great deal of time to implement. In Experiment 1, after the initial learning sessions, participants used only about 7 min per session to recall or relearn all of the items. Educators seeking to enhance student retention should consider administering multiple spaced review sessions wherein information that was previously learned is tested and relearned throughout the academic semester. The present results suggest that as long as students attempt to recall the information and feedback is provided for incorrect attempts (leading to eventual success in additional retrieval attempts), successive relearning throughout the semester could markedly improve performance on a cumulative final exam. This recommendation holds true regardless of the level of prior learning of the information, since in the present experiments successive relearning improved subsequent performance for items learned to either a lower or a higher initial criterion.

Although successive relearning has been shown to enhance memory performance, only a handful of studies beyond the present one have investigated its power (e.g., Bahrick, Bahrick, Bahrick, Bahrick, 1993; Rawson & Dunlosky, 2011, 2013). More research needs to be conducted to investigate it further, but unfortunately, successive relearning is not easy to investigate. Successive relearning involves participants completing numerous relearning sessions spaced days or weeks apart, which represents a significant time commitment for both the researcher and the participants. Moreover, prior research has primarily used key-term definitions for their learning materials, which means that responses had to be hand-scored. This burden was minimized in the present experiment by using foreign-language word pairs. Moreover, participants were able to quickly relearn the items, and their responses were computer-scored. Given that the present method minimizes the resources required to investigate successive relearning, we hope other researches will adapt the present method to further explore relearning-override effects and the power of successive relearning.

Author note The research reported here was supported by a James S. McDonnell Foundation 21st Century Science Initiative in Bridging Brain, Mind and Behavior Collaborative Award. Preparation of the manuscript was supported by a separate funding mechanism from the James S. McDonnell Foundation (awarded to Nate Kornell) to fund the first author.

Appendix

Table 2 First key-press latencies for items correctly recalled on the first trial or on a later trial during a relearning session in Experiment 1

	Session							
	2		3		4		5	
	M	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
Items Correctly Recalled on First Trial in Session *N*								
Criterion 1	4.92	.71	3.23	.45	3.21	.42	2.93	.44
Criterion 2	4.58	.60	3.76	.53	3.10	.41	2.93	.41
Criterion 3	4.23	.59	3.30	.45	2.77	.37	2.82	.41
Criterion 4	4.03	.58	3.55	.41	3.53	.64	3.16	.45
Criterion 5	3.71	.46	3.18	.40	2.89	.45	2.59	.24
Criterion 6	4.07	.46	3.25	.40	2.77	.39	2.76	.38
Criterion 7	4.14	.46	3.16	.43	2.75	.31	2.63	.35
Items Correctly Recalled on Later Trial in Session *N*								
Criterion 1	2.34	.07	2.19	.12	2.24	.14	2.19	.17
Criterion 2	2.42	.08	2.11	.09	2.15	.16	2.04	.14
Criterion 3	2.29	.09	2.22	.14	2.14	.19	2.69	.23
Criterion 4	2.33	.08	2.18	.13	1.80	.11	1.87	.17
Criterion 5	2.23	.08	2.16	.12	2.08	.11	1.92	.14
Criterion 6	2.34	.11	2.06	.12	2.21	.17	2.48	.22
Criterion 7	2.18	.09	2.03	.13	2.06	.16	2.03	.17

Table 3 First key-press latencies for items correctly recalled on the first trial or on a later trial during a relearning session in Experiment 2

	Session									
	2		3		4		5		6	
	M	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
Items Correctly Recalled on First Trial in Session *N*										
Criterion 1	3.22	.29	2.82	.12	3.22	.21	2.50	.13	3.14	.29
Criterion 2	3.28	.27	2.76	.11	2.66	.16	2.55	.20	2.55	.11
Criterion 3	3.73	.49	3.03	.18	2.63	.15	2.52	.19	2.84	.15
Criterion 4	3.42	.36	2.86	.12	2.79	.34	2.35	.12	2.63	.14
Criterion 5	3.46	.38	2.99	.20	2.81	.22	2.49	.17	2.64	.13
Criterion 6	3.07	.22	2.66	.15	2.51	.14	2.44	.10	2.75	.16
Criterion 7	3.75	.37	2.87	.14	2.59	.15	2.32	.18	2.48	.12
Items Correctly Recalled on Later Trial in Session *N*										
Criterion 1	2.42	.13	2.11	.12	1.72	.10	1.84	.14	1.96	.13
Criterion 2	2.30	.11	1.98	.13	2.13	.21	2.13	.20	2.04	.17
Criterion 3	2.23	.10	2.17	.16	2.10	.13	2.18	.19	2.30	.19
Criterion 4	2.15	.12	2.22	.24	1.94	.22	2.54	.35	2.25	.19
Criterion 5	2.33	.10	1.90	.12	1.77	.11	2.17	.26	2.20	.16
Criterion 6	2.32	.11	2.48	.24	2.51	.15	2.18	.26	2.07	.15
Criterion 7	2.14	.11	2.27	.17	2.59	.16	2.38	.20	1.99	.13

Table 4 Mean numbers of trials to criterion per item in Experiments 1 and 2

	Session											
	1		2		3		4		5		6	
	M	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
Experiment 1												
Criterion 1	2.99	.16	2.53	.08	1.73	.07	1.47	.07	1.35	.05		
Criterion 2	4.15	.18	2.27	.09	1.63	.07	1.38	.06	1.32	.05		
Criterion 3	5.16	.16	2.12	.09	1.53	.07	1.35	.06	1.27	.05		
Criterion 4	6.24	.19	2.12	.08	1.53	.07	1.36	.05	1.26	.04		
Criterion 5	7.19	.16	2.12	.10	1.51	.05	1.31	.04	1.20	.04		
Criterion 6	8.21	.18	1.93	.08	1.51	.07	1.29	.06	1.27	.05		
Criterion 7	9.19	.16	1.88	.07	1.49	.07	1.31	.06	1.22	.04		
Experiment 2												
Criterion 1	2.97	.30	2.41	.08	1.70	.07	1.41	.06	1.34	.07	1.49	.07
Criterion 2	3.79	.20	2.23	.10	1.51	.06	1.25	.03	1.21	.04	1.39	.05
Criterion 3	5.18	.25	2.20	.08	1.54	.07	1.28	.05	1.23	.03	1.43	.06
Criterion 4	6.20	.25	2.08	.09	1.43	.07	1.29	.05	1.18	.03	1.38	.05
Criterion 5	7.32	.30	2.11	.09	1.48	.08	1.32	.06	1.25	.05	1.43	.07
Criterion 6	8.40	.28	2.02	.09	1.45	.08	1.32	.05	1.19	.04	1.36	.05
Criterion 7	9.10	.26	2.03	.09	1.38	.05	1.26	.03	1.22	.03	1.36	.05

Table 5 Mean numbers of trials needed to reach the first correct recall in Session 1, for items that were correctly recalled on either the first trial or a later trial during Session 2 (in Exps. 1 and 2)

	Experiment			
	1		2	
	M	*SE*	*M*	*SE*
Items Correctly Recalled on First Trial in Session 2				
Criterion 1	2.25	.24	1.96	.36
Criterion 2	2.34	.31	1.65	.15
Criterion 3	2.54	.20	2.30	.24
Criterion 4	2.54	.22	2.25	.20
Criterion 5	2.31	.16	2.24	.32
Criterion 6	2.35	.19	2.36	.21
Criterion 7	2.55	.19	2.96	.31
Items Correctly Recalled on Later Trial in Session 2				
Criterion 1	3.06	.17	3.03	.30
Criterion 2	3.44	.20	3.12	.24
Criterion 3	3.57	.21	3.66	.32
Criterion 4	3.76	.22	3.71	.32
Criterion 5	3.82	.21	4.08	.39
Criterion 6	3.97	.27	4.28	.45
Criterion 7	3.89	.22	3.67	.33

References

Bahrick, H. P., Bahrick, L. E., Bahrick, A. S., & Bahrick, P. E. (1993). Maintenance of foreign language vocabulary and the spacing effect. *Psychological Science, 4,* 316–321. doi:10.1111/j.1467-9280.1993.tb00571.x

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

Buckner, R. L., Koutstaal, W., Schacter, D. L., Wagner, A. D., & Rosen, B. R. (1998). Functional-anatomic study of episodic retrieval using fMRI: I. Retrieval effort versus retrieval success. *NeuroImage, 7,* 151–162.

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 35,* 1563–1569. doi:10.1037/a0017021

Carpenter, S. K. (2012). Testing enhances the transfer of learning. *Current Directions in Psychological Science, 21,* 279–283.

Carpenter, S. K., & Delosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition, 34,* 268–276. doi:10.3758/BF03193405

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

Congleton, A., & Rajaram, S. (2012). The origin of the interaction between learning method and delay in the testing effect: The roles of

processing and conceptual retrieval organization. *Memory & Cognition, 40,* 528–539. doi:10.3758/s13421-011-0168-y

Cumming, G. (2008). Replication and *p* intervals: *p* values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science, 3,* 286–300. doi:10.1111/j.1745-6924.2008.00079.x

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learning and Instruction, 22,* 271–228.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14,* 4–58. doi:10.1177/1529100612453266

Francis, G. (2012). The psychology of replication and replication in psychology. *Perspectives on Psychological Science, 7,* 585–594. doi:10.1177/1745691612459520

Karpicke, J. D. (2012). Retrieval-based learning: Active retrieval promotes meaningful learning. *Current Directions in Psychological Science, 21,* 157–163.

Karpicke, J. D., & Roediger, H. L., III. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language, 57,* 151–162. doi:10.1016/j.jml.2006.09.004

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

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. doi:10.1037/a0037850

Loftus, G. R. (1985). Evaluating forgetting curves. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 11,* 397–406. doi:10.1037/0278-7393.11.2.397

Maner, J. K. (2014). Let's put our money where our mouth is if authors are to change their ways, reviewers (and editors) must change with them. *Perspectives on Psychological Science, 9,* 343–351.

Nelson, T. O., & Dunlosky, J. (1994). Norms of paired-associate recall during multitrial learning of Swahili–English translation equivalents. *Memory, 2,* 325–335. doi:10.1080/09658219408258951

Nelson, T. O., Leonesio, R. J., Shimamura, A. P., Landwehr, R. F., & Narens, L. (1982). Overlearning and the feeling of knowing. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 8,* 279–288.

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science, 7,* 531–536.

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

Pyc, M. A., & Rawson, K. A. (2012). Why is test–restudy practice beneficial for memory? An evaluation of the mediator shift hypothesis. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 38,* 737–746. doi:10.1037/a0026166

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

Rawson, K. A., & Dunlosky, J. (2013). Relearning attenuates the benefits and costs of spacing. *Journal of Experimental Psychology. General, 142,* 1113–1129. doi:10.1037/a0030498

Rawson, K. A., Vaughn, K. E., & Carpenter, S. K. (2015). Does the benefit of testing depend on lag, and if so, why? Evaluating the elaborative retrieval hypothesis. *Memory & Cognition, 43,* 619–633. doi:10.3758/s13421-014-0477-z

Roediger, H. L., III, & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15,* 20–27. doi:10.1016/j.tics.2010.09.003

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

Slamecka, N. J., & Katsaiti, L. T. (1988). Normal forgetting of verbal lists as a function of prior testing. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 14,* 716–727. doi:10.1037/0278-7393.14.4.716

Slamecka, N. J., & McElree, B. (1983). Normal forgetting of verbal lists as a function of their degree of learning. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 9,* 384–397. doi:10.1037/0278-7393.9.3.384

van den Broek, G. S., Segers, E., Takashima, A., & Verhoeven, L. (2014). Do testing effects change over time? Insights from immediate and delayed retrieval speed. *Memory, 22,* 803–812.

Vaughn, K. E., & Rawson, K. A. (2011). Diagnosing criterion-level effects on memory: What aspects of memory are enhanced by repeated retrieval? *Psychological Science, 22,* 1127–1131. doi:10.1177/0956797611417724

Wheeler, M., Ewers, M., & Buonanno, J. (2003). Different rates of forgetting following study versus test trials. *Memory, 11,* 571–580.

Wissman, K. T., Rawson, K. A., & Pyc, M. A. (2012). How and when do students use flashcards? *Memory, 20,* 568–579. doi:10.1080/09658211.2012.687052

Wixted, J. T., & Rohrer, D. (1993). Proactive interference and the dynamics of free recall. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 19,* 1024–1039. doi:10.1037/0278-7393.19.5.1024