Journal of Experimental Psychology: General
2011, Vol. 140, No. 3, 283–302

0096-3445/11/$12.00 DOI: 10.1037/a0023956

Optimizing Schedules of Retrieval Practice for Durable and Efficient Learning: How Much Is Enough?

Katherine A. Rawson and John Dunlosky
Kent State University

The literature on testing effects is vast but supports surprisingly few prescriptive conclusions for how to schedule practice to achieve both durable and efficient learning. Key limitations are that few studies have examined the effects of initial learning criterion or the effects of relearning, and no prior research has examined the combined effects of these 2 factors. Across 3 experiments, 533 students learned conceptual material via retrieval practice with restudy. Items were practiced until they were correctly recalled from 1 to 4 times during an initial learning session and were then practiced again to 1 correct recall in 1–5 subsequent relearning sessions (across experiments, more than 100,000 short-answer recall responses were collected and hand-scored). Durability was measured by cued recall and rate of relearning 1–4 months after practice, and efficiency was measured by total practice trials across sessions. A consistent qualitative pattern emerged: The effects of initial learning criterion and relearning were subadditive, such that the effects of initial learning criterion were strong prior to relearning but then diminished as relearning increased. Relearning had pronounced effects on long-term retention with a relatively minimal cost in terms of additional practice trials. On the basis of the overall patterns of durability and efficiency, our prescriptive conclusion for students is to practice recalling concepts to an initial criterion of 3 correct recalls and then to relearn them 3 times at widely spaced intervals.

Keywords: testing effects, retrieval practice, learning to criterion, relearning, retention

Supplemental materials: http://dx.doi.org/10.1037/a0023956.supp

In this journal more than 30 years ago, Harry Bahrick stated the following:

> It is disappointing that nearly 100 yr. of research have not yielded much progress toward specification of the conditions under which information, once acquired, can be maintained indefinitely. This lack of progress is particularly disappointing because few questions about memory could have more significance from either a theoretical or an applied point of view. Our educational system is designed to impart knowledge, and the significance of this great effort depends on the degree of permanence of the effects that are achieved. Much of the information acquired in classrooms is lost soon after final examinations are taken, but beyond the general advice to practice and rehearse frequently, we have little to offer those who wish to minimize or prevent such losses. (Bahrick, 1979, p. 297)

Unfortunately, not much progress has been made since Bahrick's (1979) clarion call, despite an increasing amount of research investigating conditions that promote memory. Perhaps one of the more potent and highly investigated mnemonic strategies is *retrieval practice*. To date, more than 300 experiments in more than 150 articles dating back more than 100 years (Abbott, 1909) have shown that practice tests that require recall of target information from memory (i.e., retrieval practice) improve memory. But despite the wealth of research on retrieval practice effects, this literature supports surprisingly few specific prescriptions for how learners should most effectively schedule retrieval practice to achieve long-term retention.

To illustrate the point, imagine a student seeking advice from an instructor on how she should go about using retrieval practice to most effectively learn course material. At best, the instructor might respond with a general recommendation based on the general qualitative conclusions that are supported by the retrieval practice literature—"More practice is better, and it should be spread out over time." Knowing that she has exams coming up in five classes and is expected to know a great deal of material for each class, the student might then ask, "Yes, but how much is enough? How much do I need to practice in each study session? And how many different study sessions do I need?" Unfortunately, the instructor would be unable to provide the student with any further practical advice, because the literature currently provides no answers to these questions. Furthermore, although students are usually (and understandably) focused on enhancing their memory over relatively short intervals of time (e.g., a span of days, enough to do well on an exam), ideally students would also be structuring their retrieval practice in such a way as to achieve long-term retention

This article was published Online First June 27, 2011.

Katherine A. Rawson and John Dunlosky, Department of Psychology, Kent State University.

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grants R305H050038 and R305A080316 to Kent State University. The opinions expressed are those of the authors and do not represent views of the institute or the U.S. Department of Education. For assistance with heroic amounts of data collection, data scoring, and data entry, special thanks go to Mike Appleman, Nicole Gonzalez, Phil Grimaldi, Tonya Hardway, Caitlin Metelko, Daniel Molnar, Rochelle O'Neil, Danielle Roberto, Melissa Roderick, Sara Smith, and Katie Wissman.

Correspondence concerning this article should be addressed to Katherine A. Rawson, Kent State University, Department of Psychology, P.O. Box 5190, Kent, OH 44242-0001. E-mail: krawson1@kent.edu

beyond the more proximal time point of the exam. As Bahrick (1979) noted, a critical goal of education is to equip students with a durable knowledge base, as is the case in many training environments more generally.

Although we agree with Bahrick (1979) regarding the importance of investigating conditions that promote durable learning, achieving durable learning is trivially easy in principle—rehearse the target information indefinitely. Unfortunately, this strategy is not possible in practice, because learners cannot dedicate an unlimited amount of time to retaining any one piece of information. This point highlights the often overlooked companion goal to durable learning, namely, efficient learning. How best to achieve durable learning is a difficult question only because most real-world learning environments require the learning of large amounts of information while simultaneously placing considerable constraints on how much time can be invested in learning. Thus, we propose an essential addendum to Bahrick's call, such that the problem is to identify conditions that yield both durable and efficient learning.

The highest level goal of the current work is to provide a substantial first step in this direction, which is critical for actualizing the potential of retrieval practice in many formal and informal learning environments. In the following sections, we first describe the major empirical limitations of the extant literature that motivated the current work. We then report three experiments designed to overcome these limitations and significantly expand the scope of the retrieval practice literature in order to answer the key question: How much retrieval practice is enough to achieve durable and efficient learning?

Major Empirical Limitations of the Retrieval Practice Literature

As an important preface to this section, we emphasize that we are not claiming that previous research on retrieval practice has been flawed or off-target. To the contrary, a great deal of high-quality research has established a solid foundation for future research directed at expanding the scope of the literature. Nevertheless, the literature does have some major empirical limitations. To illustrate the extent of the limitations that are discussed later, Table S1 in the online supplemental material to this article provides a summary of the methods that have been used in recent research on testing effects. Table S1 includes only studies published within the last 10 years, given that contemporary research should most accurately reflect the current state of the field. Table S2 in the supplemental material includes the list of studies published prior to 2000 that were also examined for purposes of summarizing the state of the literature in this article. Our review encompassed only studies directly exploring testing effects (including but not limited to practice test formats that most heavily involve retrieval practice, such as cued recall and free recall). Although hundreds of other studies reported in related literatures have also included conditions that involved some form of retrieval practice (including studies of retrieval-induced forgetting, generation effects, adjunct questions, hypercorrection effects, underconfidence with practice, and hypermnesia), these studies were not designed to directly explore testing effects and generally just provide converging evidence for conclusions supported by studies directly exploring testing effects.

Despite limiting our summary to direct explorations of testing effects in the past 10 years, the substantial number of studies included in Table S1 (82 articles, 168 experiments) makes clear that testing effects enjoy a widespread and growing interest. Next, we summarize the state of the literature on retrieval practice with respect to the two empirical limitations of greatest interest for answering our key question: How much is enough?

Limitation 1: Emphasis on Fixed Amounts of Practice Versus Learning to Criterion

With few exceptions, experimental manipulations in prior research on testing effects have concerned the number of practice trials during initial learning, regardless of the level of performance achieved during practice. As summarized in Table S1, many prior studies involved zero versus one practice test, and many others involved equal numbers of test versus restudy trials. Much of this work was designed to establish the efficacy of practice tests relative to no practice tests or relative to restudying for the same amount of time. An increasing amount of research has also been directed at exploring the most effective timing for a fixed number of practice trials. For example, researchers have compared massed versus spaced trials, short versus long lags between trials, and fixed versus expanding lags between trials (for recent reviews of these and other foci in testing effects research, see Roediger & Butler, 2011; Roediger & Karpicke, 2006).

Although experiments involving a fixed number of practice trials are useful for establishing testing effects, the limitation here is that they are not particularly informative for prescriptive purposes—none of the prior research provides an answer to our hypothetical student's question concerning how much practice is enough. Students typically do not (and arguably should not) simply pick an arbitrary number of trials to complete when regulating their own retrieval practice (e.g., deciding to go through a stack of flashcards twice regardless of whether items are correctly recalled). Recent survey studies (see e.g., Kornell & Bjork, 2007, 2008) have supported the intuitive assumption that students who engage in retrieval practice using flashcards do not practice for a fixed number of trials but rather continue to practice items until a desired criterion level of performance is achieved (e.g., one correct recall). But what is the appropriate criterion level for achieving both durable and efficient learning?

To our knowledge, only three out of the hundreds of testing effects studies have included experimental manipulations to examine the effects of the criterion level achieved in an initial learning session on subsequent memory performance. Importantly, note that the necessary design feature involves a *manipulation* of initial learning criterion. Although several studies have used a fixed criterion (e.g., all participants learn all items to one correct recall; Bahrick & Hall, 2005; Glover, Krug, Hannon, & Shine, 1990; Karpicke, 2009; Pyc & Rawson, 2011), few studies have manipulated criterion to compare the effects of different initial criteria on subsequent retention. Even within the literature on overlearning, all but one study (described later) examined retention as a function of the amount of practice time or practice trials completed after a fixed criterion was reached, rather than manipulating criterion (for

relevant discussion of the methods commonly used in overlearning research, see Pyc & Rawson, 2009).

In the first of the three studies involving a manipulation of initial learning criterion, Nelson, Leonesio, Shimamura, Landwehr, and Narens (1982) had participants learn number–word pairs (e.g., 76-BOOK) via cued recall practice with restudy until each item was correctly recalled 1, 2, or 4 times. Performance on a final cued recall test 4 weeks later increased monotonically as a function of initial learning criterion (18%, 29%, and 39%, respectively). Similarly, Pyc and Rawson (2007, Experiment 1) reported a monotonic increase in retention from one versus two correct recalls during practice. Pyc and Rawson (2009) examined a greater range of initial learning criteria, having participants practice Swahili–English word pairs until they were correctly recalled 1, 3, 5, 6, 7, 8, or 10 times during the initial learning session. Recalling items more than once during practice improved final test performance by 25%–190%, relative to performance for items that had been correctly recalled only once. Important for present purposes, Pyc and Rawson (2009) also established that the relationship between the number of correct recalls during practice and the incremental gain in final test performance was curvilinear, with incremental gains in final test performance from the first few correct recalls during practice but then minimal further gains in final test performance as the number of correct recalls during practice increased beyond an intermediate number.[1]

Collectively, these studies provide the only experimental evidence concerning the effects of criterion level during initial learning on subsequent retention, which is requisite for drawing strong prescriptive conclusions about optimizing schedules of retrieval practice for both durable and efficient learning. Note that all of these studies involved learning of relatively simple material, and none included conditions of relearning (the importance of which we explain later). To foreshadow, in the current experiments we manipulated initial learning criterion for more complex material followed by one or more relearning sessions, to address an important issue concerning the extent to which the effects of initial criterion persist with subsequent relearning.

Limitation 2: Emphasis on Single Learning Sessions Versus Relearning Sessions

Arguably, the greatest limitation of the current literature is that in the vast majority of prior testing effect studies, items have been practiced in one and only one learning session (see Table S1). The key here concerns the number of sessions in which the same items are practiced (as opposed to studies in which learners complete more than one practice session but with different sets of items in each session). Importantly, of those studies in which items were practiced in more than one session (those indicated in Table S1, as well as Atkinson & Paulson, 1972; Glover, 1989; Raffel, 1934; Rose, 1992; Sones & Stroud, 1940; Spitzer, 1939), the additional sessions usually involved a fixed number of practice trials rather than learning to a criterion. For the matter at hand, this design feature is just as important as with initial learning—learners would be ill advised to simply choose an arbitrary number of trials to complete during later practice sessions. Finally, of the four multiple-session studies that involved learning to a criterion during later practice sessions (Bahrick, 1979; Bahrick, Bahrick, Bahrick, & Bahrick, 1993; Bahrick & Hall, 2005; Pyc & Rawson, 2011), two studies fixed rather than manipulated the number of relearning sessions.

In the first of only two studies to manipulate the number of relearning sessions, Bahrick (1979) introduced what he referred to as the *method of successive relearning,* which involves having individuals engage in multiple sessions including alternating test and restudy trials until some prespecified criterion is reached for each item in each session. In Bahrick's study, participants learned English–Spanish word pairs, with items dropping from further practice within a session after one correct recall. After the initial learning session, items were relearned in either two or five subsequent sessions (spaced at intervals from 0 to 30 days). A final test was administered 30 days after the last relearning session. Increasing the number of relearning sessions significantly improved long-term retention (from 56% to 83% with two vs. five relearning sessions, respectively), although the possibility remains that a similarly high level of performance might have been achieved with fewer than five relearning sessions.

In what is arguably the most ambitious study to date on retrieval practice schedules (Bahrick et al., 1993), four members of the Bahrick family were the participants in one Ebbinghausian experiment. Each of them learned 300 foreign language word pairs, with items dropping from further practice within a session after one correct recall. Six conditions were defined by the number of relearning sessions (13 or 26) and the spacing between sessions (2, 4, or 8 weeks between each next session). Final retention tests were administered 1, 2, 3, or 5 years after training. Of most interest here, retention was greater after 26 versus 13 relearning sessions (67% vs. 50%, collapsing across retention interval and spacing). However, given that the training schedules examined in this study took up to 5 years to complete, the implications for making practicable prescriptive conclusions are somewhat limited.

Given that these two studies provide the only information to date concerning the relationship between amount of relearning and retention, further research on the effects of relearning criterion is clearly needed. Whereas research has almost exclusively focused on the effects of factors that are manipulated within a single session (e.g., the number, kind, or timing of practice trials in the initial and only practice session), Bahrick (1979) cogently noted that

> much of what is learned during a first exposure is forgotten during the interval between exposures and must be relearned after to become a part of semipermanent knowledge. One can, therefore, conceive of the total acquisition process as a cycle of acquisition, loss, and reacquisition of information. . . . Yet, memory research has not dealt directly with *repeated* acquisition processes. (pp. 297–298, italics in original)

[1] The only other evidence concerning the effects of initial learning criterion comes from studies involving a fixed number of practice trials that reported the results of post hoc analyses examining final retention as a function of the number of times items were correctly recalled during practice (Pyc & Rawson, 2011; Roediger & Karpicke, 2007). Although these results largely converge with those described earlier in this article, they should be interpreted with some caution, given that items were not assigned to criterion levels and thus performance differences as a function of criterion may reflect item selection effects.

Investigating the Combined Effects of Initial Learning and Relearning

Not only are relearning effects of interest in their own right, given the almost exclusive emphasis of the literature on conditions of initial learning, another focal question arises: To what extent do the effects of a variable that arise during an initial practice session persist across subsequent relearning sessions? Of greatest interest here is the effect of initial learning criterion, because of the importance of this variable for optimizing schedules of retrieval practice. Thus, answering the question of how much retrieval practice is enough to achieve efficient and durable learning will require examination of both initial learning criterion and relearning criterion, to determine whether their effects are additive or interactive. To date, no study has examined both of these factors.

Given the sheer number of retrieval practice studies (many of which are published in top-tier journals and well cited), it is perhaps surprising that this literature is largely atheoretical. Nonetheless, some of the theoretical frameworks that have more recently begun to emerge from the literature point to possible answers to the question concerning the combined effects of initial learning and relearning. Our goal here is not to competitively evaluate these theories but rather to motivate plausible outcomes. The reason for such conservatism is that these theories were not developed to explain the degree to which relearning moderates the effects of initial learning (or any condition of initial learning more generally), in large part because empirical evidence relevant to this issue is not available in the literature.

Arguably the most prominent theoretical framework in the literature, the *desirable difficulty* account (see e.g., Bjork, 1994, 1999; Schmidt & Bjork, 1992) suggests one reason why initial criterion and relearning might have superadditive effects. The central tenet of this account is that memory is enhanced by successful but effortful encoding (to a greater extent than either successful but easy encoding or effortful but unsuccessful encoding). As applied to retrieval practice in particular, the *retrieval effort hypothesis* (Pyc & Rawson, 2009) states that memory will be enhanced by successful but effortful retrieval during practice. During relearning sessions in the current experiments, all items were practiced until they were correctly recalled once. Although success of retrieval is thus held constant during relearning, the difficulty of the successful retrievals during a relearning session may still vary systematically as a function of initial learning criterion. In particular, correctly recalling items multiple times versus only once during the initial learning session would likely increase the proportion of items successfully retrieved on the first trial during relearning. Retrieving an item on the first trial (after a delay of several days since the most recent restudy opportunity) may be more difficult than retrieving an item on a later trial during relearning (after a delay of only a few minutes since the most recent restudy opportunity). If so, items practiced to a higher criterion during initial learning will accrue further benefit from more effortful retrieval during subsequent relearning.

Alternatively, strategy-based accounts of practice effects (Bahrick & Hall, 2005; Pyc & Rawson, 2010) suggest one reason why initial criterion and relearning might have subadditive effects. According to Bahrick and Hall's (2005) *strategy shift hypothesis*, attempting to retrieve information allows learners to evaluate the effectiveness of encoding strategies. In particular, the hypothesis states that when learners experience a retrieval failure, they are more likely to shift from a less effective encoding strategy to a more effective encoding strategy during subsequent restudy. To the extent that a lower initial learning criterion leads to more retrieval failures at the outset of a relearning session, learners may shift to more effective encoding of these items during restudy, which will improve the likelihood that these items are successfully recalled in a subsequent relearning session.

Of course, retrieval effort and strategy shifting are not mutually exclusive factors. Both may play a role, and to the extent that their contributions offset one another, additive effects of initial learning and relearning would be expected as well.

Experimental Overview

We conducted three experiments to explore the effects of initial learning criterion and relearning on the durability and efficiency of learning, the results of which will have important implications for both theory and practice. As summarized earlier, two key limitations of the retrieval practice literature are that few studies have manipulated either initial learning criterion or relearning criterion, and none have examined both. Furthermore, among the handful of studies including any kind of criterion manipulation, none have involved the learning of materials more complex than word pairs. This state of affairs is perhaps not surprising, given the difficulties that arise with assessing criterion learning with more complex materials. Whereas computers can reliably score word-length responses online, longer responses typically require offline handscoring. In particular, an automated scoring system with sufficient reliability does not currently exist for sentence-length responses, which are of greatest interest here. Our target materials involved key term definitions, which represent an integral part of the foundational knowledge that students need to acquire in almost any content domain. Fortunately, a technique that supports highly accurate monitoring judgments has recently been developed (described in detail later; Dunlosky, Hartwig, Rawson, & Lipko, 2011; Lipko et al., 2009; Rawson & Dunlosky, in press), such that online tracking of the criterion level attained for each item can be based on learners' judgments about the correctness of their own recall responses. Using learners' judgments in this way permitted us to instantiate Bahrick's (1979) method of successive relearning with these more complex materials.

In all experiments, students engaged in retrieval practice with restudy for key term definitions until they achieved a prespecified criterion of learning for each item. All three experiments included a manipulation of the criterion during an initial learning session. Experiment 1 included one relearning session, and Experiments 2 and 3 manipulated the number of relearning sessions. All three experiments included final retention tests administered after meaningfully long retention intervals (from 1 to 4 months). For example, most students experience about a 1-month retention interval between taking a lower level course during a fall semester and an advanced level course in the following spring semester. Likewise, a 4-month interval captures the typical delay between spring and fall semesters. The importance here is that acquiring knowledge is a cumulative process (Hambrick, 2003), such that students must retain information learned in introductory courses upon which to build their knowledge in advanced courses within that topic.

Concerning efficiency, the primary dependent measure concerns the number of practice trials needed to reach criterion in each learning session. Concerning durability, the primary dependent measures include performance on a final cued recall test and rate of relearning after final cued recall. Note that none of the prior testing effect research has used relearning rate as a measure of retention, which is perhaps not surprising given that (a) it requires learning to criterion and (b) prior research has been largely silent concerning issues of efficiency. For practical purposes, relearning rate is equally as important as final recall—even under the best practice conditions, it is implausible to expect students to recall 100% of the information learned in an introductory course when they begin an advanced course several months later. Thus, the speed with which they can relearn the unrecallable information will also have significant bearing on their success in the advanced course (e.g., time required to relearn old information reduces time available to spend learning new information). In addition to examining cued recall and relearning rate during final test sessions, we also examine performance in each relearning session to provide interim measures of retention. Taken together, these experiments provide the first examination of the extent to which the effects of initial learning and relearning are additive or interactive, as well as the first empirical investigation that compares the efficiency of retrieval practice schedules for achieving durable learning.

Experiment 1

The primary goal of Experiment 1 was to explore the effects of initial learning criterion on subsequent memory, both prior to relearning and after relearning. In this experiment, students practiced items until they were correctly recalled 1, 2, 3, or 4 times during Session 1. During Session 2, 2 days later, all students practiced items until each item was correctly recalled once. Approximately 6 weeks later, Session 3 began with a cued recall test, and then items were again relearned to a criterion of one correct recall.

Method

Participants and design. Participants included 130 undergraduates at Kent State University who participated either for course credit or for monetary compensation. Participants were randomly assigned to one of four groups, defined by the criterion level for practice during Session 1 (1, 2, 3, or 4 correct recalls).

Materials. Materials included a short passage on human memory, adapted from General Psychology textbooks (426 words, Flesch–Kincaid Grade Level 12.0). The passage included eight key terms and their definitions (e.g., "Sensory memory is a memory system that retains large amounts of sensory input for very brief periods of time," "Declarative memory is memory for specific facts and events that can be stated verbally"). Our primary reason for using course-relevant material (rather than completely unrelated content) was to enhance the motivation of participants to do their best during the learning tasks, with the idea that doing so might benefit them in their psychology course.

Procedure. In Session 1, participants were told that the study was investigating the effectiveness of different study strategies and schedules that students use to learn course material and that they would be asked to learn eight key term definitions using three different tasks (studying, recall practice, and monitoring learning). The computer provided instructions for each of the tasks, including a sample item illustrating the format of the monitoring task. Participants were told that they would practice recalling items until they had recalled them "enough times," although participants were not informed of their specific assigned criterion level. Participants were told that they would have up to 90 min to learn all eight items and that if they finished sooner they could leave early. Participants were also told that the computer would be scoring the accuracy of their responses (although in actuality, online recall accuracy was based on the participants' monitoring judgments, as described next). After participants finished reading all instructions, the experimenter confirmed that they understood the task instructions and then advanced the program to begin the experiment.

Participants were first given 4 min to study the passage. Next, the eight key term definitions were presented one at a time for a self-paced study trial. The eight items were then presented one at a time for a self-paced retrieval-monitoring-feedback (RMF) trial. Each RMF trial began with the presentation of a key term along with a text field in which the participant was prompted to type the definition. When participants clicked a button to indicate they were done with recall, they were prompted to make a monitoring judgment. As illustrated in Figure 1, the key term was presented at the top of the screen, and the participant's recall response was presented in an uneditable field. The correct definition was broken down into its main ideas, and each idea was presented in a separate field with corresponding *yes* and *no* buttons. For each idea, participants were instructed to click *yes* or *no* to indicate whether their recall response contained that idea. Participants then clicked on a button at the bottom of the screen to submit their judgments. To enhance the accuracy of the judgments, on trials in which participants indicated that they had recalled all of the idea units but their response had fewer than half the number of characters as in the correct definition, the program displayed the following message: *Your decision that your response includes all of the idea units is incorrect. Please revisit your idea-unit judgments and revise them accordingly.* After the idea-unit judgment was completed, a feedback screen presented the key term and intact definition for self-paced restudy.

During the experiment, a recall response was counted as correct if participants judged that all of the idea units were contained in their response.[2] The program tracked how many times each item had been judged as entirely correct. On any given RMF trial, if participants did not judge the recall response as completely correct,

[2] On the basis of our postexperiment scoring of participants' responses, the accuracy of the responses that participants judged as completely correct was 77% ($SE = 2$) in Session 1 and 80% ($SE = 2$) in Session 2, with no significant differences between groups in either session (Fs < 1.57). Participants' modest overconfidence in the accuracy of their responses is consistent with results of earlier work (Dunlosky et al., 2011). To foreshadow, the accuracy of responses that participants judged as completely correct in Experiment 2 was similar—81% ($SE = 2$) in Session 1, 88% ($SE = 2$) in Session 2, 93% ($SE = 2$) in Session 3, and 92% ($SE = 2$) in Session 4—with no significant differences between groups in any session (ts < 0.87). Similarly, the accuracy of responses that participants judged as completely correct in Experiment 3 ranged from 87% to 92% (SEs $=$ 0.7–0.9) across sessions, with minimal differences between groups or conditions in each session.

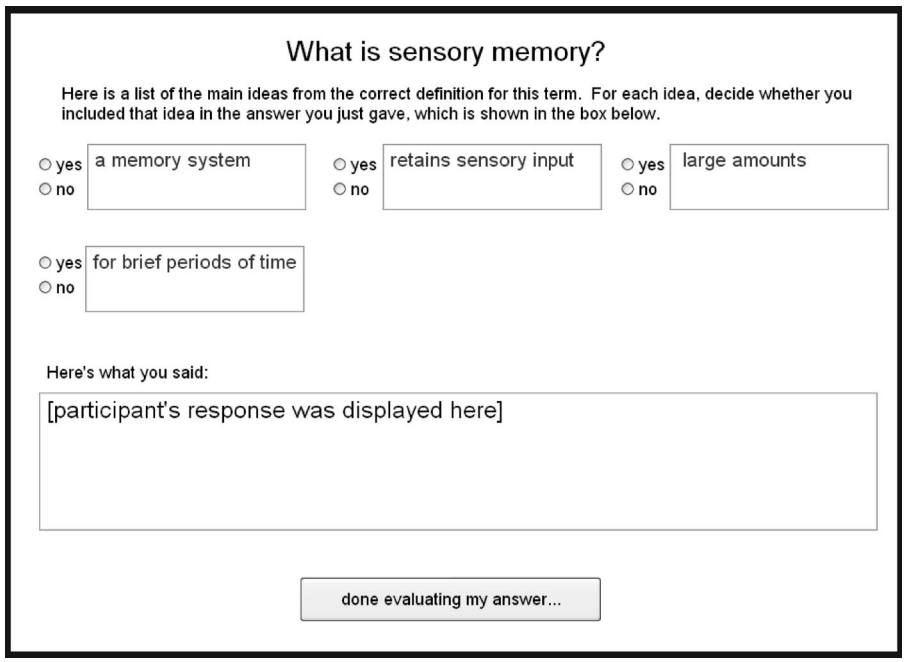

Figure 1. Illustration of the method used in all three experiments to collect participants' monitoring judgments about the accuracy of their recall responses.

that item was placed at the end of the list for another RMF trial later. Likewise, if participants judged the recall response as completely correct but the number of times the item had been judged as correctly recalled was below the assigned criterion level, the item was placed at the end of the list for another RMF trial later. Once an item was judged as completely correct the assigned number of times (e.g., the third time in the Criterion 3 group), practice for that item was discontinued for the remainder of the session. Once all items had reached the assigned criterion level (or once 90 min had elapsed), participants were dismissed and reminded to return 2 days later.

Session 2 began with one RMF trial for each of the eight key terms. Note that no restudy opportunity was provided prior to the first recall attempt, and thus performance on the first RMF trial serves as an interim retention test. Each item continued to be presented for RMF trials until a criterion level of one correct recall was reached (on the basis of participants' monitoring judgments, as in Session 1).

Session 3 took place approximately 6 weeks later ($M = 39.6$ days, $SD = 4.0$). Session 3 began with one cued recall test for each item, without restudy. Each item was then presented for RMF trials until a criterion of one correct recall was reached. Participants were also asked to indicate for each item whether they had encountered that item in their General Psychology class. On average across groups, participants reported having encountered 40.5% ($SD = 41.2\%$) of the items in their class, with no significant differences between groups ($F < 1$).

Scoring. We describe the scoring procedure used for all three experiments here. All recall responses generated during practice and final test sessions were scored. Each response was assigned a recall score based on the percentage of main ideas from the definition that the response contained (the main ideas were the same as those presented to participants for monitoring judgments during practice). Responses were counted as correct if they included either verbatim restatements or paraphrases that preserved the meaning of the definition.

Given the sizeable number of recall responses to be hand-scored across experiments (8,069 responses in Experiment 1; 6,453 responses in Experiment 2; 86,010 responses in Experiment 3), multiple raters were trained to complete the scoring. A total of four sets of experimental items were used across experiments. For each item set, we chose a random sample of protocols to serve as the training set. Each rater scored the training set, and the reliability of his or her scores was checked against the scores of other raters for that item set. Given that reliability was consistently high across raters and item sets, each remaining protocol was then scored by one of the trained raters.

Results

Data for 17 participants who failed to return for Session 2 were excluded from analyses. Although 25 other participants did not complete Session 3 (10 due to attrition and 15 who were not scheduled to complete Session 3 due to end of the semester), their data were not excluded from analyses because Session 2 also included an interim test, and thus their data are still informative. As a result, for Initial Criterion Groups 1–4, analyses are based on respective *n*s of 28, 30, 29, and 26 for Session 2 and *n*s of 22, 26, 22, and 19 for Session 3. Effect size estimates reported for all experiments are based on pooled variance.

Durability of learning.

Effects of initial learning criterion on cued recall performance in Sessions 2 and 3. For each participant, we computed the mean percentage recall across items on the first test trial in Session

2 and in Session 3. Mean performance across participants in each criterion level group is reported in Figure 2. A 4 (initial criterion level) × 2 (session) mixed-factor analysis of variance (ANOVA) yielded a significant interaction, $F(3, 84) = 2.86$, $MSE = 216.18$, $p = .042$, $\eta_p^2 = .09$. To diagnose the nature of the interaction, we conducted follow-up analyses for each session. For Session 2, a one-way ANOVA indicated that recall performance differed significantly as a function of Session 1 criterion level, $F(3, 109) = 4.54$, $MSE = 436.31$, $p = .003$, $\eta_p^2 = .11$. Most important, recall was significantly poorer when items had been correctly recalled only once during practice versus two times, $t(56) = 1.94$, $p = .028$, $d = 0.52$; three times, $t(55) = 3.35$, $p = .001$, $d = 0.90$; or four times, $t(52) = 3.06$, $p = .002$, $d = 0.85$. However, a pattern of diminishing returns was observed: Recall tended to be lower after two correct recalls during practice versus three or four, $t(57) = 1.35$, $p = .091$, $d = 0.36$, and $t(54) = 1.17$, $p = .123$, $d = 0.32$, respectively, and the latter two groups did not differ significantly, $t(53) = 0.11$. Overall, the group means were better fit by the curvilinear logarithmic function than a linear function ($R^2 = .95$ vs. .84, respectively), confirming that the incremental benefit to recall performance diminished as initial criterion level increased.

The pattern of results for Session 3 contrast strikingly with those for Session 2. Relearning items to one correct recall during Session 2 almost completely eliminated the effects of Session 1 criterion level. A one-way ANOVA showed no effect of Session 1 criterion level on cued recall in Session 3 ($F < 1$), and pairwise comparisons revealed no significant differences between any groups (all ts < 0.64, dfs ≥ 39). These results provide the first indication that the effects of initial criterion and relearning are subadditive, although this pattern is examined more extensively in Experiments 2 and 3.

Effects of initial learning criterion on relearning rate in Sessions 2 and 3. As a converging measure of memory, the number of RMF trials required to reach one correct recall for each item in Sessions 2 and 3 was examined. Mean number of trials per item across participants in each group is reported in Figure 3 (trials per item in Session 1 are also reported in Figure 3, but these results are relevant to the issue of learning efficiency and thus are addressed in the next section). Three individuals' values from Session 2 and five values from Session 3 were more than 3 *SD*s above the group mean and were therefore excluded from analyses as outliers. A 4 (initial criterion level) × 2 (session) mixed-factor ANOVA yielded main effects of session and initial criterion level, $F(1, 79) = 14.54$, $MSE = 0.49$, $p < .001$, $\eta_p^2 = .16$, and $F(3, 79) = 2.68$, $MSE = 0.86$, $p = .052$, $\eta_p^2 = .09$, respectively; the interaction was not significant, $F(3, 79) = 2.01$, $MSE = 0.49$, $p = .119$, $\eta_p^2 = .07$. To parallel the analyses reported for cued recall, we report follow-up tests to explore the effect of initial criterion level. More trials were needed to reach criterion in Session 2 when items had been correctly recalled only once during Session 1 versus two times, $t(54) = 2.68$, $p = .005$, $d = 0.74$; three times, $t(52) = 1.84$, $p = .035$, $d = 0.51$; or four times, $t(50) = 3.05$, $p = .002$, $d = 0.86$. The number of trials needed to reach criterion in Session 2 did not differ significantly for the latter three groups (ts < 1.58, dfs ≥ 52, ps $> .06$). Overall, the group means were better fit by a logarithmic function than a linear function ($R^2 = .76$ vs. .65, respectively), confirming that the incremental benefit to relearning rate diminished as criterion level increased.

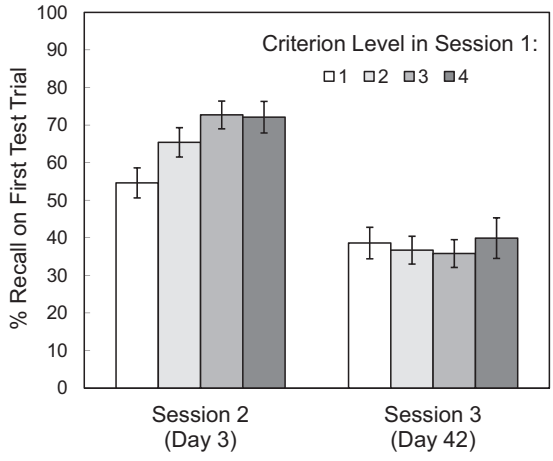

Figure 2. Experiment 1: Mean percentage recalled on the first test trial in Session 2 and Session 3 as a function of criterion level during Session 1. Error bars represent standard error of the mean.

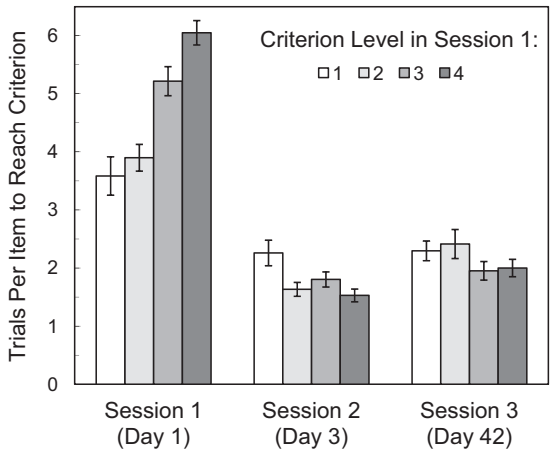

Figure 3. Experiment 1: Mean number of retrieval-monitoring-feedback trials per item needed to reach criterion during initial learning (Session 1) and during relearning (Sessions 2 and 3), as a function of criterion level during Session 1. Criterion level during Session 1 was one, two, three, or four correct recalls; criterion in Sessions 2 and 3 was one correct recall. Error bars represent standard error of the mean.

For Session 3 (the rightmost set of bars in Figure 3), none of the pairwise comparisons reached statistical significance (all ts < 1.48, dfs ≥ 36, ps $> .07$). However, a pairwise comparison collapsing the Criterion 1–2 groups and the Criterion 3–4 groups confirms the pattern apparent from visual inspection of Figure 3, such that the lower criterion groups required more relearning trials than did the higher criterion groups, $t(83) = 1.97$, $p = .026$, $d = 0.43$. Thus, whereas cued recall showed no effect of initial criterion on Session 3 performance, relearning rate provides some evidence for a small but persistent effect of initial criterion.

Efficiency of learning. To summarize the results thus far, higher criterion levels during Session 1 led to greater recall during Session 2 and faster relearning overall but had minimal effects on recall in Session 3. However, obtaining higher criterion levels in

Session 1 required more RMF trials during Session 1, as shown in Figure 3 (left set of bars). To evaluate the overall efficiency of the practice schedules, for each individual we computed the total number of RMF trials per item combined across Sessions 1 and 2. Means across individual values in each group are reported in Table 1. Mean recall in Session 3 is also included in the Final Cued Recall row of the table for purposes of evaluating the optimality of each schedule (e.g., one might choose the most efficient schedule from among those that achieve some minimum level of final recall). For total trials per item, a one-way ANOVA revealed a significant effect of initial learning criterion, $F(3, 102) = 6.67$, $MSE = 3.50$, $p < .001$, $\eta_p^2 = .16$. Total trials per item was significantly lower in the Criterion 1 group than in the Criterion 3 and 4 groups, $t(49) = 1.79$, $p = .040$, $d = 0.51$, and $t(49) = 3.20$, $p = .001$, $d = 0.91$, respectively. Likewise, total trials per item was significantly lower in the Criterion 2 group than in the Criterion 3 and 4 groups, $t(53) = 2.72$, $p = .005$, $d = 0.75$, and $t(53) = 4.77$, $p < .001$, $d = 1.32$, respectively. The difference between the latter two groups was also significant, $t(48) = 1.86$, $p = .035$, $d = 0.54$.

Experiment 2

Although we refrain from drawing strong prescriptive conclusions concerning schedule optimality until the results of all three experiments have been presented, the results of Experiment 1 suggest that with respect to initial learning criterion, more is not necessarily better when both durability and efficiency are taken into account. Although performance in Session 2 increased as initial learning criterion increased, diminishing returns were apparent: Four versus three correct recalls during Session 1 was no better with respect to durability and was worse with respect to efficiency. Even more striking, whereas strong effects of initial learning criterion on memory were observed prior to relearning, one relearning session was sufficient to markedly reduce this advantage. When considered in light of the total number of practice trials spent during practice, practice schedules involving higher initial criterion levels were relatively inefficient schedules of learning.

To further explore the combined effects of initial criterion and relearning, we designed Experiment 2 to provide a partial replication of Experiment 1 (initial learning criterion of one vs. three correct recalls) as well as an important extension to manipulate the number of relearning sessions (one vs. three). In this experiment, during initial learning, students practiced items until they were correctly recalled either one or three times. Two days later, all items were relearned to a criterion of one correct recall. Half of the items were relearned a second time 5 days later and then a third time 2 days after that. Students completed a final test approximately 5 weeks after the last relearning session.

Table 1
Mean RMF Trials per Item Across All Practice Sessions and Final Cued Recall as a Function of Practice Schedule in Experiment 1

Variable	Initial criterion level			
	One	Two	Three	Four
Total trials per item[a]	5.7 (0.5)	5.5 (0.3)	6.8 (0.3)	7.5 (0.2)
Final cued recall %	38.6 (4.2)	36.6 (3.7)	35.8 (3.7)	39.9 (5.4)

Note. Initial criterion level refers to the number of times an item was correctly recalled during Session 1. Values in parentheses are standard errors of the mean. RMF = retrieval-monitoring feedback.
[a] Computed as the total number of RMF trials per item across practice sessions (i.e., including trials to criterion in Session 1 plus trials for subsequent relearning in Session 2).

Method

Participants and design. Participants included 68 undergraduates at Kent State University who participated either for course credit or for monetary compensation. Participants were randomly assigned to one of two groups, defined by the criterion level for practice during Session 1 (one or three correct recalls). Number of relearning sessions (one or three) was a within-subject manipulation.

Materials. Materials included the passage and items used in Experiment 1, along with a second material set. The new passage was adapted from General Psychology textbooks (605 words, Flesch–Kincaid Grade Level 12.0) and included eight key terms about attribution (e.g., "Attribution is the process through which we seek to determine the causes behind people's behavior," "The just-world hypothesis is the strong desire or need people have to believe that the world is an orderly, predictable, and just place, where people get what they deserve").

Procedure. In addition to describing the practice schedule in more detail in this section, we summarize it in Table 2. The procedure in Session 1 was the same as in Experiment 1, with participants practicing items until each was judged as correctly recalled either once or three times depending on group assignment. Participants were asked to learn only one of the two sets of eight items (as described later, the other set of items was presented only during the final cued recall test, to serve as a baseline knowledge condition). Assignment of item set to practice or baseline condition was counterbalanced across participants.

Two days later, participants returned to complete Session 2, which was the same as in Experiment 1 (briefly, practice began with a self-paced RMF trial for each item and continued until each item had been correctly recalled once). Five days after Session 2, participants completed Session 3. The procedure in Session 3 was the same as in Session 2, except that only four of the eight items were presented for RMF trials until each one had been correctly recalled once. Two days later, participants completed Session 4. The procedure in Session 4 was the same as in Session 3, with the same set of four items presented until each one had been correctly recalled once. The subset of items assigned to one versus three relearning sessions was counterbalanced across participants. Note that no restudy opportunity was provided prior to the first recall attempt in Sessions 2–4, and thus performance on the first RMF trial in each session also serves as an interim retention test.

Approximately 5 weeks after Session 4 ($M = 35.7$ days, $SD = 3.3$ days), participants returned for Session 5, which began with a

Table 2
Presentation Schedule for Items in Experiment 2 as a Function of Relearning Condition

	No. of relearning sessions	
Schedule and test	One	Three
Practice schedule		
Session 1 (Day 1)[a]	Items 1–4	Items 5–8
Session 2 (Day 3)	Items 1–4	Items 5–8
Session 3 (Day 8)		Items 5–8
Session 4 (Day 10)		Items 5–8
Final test		
Session 5 (Day 45)[b]	Items 1–4	Items 5–8
	Items 9–12	Items 13–16

Note. The actual items designated as Items 1–4, Items 5–8, Items 9–12, and Items 13–16 were counterbalanced across participants.
[a] On Day 1, items were practiced until they were correctly recalled either one or three times, depending on assignment to initial criterion group, which is not represented in this table. In all subsequent relearning sessions, all participants practiced items until they were correctly recalled once. [b] The day of the final test is approximate, with some minor deviation to accommodate participants' schedules.

final cued recall test.[3] All 16 key terms (the eight practiced items and the eight baseline control items that had not been presented at any point during practice) were presented one at a time in random order, and participants were prompted to type the definition of each term into a response field. Participants were told that for items that had not been practiced during the experiment, they should respond on the basis of anything they had learned about that item in their General Psychology class. After the recall response for each of the 16 items, participants were asked to indicate whether they had encountered that item in their General Psychology class. On average, participants reported having encountered 36.7% (SD = 25.6%) of the items in their class, with no significant main effects or interaction as a function of criterion level or number of relearning sessions (Fs < 1.03).

One other modification of the method in Experiment 2 concerned the converging measure of memory collected after the final cued recall test. To allow for the possibility that a recognition measure might be more sensitive than relearning rate for detecting marginal knowledge (Berger, Hall, & Bahrick, 1999), we replaced the relearning task with a multiple-choice test of definition recognition. For each question, a key term was presented as the prompt, followed by four definition alternatives (the correct definition along with definitions of other terms and/or incorrect paraphrases of the definition).

Results

Data from 10 participants were excluded from analyses because they did not complete Session 2, because they did not reach at least 80% criterion across sessions, or because they failed to comply with instructions. Data for seven participants who did not complete Sessions 3, 4, and/or 5 were not excluded from analyses, because their data for interim tests in the relearning sessions were still informative. For Initial Criterion Groups 1 and 3, analyses are based on respective ns of 30 and 28 for Session 2; 29 and 26 for Session 3; 27 and 26 for Session 4; and 26 and 26 for Session 5.

Durability of learning

Effects of initial learning criterion on cued recall performance on the interim tests in Sessions 2–4. For each participant, we computed the mean percentage recall across items on the first test trial for Sessions 2, 3, and 4. Mean recall performance across participants is reported in Figure 4. A 2 (one or three correct recalls in Session 1) × 3 (first trial in Session 2, 3, or 4) mixed-factor ANOVA indicated a significant main effect of session, $F(2, 100) = 17.59$, $MSE = 201.74$, $p < .001$, $\eta_p^2 = .26$, and a trend toward a main effect of Session 1 criterion, $F(1, 50) = 3.41$, $MSE = 1{,}673.53$, $p = .071$, $\eta_p^2 = .06$; the interaction was not significant, $F(2, 100) = 1.91$, $p = .154$, $\eta_p^2 = .04$.

As in Experiment 1, correctly recalling items three times versus one time during Session 1 led to significantly greater recall performance on the first trial in Session 2, $t(56) = 2.34$, $p = .011$, $d = 0.63$. Despite the fact that all items were relearned to a criterion of one correct recall during Session 2, the effect of Session 1 criterion level tended to persist on the first trial in Session 3, $t(53) = 1.40$, $p = .083$, $d = 0.39$. Thus, in contrast to in Experiment 1, the effect of initial learning criterion in Experiment 2 was not completely eliminated by one relearning session, suggesting that Experiment 1 may have underestimated the persistence of initial criterion effects. We have no explanation for this quantitative difference in the degree of attenuation of initial criterion effects, although the most obvious methodological difference concerns the retention interval between Sessions 2 and 3 (6 weeks in Experiment 1 versus 5 days in Experiment 2; to foreshadow, the results of Experiment 3 replicate the qualitative pattern in Experiment 2). Likewise, despite the fact that items were relearned a second time during Session 3, a numerical trend for an effect of Session 1 criterion level was still evident on the first trial in Session 4, $t(51) = 0.94$, $p = .177$, $d = 0.26$.

Effects of initial learning criterion on relearning rate in Sessions 2–4. To further explore the extent to which effects of initial learning criterion persist, we again examined relearning rate as a converging measure of memory. For each participant, we computed the mean number of RMF trials per item needed to reach one correct recall in Sessions 2, 3, and 4. Mean trials per item across participants is reported in Figure 5. A 2 (one or three correct recalls in Session 1) × 3 (Session 2, 3, or 4) mixed-factor ANOVA indicated a significant main effect of session, $F(2, 100) = 17.01$, $MSE = 0.37$, $p < .001$, $\eta_p^2 = .25$; a main effect of Session 1 criterion, $F(1, 50) = 6.34$, $MSE = 3.38$, $p = .015$, $\eta_p^2 = .11$; and a significant interaction, $F(2, 100) = 5.96$, $p = .004$, $\eta_p^2 = .11$. Replicating Experiment 1, fewer trials were needed to reach criterion in Session 2 when items had been correctly recalled three

[3] Because Session 5 took place 5 weeks after the last relearning session, the retention interval was 1 week longer for items relearned once than for items relearned three times (see Table 2). Given the well-established finding that forgetting is negatively accelerated (i.e., relatively rapid forgetting shortly after learning, with increasingly minimal loss at longer delays), we assumed a minimal functional difference between a 5-week versus 6-week retention interval. To foreshadow, the nominal retention interval was equated in Experiment 3 by manipulating relearning between subjects rather than within subjects, and highly similar results were obtained.

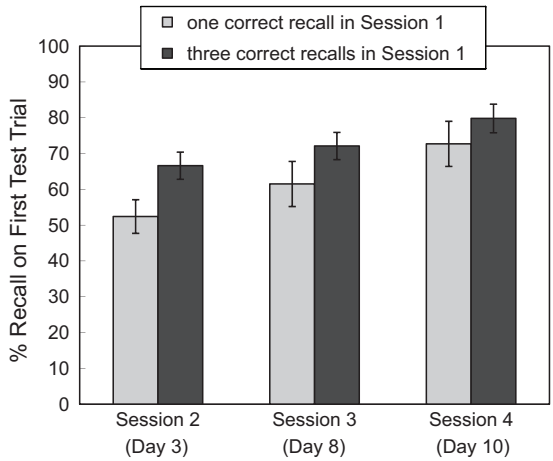

Figure 4. Experiment 2: Mean percentage recalled on the first test trial in Sessions 2–4 as a function of criterion level during Session 1. Error bars represent standard error of the mean.

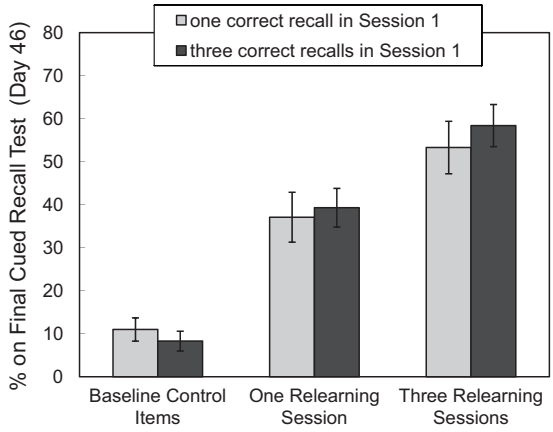

Figure 6. Experiment 2: Mean performance on the final cued recall test in Session 5 as a function of Session 1 criterion level (either one or three correct recalls) and as a function of the number of relearning sessions (practiced items were relearned either once or three times; baseline control items were not practiced in any session). Error bars represent standard error of the mean.

times versus one time during Session 1, $t(56) = 3.69$, $p < .001$, $d = 0.99$. Mirroring the pattern of recall performance, this trend persisted in Session 3, $t(53) = 2.22$, $p = .016$, $d = 0.61$, and although weaker was still apparent in Session 4, $t(51) = 1.10$, $p = .139$, $d = 0.31$.

Effects of initial learning criterion and relearning on final cued recall performance in Session 5. For each participant, we computed mean percentage correct on the final cued recall test in Session 5 for items that had been relearned once, items relearned three times, and the baseline control items that were not practiced in any session. Mean recall across participants is reported in Figure 6. Performance for baseline control items did not significantly differ as a function of initial learning criterion groups, $t(50) = 0.75$, suggesting minimal differences in baseline levels of knowledge in the two groups. Performance was significantly greater in both relearning conditions than in the baseline control condition, $t(51) = 6.59$ with one relearning session and $t(51) = 9.75$ with three relearning sessions, establishing significant effects of practice beyond any learning of items acquired outside of the experiment.

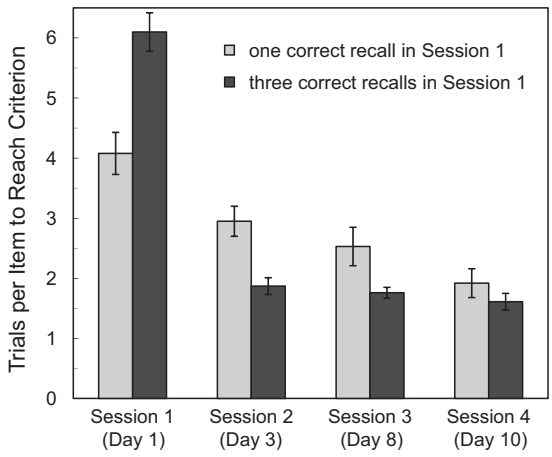

Figure 5. Experiment 2: Mean number of retrieval-monitoring-feedback trials per item needed to reach criterion during initial learning (Session 1) and during relearning (Sessions 2–4) as a function of criterion level during Session 1. Criterion level during Session 1 was one or three correct recalls; criterion in Sessions 2–4 was one correct recall. Error bars represent standard error of the mean.

A 2 (one or three correct recalls in Session 1) × 2 (one or three relearning sessions) mixed-factor ANOVA revealed only a main effect of the number of relearning sessions, $F(1, 50) = 42.38$, $MSE = 191.41$, $p < .001$, $\eta_p^2 = .46$ (other Fs < 1). Thus, the trend for persistent effects of initial learning criterion on the interim tests in the relearning sessions continued to weaken on the final test (45% vs. 49%, $d = 0.15$). To directly compare the effects of initial criterion prior to relearning versus after relearning, a 2 (one or three correct recalls in Session 1) × 2 (recall on the first trial of Session 2 vs. recall in Session 5) ANOVA yielded a significant interaction, $F(1, 49) = 6.12$, $MSE = 252.47$, $p = .017$, $\eta_p^2 = .11$, which further establishes the attenuated effects of initial learning criterion after relearning sessions.

In contrast to the modest and waning effects of initial criterion, the number of relearning sessions produced sizeable effects on memory. Final recall performance was substantially greater for items that were relearned three times than for items relearned once (56% vs. 38%, $d = 0.65$). Thus, increasing the relearning criterion had substantial effects on retention.

Effects of initial learning criterion and relearning on multiple-choice performance in Session 5. On the multiple-choice test, performance for the baseline control items did not significantly differ as a function of initial learning criterion (Criterion 1 = 45.2%, $SE = 5.4$, vs. Criterion 3 = 42.3%, $SE = 4.0$), $t(50) = 0.67$, $d = 0.12$. For practiced items, a 2 × 2 mixed-factor ANOVA revealed no main effects or interaction (Fs < 1). Thus, performance for practiced items did not differ as a function of initial criterion (Criterion 1 = 72.1%, $SE = 3.2$; Criterion 3 = 70.7%, $SE = 3.9$, $d = 0.08$) or relearning (one relearning session = 69.2%, $SE = 3.7$; three relearning sessions = 73.6%, $SE = 3.0$,

$d = 0.18$), although performance in all practice conditions was significantly greater than for baseline items, $ts(25) > 3.75$.

Although these null effects may seem surprising, other research has shown that testing manipulations can exert robust effects on recall measures but weaker or nonexistent effects on recognition measures. For example, Darley and Murdock (1971) had participants study word lists that either received a practice recall test or included no practice test prior to a final test. When the final test was free recall, memory was greater for items that were tested versus untested during practice, but no effect was observed when the final test was recognition. On the basis of these results, Darley and Murdock concluded that practice tests primarily affect the accessibility rather than the availability of information in memory. Similarly, the current pattern may suggest that manipulations of initial criterion and relearning primarily affect accessibility rather than availability.

Because recognition appears to be less rather than more sensitive to criterion effects, we returned to use of the relearning measure after the final cued recall test in Experiment 3. However, one final note is in order. Given that many undergraduate courses involve multiple-choice tests, one might wonder about the educational recommendations concerning increasing initial criterion or relearning. Importantly, the highest level goal of this research was to examine how much retrieval practice is enough to achieve durable and efficient learning, not how much is needed for doing well on a course exam per se. Indeed, some research would suggest that if the goal is simply to do well on an exam, a relatively effective practice schedule is to cram immediately prior to the test (see e.g., Glenberg & Lehmann, 1980; Rawson & Kintsch, 2005), but this schedule is notoriously bad for durable learning. In the General Discussion, we consider the extent to which the educational recommendations that follow from this research will satisfy students' short-term goals of doing well on exams as well as the longer term goal of retaining learning well beyond an exam.

Efficiency of learning. To summarize the results thus far, we found that practicing until items were correctly recalled three times versus once during Session 1 produced relatively modest improvements in recall in later sessions. Of course, more practice trials were required to obtain the higher criterion in Session 1, as shown in Figure 5 (left set of bars), but this additional cost was partially offset by faster relearning in subsequent practice sessions. Concerning the effects of number of relearning sessions, three versus one relearning session produced significant improvements in final recall performance, but at the obvious expense of additional practice trials in those sessions.

To estimate the overall efficiency of these practice schedules, we computed two measures for each participant: the total number of RMF trials per item needed to reach criterion in Sessions 1 and 2 (for the condition in which items were relearned only once) and the total number of RMF trials per item needed to reach criterion in Sessions 1–4 (for the condition in which items were relearned three times). Means across individual values in each group are reported in Table 3. Mean recall in Session 5 is also included in the Final Cued Recall row of the table for purposes of evaluating the optimality of each schedule.

A 2 (one or three correct recalls in Session 1) × 2 (one or three relearning sessions) mixed-factor ANOVA revealed no effect of initial learning criterion ($F < 1$); a significant main effect of relearning criterion, $F(1, 50) = 160.35$, $MSE = 2.54$, $p < .001$, $\eta_p^2 = .76$; and a marginal interaction, $F(1, 50) = 3.36$, $p = .073$, $\eta_p^2 = .06$. Thus, the cost of additional practice trials to reach a higher criterion in Session 1 was largely recouped across subsequent relearning opportunities. Nevertheless, although the ultimate costs associated with a higher initial criterion were minimal, the gains in final test performance were also minimal. In contrast, significant cost was associated with increasing the number of relearning sessions from one to three, but final test performance was also improved substantially. To consider the costs and benefits of additional relearning sessions using other metrics, we found that relearning items during two additional sessions amounted to a total of less than 3 min of additional practice time per item ($M = 2.7$ min, $SE = 0.1$) but yielded a 46% relative improvement in recall over a single relearning session.

Table 3
Mean RMF Trials per Item Across All Practice Sessions and Final Cued Recall as a Function of Practice Schedule in Experiment 2

	One relearning session		Three relearning sessions	
Variable	One	Three	One	Three
Total trials per item[a]	7.0 (0.5)	8.1 (0.4)	11.7 (1.1)	11.2 (0.6)
Final cued recall %	37.1 (5.8)	39.3 (4.5)	53.3 (6.1)	58.4 (4.9)

Note. *One* and *Three* indicate the initial criterion level, which refers to the number of times an item was correctly recalled during Session 1. Values in parentheses are standard errors of the mean. RMF = retrieval-monitoring feedback.
[a] Computed as the total number of trials per item across practice sessions (i.e., including trials to criterion in Session 1 plus trials in any subsequent relearning session).

Experiment 3

Replicating the results of Experiment 1, Experiment 2 showed that initial learning criterion had sizeable effects on retention prior to relearning but that these effects were attenuated after relearning. However, Experiment 2 provided some evidence that relearning may not completely eliminate initial criterion effects. One goal of Experiment 3 was to further explore this possibility using a much larger sample.

Perhaps of greater importance, increasing the number of relearning sessions had a striking effect on long-term retention, with two additional relearning sessions yielding a relative increase of 46% over just one relearning session. The second goal of Experiment 3 was to further explore effects of relearning by examining a larger range of relearning sessions. In so doing, Experiment 3 represents the first experiment to examine more than two levels of relearning, which in turn permits an examination of the extent to which the effects of relearning on retention are linear or curvilinear. Will retention show similar incremental benefits from each next relearning session, or will it exhibit a pattern of diminishing returns such as those observed with initial learning criterion?

In this experiment, during initial learning, students practiced items until they were correctly recalled either one or three times. Students then practiced items to one correct recall in 1, 2, 3, 4, or 5 subsequent relearning sessions. Final tests were

completed 1 month and 4 months after the last relearning session.

Method

Participants and design. Participants included 335 undergraduates enrolled in General Psychology at Kent State University who participated for monetary compensation. Participants were randomly assigned to one of 10 groups, defined by the factorial combination of initial learning criterion (one or three correct recalls) and number of relearning sessions (1, 2, 3, 4, or 5).

Materials. Materials included the two passages and item sets used in Experiment 2, along with two new material sets. The new passages were adapted from General Psychology textbooks (583 and 575 words, Flesch–Kincaid Grade Levels of 10.6 and 11.7). One passage included eight key term definitions about depth perception (e.g., "Convergence refers to information from the ocular muscles about how far the eyes have to turn in to focus on an object") and the other passage included eight key term definitions about the nervous system (e.g., "The parasympathetic nervous system is the system that is active when the body is in a relaxed state, working to calm the body and conserve energy").

Procedure. In each of the 10 groups, participants practiced two sets of items, and the other two sets of items were presented only during the final test sessions to serve as a baseline control condition. To keep the length of practice sessions reasonable, we introduced the two sets of practiced items on different days. Because of the number of experimental groups and the introduction of items on different days, the practice schedule for items was somewhat complex, so we summarize it in Table 4 in addition to describing it in more detail next.

The procedure on Day 1 was the same as in Experiment 2, with participants studying and then practicing the first set of eight items until each was judged as correctly recalled either one or three times depending on group assignment. The procedure on Day 3 was the same as Session 2 in Experiment 2, with RMF trials until each item had been judged as correctly recalled once. The procedure on Day 8 was the same as on Day 1 except that initial study and RMF practice was completed for the second set of eight items. Day 10 began with RMF trials for the second set of eight items until each item was judged as correctly recalled once. For participants assigned to relearn items more than once, the first set of eight items was then presented for RMF trials until each item was again correctly recalled once. Blocks of practice were completed for each set of items in a similar manner on subsequent days as illustrated in Table 4, with relearning always involving RMF trials to one correct recall for each item. As in the previous experiments, no restudy opportunity was provided prior to the first recall attempt for an item in any relearning session, and thus performance on the first RMF trial in each session served as an interim retention test.

Approximately one month ($M = 29.1$ days, $SD = 3.1$) after the last practice session in each group, participants returned for the 1-month test session. The session began with one cued recall test (without restudy) for half of the items in each set of practiced items, as well as a cued recall test for half of the baseline control items. As in Experiment 2, participants were told that for items that they had not practiced during the experiment, they should respond on the basis of anything they had learned about that item in their General Psychology class. Items were presented one at a time, and after entering a recall response for an item, participants indicated whether they had encountered that item in their General Psychol-

Table 4
Presentation Schedule for Items in Experiment 3 as a Function of Group

	No. of relearning sessions				
Practice and test	One	Two	Three	Four	Five
Practice schedule					
Day 1[a]	Items 1–8	Items 1–8	Items 1–8	Items 1–8	Items 1–8
Day 3	Items 1–8	Items 1–8	Items 1–8	Items 1–8	Items 1–8
Day 8	Items 9–16	Items 9–16	Items 9–16	Items 9–16	Items 9–16
Day 10	Items 9–16	Items 9–16	Items 9–16	Items 9–16	Items 9–16
		Items 1–8	Items 1–8	Items 1–8	Items 1–8
Day 17		Items 9–16	Items 1–8	Items 1–8	Items 1–8
			Items 9–16	Items 9–16	Items 9–16
Day 24			Items 9–16	Items 1–8	Items 1–8
				Items 9–16	Items 9–16
Day 31				Items 9–16	Items 1–8
					Items 9–16
Day 38					Items 9–16
1-month test[b]	Day 40	Day 47	Day 54	Day 61	Day 68
4-month test[b]	Day 130	Day 137	Day 144	Day 151	Day 158

Note. The actual items designated as Items 1–8, Items 9–16, and so on were counterbalanced across participants in each group. For all five groups, the 1-month test involved a cued recall test for Practice Items 1–4 and 9–12 and for Baseline Control Items 17–20 and 24–28 and then relearning for Items 1–4 and 9–12. For all five groups, the 4-month test involved a cued recall test for Items 1–32 (i.e., all practiced and baseline items) and then relearning for Items 1–16.

[a] On Day 1, items were practiced until they were correctly recalled either one or three times, depending on assignment to initial criterion group, which is not represented in this table. In all subsequent relearning sessions, all participants practiced items until they were correctly recalled once. [b] Days are approximate, with some minor deviation to accommodate participants' schedules.

ogy class. Overall, participants reported having encountered 49.1% (*SD* = 21.9%) of the items in their class.[4] After the cued recall test, each of the eight practiced items tested in that session were then presented for RMF practice trials until a criterion of one correct recall was reached.

Approximately four months (*M* = 120.0 days, *SD* = 9.3) after the last practice session in each group, participants returned for the 4-month test session. The session began with a cued recall test for all 32 items. The procedure for this test was the same as in the 1-month test session. Overall, participants reported having encountered 58.0% (*SD* = 23.1%) of the items in their class. After the cued recall test, all 16 practiced items were then presented for RMF trials until each item was correctly recalled once.

Despite the complexity of the schedule, we counterbalanced (a) assignment of item set to practice or baseline condition, (b) which of the two sets of practiced items was introduced first, and (c) which half of the items in each set was tested on the 1-month test. Finally, immediately after the RMF trial in which an item reached criterion in each session, participants were asked to predict how well they would remember that item at the beginning of the next session. These data are not of interest for present purposes and are not discussed further.

Results

In summary, practice for the two sets of practiced items spanned different days, but both sets of items were relearned the same number of times. For all analyses in the following sections, we collapsed across the two sets of items. In keeping with our terminology in Experiments 1–2, we use the term *Session 1* to refer to the initial learning session (i.e., combining Day 1 for the first set of practiced items and Day 8 for the second set; see Table 4) and *Sessions 2–6* to refer to performance during the relearning sessions.

Data were excluded from analyses for 22 participants who did not complete Session 2 and for 10 participants who did not reach at least 80% criterion across sessions. Data for participants who failed to complete one or more of the later sessions were not excluded from analyses, because these sessions also included interim retention tests, and thus their data are still informative. As a result, analyses are based on group *n*s of 26–34 for Sessions 2–4; 25–33 for Session 5; 33 for Session 6; 24–31 for the 1-month test session; and 21–30 for the 4-month test session.

Durability of learning.

Effects of initial learning criterion on cued recall performance in Sessions 2–6. For each participant, we computed the mean percentage recall across items on the first test trial in each of the relearning sessions. Figure 7 reports mean performance on the first test trial in each relearning session as a function of initial criterion, collapsed across relearning criterion groups. Collapsing across relearning groups simplifies examination of the effects of initial criterion and is warranted on the basis of the outcomes of separate factorial ANOVAs (Initial Learning Criterion × Relearning Criterion) conducted for each session, which revealed no significant main effects of relearning or any interactions (all *F*s < 1). (Because participants completed different numbers of relearning sessions according to group assignment, the full mixed-factor ANOVA Initial Learning Criterion × Relearning Criterion × Sessions 2–6—would not be meaningful, because it would include only participants in the group that completed all five relearning sessions.)

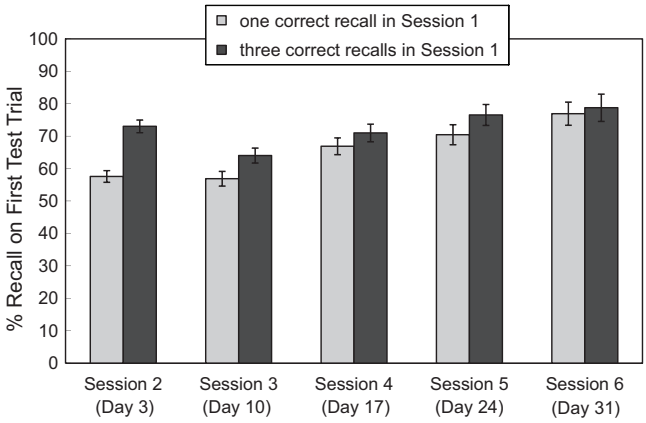

Figure 7. Experiment 3: Mean percentage recalled on the first test trial of each relearning session as a function of criterion level during Session 1. Error bars represent standard error of the mean.

As in Experiments 1 and 2, correctly recalling items three times versus one time during practice in Session 1 led to significantly greater recall performance on the first trial in Session 2, $t(299) = 5.83$, $p < .001$, $d = 0.67$. The issue of greater interest concerns the extent to which this effect persisted after subsequent relearning. To compare the effects of initial criterion prior to relearning versus after relearning, we conducted a 2 (initial criterion level) × 2 (first trial of Session 2 vs. first trial of last relearning session) ANOVA. For the latter variable, note that participants assigned to the group who completed only one relearning session were excluded from this analysis, given that the first trial in Session 2 was also the first trial of their last relearning session. Across participants in the remaining groups, the interaction was significant, $F(1, 234) = 30.25$, $MSE = 107.85$, $p < .001$, $\eta_p^2 = .11$, indicating that the effect of initial criterion level was weaker after relearning than prior to relearning. As in Experiment 2, the effect of Session 1

[4] Although reports of prior encounter did not differ significantly as a function of initial learning criterion or relearning criterion, participants unexpectedly reported having encountered more of the practiced items than baseline control items at the 1-month test (55.3% vs. 43.0%), $t(282) = 4.25$, $p < .001$. Given that we counterbalanced assignment of item sets to practice versus baseline control conditions, it is implausible that students actually encountered practiced items more frequently in class. Rather, we assume that the inflated estimates of encounter for the practiced items reflect source-monitoring error. The inflation tended to be greater with an initial learning criterion of one (61% vs. 42%) versus three (49% vs. 44%), $F(1, 273) = 5.85$, $p = .016$. No other effects or interactions were significant. A similar pattern of inflation was observed for the reports of prior encounter at the 4-month test, with participants again reporting having encountered more of the practiced items than baseline control items (65.5% vs. 50.5%), $t(262) = 6.38$, $p < .001$. No effects of initial learning criterion or relearning criterion or interactions were significant. However, report of prior encounter differed as a function of whether items had been tested in the 1-month test session (61.3% vs. 54.8%), $t(262) = 7.40$, $p < .001$. Given that we counterbalanced assignment of item sets to be tested in the 1-month session, this pattern further suggests that estimates of encounter in part reflect source-monitoring error.

criterion level was still significant on the first trial in Session 3, $t(238) = 2.21$, $p = .014$, $d = 0.29$. However, the numerical trends in subsequent sessions did not reach significance: Session 4: $t(175) = 1.10$, $p = .137$, $d = 0.16$; Session 5: $t(116) = 1.37$, $p = .089$, $d = 0.25$; Session 6: $t(60) = 0.33$, $p = .373$, $d = 0.08$. Thus, the strong effects of initial criterion prior to relearning were once again attenuated after relearning.

Effects of initial learning criterion on relearning rate in Sessions 2–6. As in Experiment 1, our converging measure of memory was relearning rate. For each participant, we computed the mean number of RMF trials per item needed to reach one correct recall for each item in each relearning session. Separate factorial ANOVAs (Initial Learning Criterion × Number of Relearning Sessions) for each session revealed no significant main effects of relearning or any interactions (all Fs < 2.72). Thus, mean RMF trials to reach criterion in each relearning session is shown in Figure 8 as a function of Session 1 criterion level, collapsed across relearning groups.

Replicating Experiments 1–2, fewer trials were needed to reach criterion in Session 2 when items had been correctly recalled three times versus only once during Session 1, $t(299) = 3.98$, $p < .001$, $d = 0.46$. Mirroring the pattern of recall performance, this effect diminished after subsequent relearning. A 2 (initial criterion level) × 2 (number of trials in Session 2 vs. number of trials in the last relearning session) ANOVA yielded a significant interaction, $F(1, 234) = 17.05$, $MSE = 0.30$, $p < .001$, $\eta_p^2 = .07$. Although the effect of initial criterion on relearning rate persisted in Session 3, $t(238) = 1.98$, $p = .025$, $d = 0.26$, the trends were not significant in subsequent sessions: Session 4: $t(175) = 1.11$, $p = .134$, $d = 0.17$; Session 5: $t(116) = 0.86$, $p = .195$, $d = 0.16$; Session 6: $t(60) = 0.61$, $p = .274$, $d = 0.15$.

Effects of initial learning criterion and number of relearning sessions on cued recall performance in the 1-month and 4-month test sessions. For the 1-month test session, mean percentage recall on the first test trial for practiced items is presented in Figure 9. A 2 (one or three correct recalls in Session 1) × 5 (number of relearning sessions) factorial ANOVA revealed a significant main effect of initial learning criterion, $F(1, 273) = 4.81$, $MSE = 452.29$, $p = .029$, $\eta_p^2 = .02$, and a significant main effect of relearning criterion, $F(4, 273) = 16.77$, $p < .001$, $\eta_p^2 = .20$, with no interaction ($F < 1$).

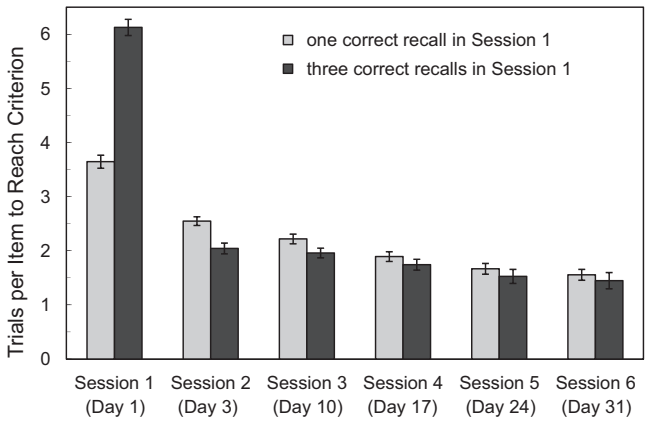

Figure 8. Experiment 3: Mean number of retrieval-monitoring-feedback trials per item needed to reach criterion during initial learning (Session 1) and during relearning (Sessions 2–6) as a function of criterion level during Session 1. Criterion level during Session 1 was one or three correct recalls; criterion in Sessions 2–6 was one correct recall. Error bars represent standard error of the mean.

A parallel ANOVA for baseline control items unexpectedly revealed a main effect of initial learning criterion, $F(1, 273) = 16.09$, $MSE = 108.12$, $p < .001$, $\eta_p^2 = .06$. Given that the main effect of relearning and the interaction were not significant (Fs < 1), performance for the baseline control items reported in Figure 9 was collapsed across relearning group. Concerning the significant effect of initial learning criterion, enhanced performance for the baseline items in the group with a higher initial learning criterion might reflect a form of retrieval-induced facilitation. Chan and colleagues (see e.g., Chan, 2009; Chan, McDermott, & Roediger, 2006) have recently shown that retrieval practice for a subset of information contained in a text can facilitate performance for related but untested information. By analogy, perhaps enhanced initial learning of items in the current experiment facilitated students' learning of some of the baseline concepts they encountered in class. Although this is an intriguing possibility, we adopt the more conservative interpretation that the effect of initial criterion on baseline performance reflected error due to some other factor rather than reflecting another form of systematic learning gains, and we controlled for this difference accordingly in analyses explained next.

Concerning the significant effect of initial learning criterion, recall for practiced items was greater when items had been recalled three times versus one time in Session 1 (54.8% vs. 48.4%, $d = 0.27$). However, as mentioned earlier, recall of the baseline control items also differed as a function of initial learning criterion. When recall for practiced items was reanalyzed with recall of baseline control items as a covariate, the main effect of initial learning criterion was no longer significant, $F(1, 272) = 2.04$, $MSE = 438.38$, $p = .155$, $\eta_p^2 = .01$. Further supporting the conclusion that the effects of initial criterion on retention were attenuated by subsequent relearning, a 2 (initial criterion level) × 2 (first trial of Session 2 vs. first trial of 1-month test session) ANOVA yielded a significant interaction, $F(1, 281) = 11.09$, $MSE = 254.43$, $p = .001$, $\eta_p^2 = .04$.

In contrast, the main effect of relearning on recall of practiced items remained significant after entering baseline performance as a covariate, $F(4, 272) = 17.25$, $p < .001$, $\eta_p^2 = .20$. Recall was significantly poorer with only one relearning session versus two, three, four, or five relearning sessions, $t(113) = 3.02$, $t(113) = 5.08$, $t(109) = 6.87$, $t(114) = 8.05$, and ds $= 0.56, 0.95, 1.30, 1.50$, respectively. Likewise, recall was significantly poorer with two relearning sessions than with three, four, or five, $t(112) = 1.98$, $t(108) = 3.38$, $t(113) = 4.37$, and ds $= 0.37, 0.65, 0.81$, respectively. In contrast, at higher levels of relearning, the differences in recall performance were less pronounced. Although recall was greater after relearning five versus three times, $t(113) = 2.14$, $p = .017$, $d = 0.40$, five was not significantly better than four, and four was not significantly better than three, $t(109) = 0.88$, $t(108) = 1.26$, ds $< .25$, respectively. Overall, the group means were better fit by the curvilinear logarithmic function than a linear function ($R^2 = .998$ vs. .949, respectively), confirming that the incremental benefit to recall performance diminished as the number of relearning sessions increased.

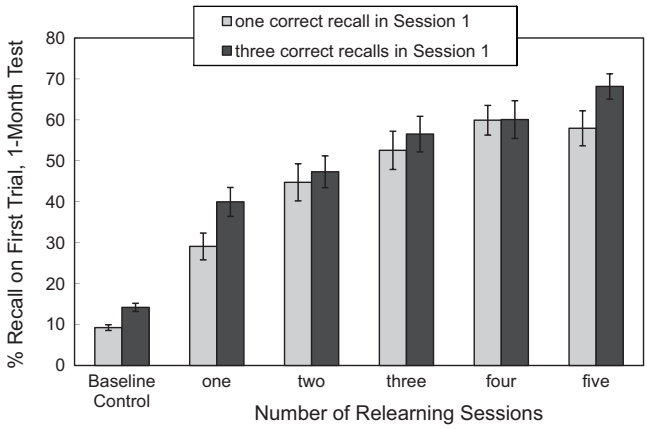

Figure 9. Experiment 3: Mean percentage recalled on the first test trial 1 month after the last practice session as a function of Session 1 criterion level (either one or three correct recalls) and as a function of the number of relearning sessions (practiced items were relearned one–five times; baseline control items were not practiced in any session). Error bars represent standard error of the mean.

The overall pattern of performance for the 4-month test session was similar. Mean percentage recall on the first test trial for practiced items is presented in Figure 10. A 2 (one or three correct recalls in Session 1) × 5 (number of relearning sessions) factorial ANOVA revealed a main effect of initial learning criterion, $F(1, 253) = 3.76$, $MSE = 365.16$, $p = .054$, $\eta_p^2 = .02$, and a significant main effect of relearning criterion, $F(4, 253) = 6.82$, $p < .001$, $\eta_p^2 = .10$, with no interaction ($F < 1$). As for the 1-month test session, a parallel ANOVA for baseline control items revealed a main effect of initial learning criterion, $F(1, 253) = 7.37$, $MSE = 78.58$, $p = .007$, $\eta_p^2 = .03$. Given that the main effect of relearning and the interaction were not significant (Fs < 1), performance for the baseline control items reported in Figure 10 was collapsed across relearning group.

Concerning the significant effect of initial learning criterion, recall for practiced items was greater when items had been recalled three times versus one time in Session 1 (38.3% vs. 33.4%). However, as mentioned earlier, recall of the baseline control items also differed as a function of initial learning criterion. When recall for practiced items was reanalyzed with recall of baseline control items as a covariate, the main effect of initial learning criterion was no longer significant, $F(1, 252) = 1.06$. Furthermore, a 2 (initial criterion level) × 2 (first trial of Session 2 vs. first trial of 4-month test session) ANOVA yielded a significant interaction, $F(1, 261) = 17.52$, $MSE = 212.79$, $p < .001$, $\eta_p^2 = .06$, again supporting the conclusion that the effects of initial criterion on retention diminished with subsequent relearning.

In contrast, the main effect of relearning on recall of practiced items remained significant after controlling for baseline performance, $F(4, 252) = 7.03$, $p < .001$, $\eta_p^2 = .10$. Recall was significantly poorer with only one relearning session versus two, three, four, or five relearning sessions, $t(102) = 1.86$, $t(104) = 2.87$, $t(104) = 4.79$, $t(104) = 5.06$, and ds = 0.36, 0.56, 0.93, 0.98, respectively. Recall was also significantly poorer with two relearning sessions than with four or five, $t(102) = 2.45$, $t(102) = 2.69$, and ds = 0.48, 0.53. In contrast, at higher levels of relearning, the differences in recall performance were less pronounced. Although recall was greater after relearning five versus three times, $t(104) = 1.69$, $p = .047$, $d = 0.33$, five was not significantly better than four, $t(104) = 0.26$, $d = 0.05$, and four was not significantly better than three, $t(104) = 1.45$, $p = .075$, $d = 0.28$. Again, the group means were better fit by a curvilinear function than a linear function ($R^2 = .99$ vs. .96).

To revisit, half of the items tested at the 4-month interval had also been tested and then relearned during the 1-month test session. Although we collapsed across this variable in the previous analyses for the sake of simplicity, post hoc analyses revealed a significant effect of tested versus not tested at 1 month (40.3% vs. 31.3%, respectively), $F(1, 253) = 108.19$, $p < .001$, $\eta_p^2 = .30$ (this variable did not interact with initial criterion or relearning). Of course, this effect must be interpreted with caution given that items relearned at 1 month functionally had a 3- versus 4-month retention interval. Nonetheless, the results suggest that diminishing returns of increasing relearning sessions may be overcome by expanding the interval between later sessions.

Effects of initial learning criterion and number of relearning sessions on relearning rate in the 1-month and 4-month test sessions. Mean number of RMF trials required to reach one correct recall for each practiced item after the 1-month cued recall test is reported in Figure 11. A 2 (one or three correct recalls in Session 1) × 5 (number of relearning sessions) factorial ANOVA revealed a significant main effect of relearning, $F(4, 273) = 12.12$, $MSE = 0.54$, $p < .001$, $\eta_p^2 = .15$, but no main effect of initial learning criterion and no interaction (Fs < 1). Thus, a higher initial learning criterion did not yield faster relearning in the 1-month test session. Further confirming that the effects of initial criterion prior to relearning significantly diminished after relearning, a 2 (initial criterion level) × 2 (number of trials in Session 2 vs. 1-month test) ANOVA yielded a significant interaction, $F(1, 281) = 14.10$, $MSE = 0.41$, $p < .001$, $\eta_p^2 = .05$.

In contrast, relearning in the 1-month test session was slower with only one prior relearning session versus two, three, four, or

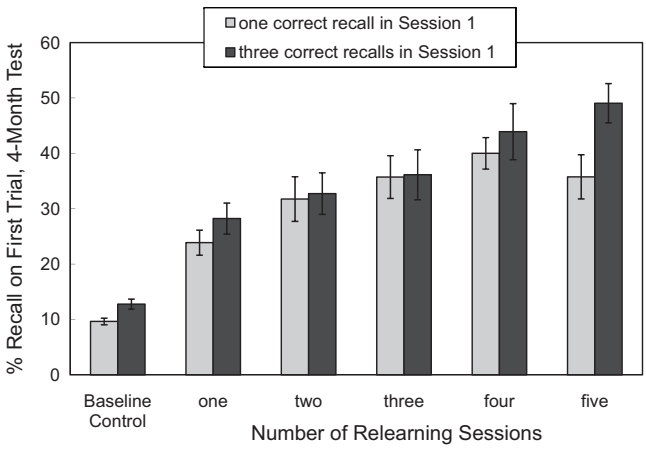

Figure 10. Mean percentage recalled on the first test trial 4 months after the last practice session as a function of Session 1 criterion level (either one or three correct recalls) and as a function of the number of relearning sessions (practiced items were relearned one–five times; baseline control items were not practiced in any session). Error bars represent standard error of the mean.

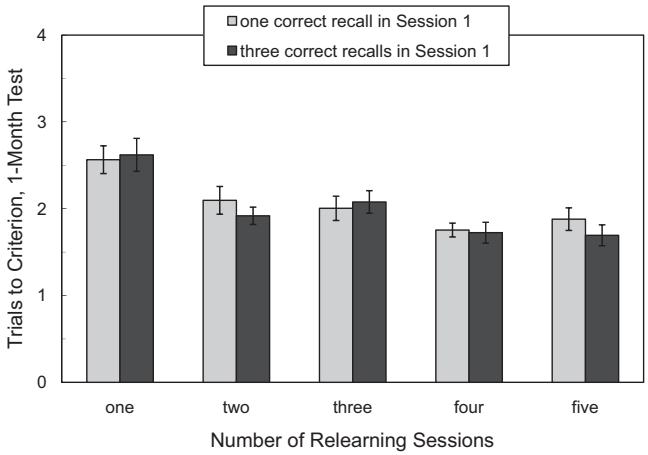

Figure 11. Experiment 3: Mean number of trials per item needed to reach a criterion of one correct recall per item during the 1-month test session. Error bars represent standard error of the mean.

five prior relearning sessions, $t(113) = 3.70$, $t(113) = 3.51$, $t(109) = 5.91$, $t(114) = 5.33$, and ds = 0.69, 0.66, 1.12, 0.99, respectively. Although relearning trials did not differ with two versus three prior relearning sessions, $t(112) = 0.19$, relearning was slower with two or three prior relearning sessions versus four or five relearning sessions (all ts > 1.78, dfs ≥ 108, ds = 0.33–0.48). Relearning trials did not differ for four versus five prior relearning sessions, $t(109) = 0.34$, $d = 0.07$. Thus, with higher levels of prior relearning, the differences in relearning trials were less pronounced. Overall, the group means were better fit by a curvilinear function than a linear function ($R^2 = .90$ vs. .78, respectively), confirming that the incremental benefit to relearning rate diminished as the number of prior relearning sessions increased.

A similar pattern of results emerged for the 4-month test session. Mean number of RMF trials required to reach one correct recall for each practiced item after the 4-month cued recall test is reported in Figure 12. A 2 (one or three correct recalls in Session 1) × 5 (number of relearning sessions) factorial ANOVA revealed a significant main effect of relearning, $F(4, 247) = 8.10$, $MSE = 1.07$, $p < .001$, $\eta_p^2 = .12$, but no main effect of initial learning criterion and no interaction (Fs < 1.45). Thus, a higher initial learning criterion did not yield faster relearning in Session 10. Further confirming that the effects of initial criterion prior to relearning significantly diminished after relearning, a 2 (initial criterion level) × 2 (number of trials in Session 2 vs. 4-month test) ANOVA yielded a significant interaction, $F(1, 255) = 15.65$, $MSE = 0.46$, $p < .001$, $\eta_p^2 = .06$.

In contrast, relearning in the 4-month test session was slower with only one versus two, three, four, or five prior relearning sessions, $t(100) = 1.99$, $t(103) = 3.43$, $t(101) = 4.96$, $t(101) = 4.26$, and ds = 0.39, 0.67, 0.98, 0.84. Relearning was also slower with two versus four or five prior relearning sessions, ts(99) > 2.05, ds = 0.53 and 0.41, respectively. No other differences were significant (all ts < 1.58, ds = 0.12–0.31). Again, the group means were better fit by a curvilinear function than a linear function ($R^2 = .95$ vs. .84), confirming that at higher levels of prior relearning, the differences in relearning rate were less pronounced.

To revisit, half of the items tested at the 4-month interval had also been tested and then relearned during the 1-month test session. Although we collapsed across this variable in the previous analyses for the sake of simplicity, post hoc analyses revealed a significant effect of tested versus not tested at 1 month (2.38 vs. 2.72 relearning trials, respectively), $F(1, 247) = 60.55$, $p < .001$, $\eta_p^2 = .20$ (this variable did not interact with initial learning criterion or relearning). Once again, this effect must be interpreted with caution given that items relearned at 1 month functionally had a 3- versus 4-month retention interval. Nonetheless, the results suggest that the diminishing returns of increasing relearning sessions may be overcome by expanding the interval between later sessions.

Efficiency of learning for the 1-month test. To summarize the results thus far, we found that practicing until items were correctly recalled three times versus once during Session 1 produced initially strong effects on retention that then diminished as the number of subsequent relearning sessions increased. More practice trials were required to obtain the higher criterion in Session 1, as shown in Figure 8 (left set of bars), but this additional cost was partially offset by faster relearning in subsequent practice sessions. Concerning relearning, increasing the number of relearning sessions produced significant but diminishing improvements in retention, at the obvious expense of additional practice trials in those sessions.

To estimate the overall efficiency of these practice schedules, for each participant we computed the total number of RMF trials per item needed to reach criterion in all sessions prior to the 1-month test (i.e., the total number of trials per item across initial learning and all subsequent relearning sessions). Means across individual values are reported in Figure 13. (The value embedded in each bar represents mean recall on the first test trial in the 1-month test session, to facilitate evaluation of the optimality of each schedule.) A 2 (one or three correct recalls in Session 1) × 5 (number of relearning sessions) factorial ANOVA revealed significant main effects of initial learning criterion, $F(1, 278) = 6.37$, $MSE = 15.93$, $p = .012$, $\eta_p^2 = .02$, and of relearning, $F(4, 278) = 25.86$, $p < .001$, $\eta_p^2 = .27$, but no interaction ($F < 1.26$).

Thus, the cost of additional practice trials to reach a higher criterion in Session 1 was largely but not entirely recouped across

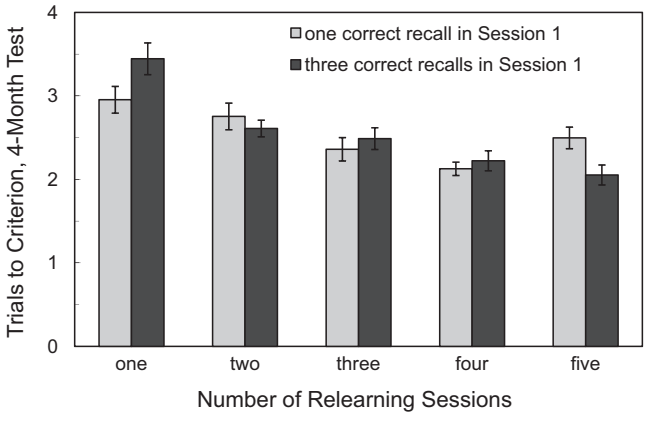

Figure 12. Experiment 3: Mean number of trials per item needed to reach a criterion of one correct recall per item during the 4-month test session. Error bars represent standard error of the mean.

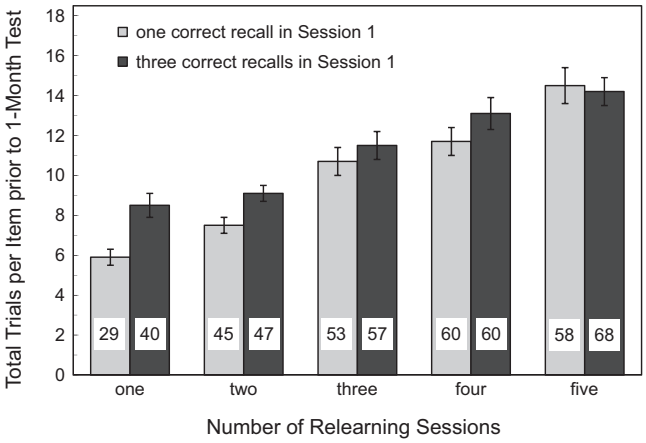

Figure 13. Mean number of retrieval-monitoring-feedback (RMF) trials per item across all practice sessions prior to the 1-month test as a function of practice schedule in Experiment 3. Trials per item is computed as the total number of RMF trials per item across sessions (i.e., including trials to criterion in Session 1 plus trials in subsequent relearning sessions). Values embedded in the bars are performance on the first test trial of the 1-month test (these are the same values reported in Figure 9, but they are included here to facilitate comparison of the optimality of each schedule). Error bars represent standard error of the mean.

subsequent relearning opportunities (three vs. one correct recalls in Session 1 resulted in approximately 11 vs. 10 total trials per item across practice sessions). However, a higher criterion in Session 1 also produced a modest improvement in performance on the 1-month test (although the differences in performance for baseline control items described earlier must be kept in mind). To consider the relative efficiency of a higher initial learning criterion using another metric, practicing to achieve two additional correct recalls in Session 1 amounted to 1.1 min of additional practice time per item across sessions (9.2 vs. 8.1 min total per item across sessions, respectively). In sum, the ultimate costs associated with a higher initial criterion were minimal, although the gains in final test performance were also minimal.

In contrast, greater cost was associated with increasing the number of relearning sessions (approximately 7, 8, 11, 12, and 14 total RMF trials per item, respectively) but with more substantial benefits to final test performance (34%, 46%, 55%, 60%, and 63%, respectively, on the 1-month cued recall test). To consider the relative efficiency of additional relearning sessions using other metrics, relearning items only once resulted in a total of 6.3 min of practice time per item. A second relearning session added only another 0.7 min of practice time per item but yielded a 35% relative improvement in performance. Three relearning sessions (vs. one) added only another 2.7 min of practice time per item but yielded a 62% relative improvement in performance. Four or five relearning sessions (vs. one) added 3.4 or 4.8 min, respectively, of additional practice time per item but with diminishing incremental benefit to retention.

Efficiency of learning for the 4-month test. To estimate overall efficiency of practice for the 4-month test, we again computed the total number of RMF trials per item needed to reach criterion in all sessions prior to the 4-month test. As a reminder, the items that were tested at 1 month were also relearned in the 1-month test session. Thus, the total number of trials for these items includes RMF trials from initial learning, all relearning sessions, and the 1 month-test session. For these items, means across individual values in each group are reported in Figure 14. The value embedded in each bar represents mean recall on the first test trial in the 4-month test session. Total trials for items that were not relearned in the 1-month test session are highly similar to those reported in Figure 13 (although not identical, because they include results from only participants who completed the 4-month test session); because they are largely redundant, we do not consider them here.

For items that were relearned in the 1-month test session, a 2 (one or three correct recalls in Session 1) $\times$ 5 (number of relearning sessions) factorial ANOVA revealed a significant main effect of initial learning criterion, $F(1, 244) = 4.86$, $MSE = 15.43$, $p = .029$, $\eta_p^2 = .02$, and a significant main effect of relearning, $F(4, 244) = 24.15$, $p < .001$, $\eta_p^2 = .28$, but no interaction ($F < 1.63$). Thus, the cost of additional practice trials to reach a higher criterion in Session 1 was still not entirely recouped across subsequent relearning opportunities (three vs. one correct recalls in Session 1 resulted in approximately 13 vs. 12 total trials per item across sessions). Although the ultimate costs associated with a higher initial criterion were minimal, the gains in final test performance were also minimal (42% vs. 38%, collapsing across relearning groups).

Greater cost was associated with increasing the number of relearning sessions (approximately 10, 10, 13, 14, and 16 total practice trials per item, collapsing across initial criterion) but with more substantial benefits to final test performance (30%, 36%, 40%, 47%, and 48%, respectively, on the 4-month cued recall test). Considering efficiency on other metrics, one relearning practice session along with relearning in the 1-month test session resulted

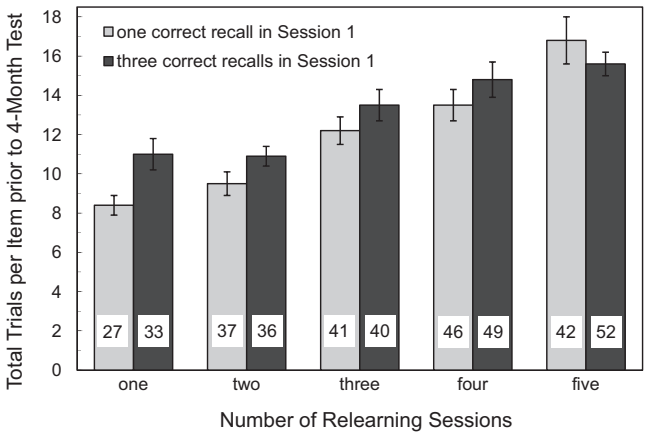

Figure 14. For items that were relearned during the 1-month test session, mean number of retrieval-monitoring-feedback (RMF) trials per item across all sessions prior to the 4-month test as a function of practice schedule in Experiment 3. Trials per item is computed as the total number of RMF trials per item across sessions (i.e., including trials to criterion in Session 1 plus trials in subsequent relearning sessions and during the 1-month test session). Values embedded in the bars are performance on the first test trial of the 4-month test. Error bars represent standard error of the mean.

in a total of 7.8 min of practice time per item. A second relearning practice session did not increase total time (7.8 min, because although additional time was spent in the second relearning session, it was offset by faster relearning after the 1 month test), but it yielded a 20% relative improvement in performance. A third relearning session added only another 2.1 min of practice time per item above a single relearning session but yielded a 33% relative improvement in performance. The fourth and fifth relearning sessions added 3.0 or 4.4 min, respectively, of total practice time per item above a single relearning session but yielded 57%–60% relative improvement.

General Discussion

The highest level goal of the current research was to investigate how to promote durable and efficient learning. To this end, we focused on retrieval practice, because a wealth of prior research has established its potency for improving memory. Going significantly beyond prior research, we examined the combined effects of both initial learning criterion and relearning to provide a seminal answer to the key question: How much retrieval practice is enough to achieve durable and efficient learning?

Qualitative Conclusions Concerning Initial Learning Criterion and Relearning

The current research supports several clear qualitative conclusions. Concerning initial learning criterion, increasing the criterion level during initial learning has strong effects on retention in the absence of relearning. However, the effects of initial learning criterion diminish markedly after relearning. Despite this fact, we would still advocate that learners practice to a higher initial criterion (in contrast to what students often spontaneously do, which is to terminate practice after one correct recall; Kornell & Bjork, 2007, 2008). This prescriptive conclusion is based on consideration of both durability and efficiency: If a learner does not engage in relearning, long-term retention will benefit from the higher initial criterion. Although reaching the higher criterion will come at the cost of significantly more practice trials during initial learning, the gains for subsequent memory are substantial enough to warrant the investment. If a learner does engage in relearning, the higher initial criterion will have increasingly minimal benefit to long-term retention, but it will also be associated with decreasing additional cost. That is, most or all of the cost of additional trials during initial practice is recouped by faster subsequent relearning. Thus, although the gains associated with a higher initial criterion may be modest, so too are the costs.

Concerning relearning, the basic qualitative conclusion is that more relearning is definitely better—increasing the number of relearning sessions had consistently substantial effects on long-term retention. However, the results of Experiment 3 suggest a qualification to this conclusion, in that increasing relearning had diminishing returns on long-term retention. That is, as the number of relearning sessions increased, the incremental benefit for long-term retention decreased. With respect to efficiency, the benefit of relearning for durability came at the cost of additional practice trials. However, at least for the first few relearning sessions, the magnitude of the gains arguably outweigh these costs.

Quantitative Conclusions: Picking a Winner

As outlined earlier, the existing literature on testing effects affords no answer to the question of how much practice is enough. One important goal of the current work was to take a significant first step in that direction by identifying the most effective schedule for promoting both durable and efficient learning. To that end, how might one go about picking a winner?

Certainly, the selection of a winner must be based on the degree of match to a learner's goals. Our assumption is that in most cases, the goal is to identify the most efficient schedule from among those that achieve a desired level of durability. Of course, one must consider both short-term goals (e.g., a student would like to earn at least a B on an upcoming exam) and long-term goals (e.g., retaining key concepts from an introductory course to facilitate cumulative learning in an advanced course in the following semester).

With both short-term and long-term goals in mind, our pick is the 3 + 3 schedule (i.e., practicing to three correct recalls during initial learning followed by three subsequent relearning sessions). For satisfying learners' short-term goals, consider the levels of performance observed on the interim tests in Experiments 2 and 3. Learners correctly recalled 80% of the content 2 days after a second relearning session (see Figure 4) and 77% 1 week after a third relearning session (see Figure 7). Assuming that students would likely schedule their last relearning session the night before an exam, and bearing in mind that many course exams involve recognition measures rather than recall, we are reasonably confident that the 3 + 3 schedule would satisfy most students' short-term retention goals. With respect to efficiency, the 3 + 3 schedule is preferable to schedules involving more relearning sessions (e.g., as is apparent in Figure 7, an additional relearning session increased recall by only 2%).

Concerning long-term retention goals, we emphasize the importance of considering both recall and relearning rate. Arguably, achieving 100% correct recall after lengthy delays is an unreasonable goal, and thus the need for some relearning is to be expected. As Nelson (1985) noted,

> Educators generally realize that most of what they teach their students does not remain recallable, or even recognizable, for extended periods of time. Rather the hope is that their teaching will have been beneficial because the student could relearn the information relatively quickly. (p. 478)

Accordingly, an effective schedule is one that most efficiently yields a reasonably high level of long-term retention along with rapid relearning for inaccessible content. Again, the 3 + 3 schedule effectively meets these goals. Correct recall was around 57% after 4–6 weeks (see Figures 6 and 9), and relearning of unrecallable items was relatively rapid. Furthermore, the 3 + 3 schedule required fewer total practice trials across sessions than did other schedules that yielded similar levels of performance.

Finally, although the descriptive conclusions that follow from the current findings are relatively clear, we realize that our prescriptive conclusions favoring the 3 + 3 schedule may not meet the full consensus of researchers in this field (or of researchers in related fields or students and educators more generally). However, the present research will have served its purpose if it engenders healthy debate and increases the attention of the larger research community to the importance of further investigating these issues for both theory and practice.

Future Directions: A New Agenda for Testing Effects Research

The current research represents the first examination of the combined effects of initial learning criterion and relearning. Further work will be needed to evaluate whether our prescriptive conclusions generalize to other kinds of material, other forms of criterion task, and to other populations of learners. Regarding materials, the optimal criterion levels for efficient and durable learning of simpler items (e.g., paired associates), more complex materials (e.g., lengthier text), or nonverbal material (e.g., diagrams) may differ from those identified here for key concepts. Regarding criterion task, Experiment 2 provided evidence that recognition-based measures (such as the multiple-choice questions we used here) may be less sensitive to effects of initial criterion and relearning. In contrast, the sizeable effects on recall suggest that similar effects may obtain in other productive tasks (e.g., application tasks). Regarding other learners, we are currently examining the effects of initial learning criterion and relearning with younger learners (middle-school students).

Another important direction involves examining the extent to which effects of the amount of practice (initial criterion and relearning) depend on the timing of practice. The two prior studies that manipulated relearning also included manipulations of the timing of those sessions. Bahrick (1979) administered two or five relearning sessions that were 0, 1, or 30 days apart, and relearning and spacing appeared to have additive effects. Bahrick et al. (1993) completed 13 or 26 relearning sessions that were 2, 4, or 8 weeks apart and also found additive effects of relearning and spacing. Both of these studies involved fixed intervals between sessions. An intriguing possibility suggested by secondary results of Experiment 3 is that expanding the interval between relearning sessions may improve the durability and/or efficiency of learning.

Future research exploring the underlying bases of these effects would also be useful, although the current results do have implications for theory. As mentioned earlier, the empirical literature on testing effects is substantial, but theoretical work on these effects has lagged behind considerably. The current results provide important empirical constraints as the theoretical accounts that have recently emerged from the literature continue to be developed and tested. For example, the subadditive effects of initial learning and relearning on recall performance are inconsistent with the superadditive pattern that would reasonably have been predicted by the retrieval effort hypothesis (outlined in the introduction). In contrast, the subadditive effects are consistent with the strategy shift hypothesis. A lower initial criterion led to more retrieval failures in a subsequent relearning session. According to the strategy shift hypothesis, retrieval failures allow learners to evaluate the effectiveness of encoding strategies and to shift to more effective encoding strategies, which would help overcome the initial disadvantage of lower criterion items in later sessions. Although the current experiments were not designed to directly test this account, one additional outcome of Experiment 3 is consistent with the more general idea that learners shift strategies with practice. Recall that learners practiced a total of 16 items, but items were separated into two sets with initial learning on two different days (see Table 4). Learners needed significantly fewer trials to reach criterion for the second set versus first set of items (5.3 vs. 4.8 trials per item), $t(300) = 3.91$, $p < .001$, which may reflect the use of more effective strategies in later learning sessions.

Taken together, the summary of the extant literature provided here and the novel outcomes of the current research support important general prescriptions for future research. Namely, we believe that the agenda for testing effects research should increasingly include (a) less reliance on single-session schedules involving fixed amounts of practice and greater attention to criterion learning in multiple sessions and (b) consideration of efficiency as a primary outcome, in addition to durability across meaningful retention intervals. Not only can work of this kind have significant theoretical implications, but it is arguably the critical next step to move the field beyond the plateau on which it currently rests with respect to real-world implications for learning.

References

Abbott, E. E. (1909). On the analysis of the factors of recall in the learning process. *Psychological Monographs, 11,* 159–177.

Atkinson, R. C., & Paulson, J. A. (1972). An approach to the psychology of instruction. *Psychological Bulletin, 78,* 49–61. doi:10.1037/h0033080

Bahrick, H. P. (1979). Maintenance of knowledge: Questions about memory we forgot to ask. *Journal of Experimental Psychology: General, 108,* 296–308. doi:10.1037/0096-3445.108.3.296

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

Bahrick, H. P., & Hall, L. K. (2005). The importance of retrieval failures to long-term retention: A metacognitive explanation of the spacing effect. *Journal of Memory and Language, 52,* 566–577. doi:10.1016/j.jml.2005.01.012

Berger, S. A., Hall, L. K., & Bahrick, H. P. (1999). Stabilizing access to marginal and submarginal knowledge. *Journal of Experimental Psychology: Applied, 5,* 438–447. doi:10.1037/1076-898X.5.4.438

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

Bjork, R. A. (1999). Assessing our own competence: Heuristics and illusions. In D. Gopher & A. Koriat (Eds.), *Attention and Performance XVII: Cognitive regulation of performance: Interaction of theory and application* (pp. 435–459). Cambridge, MA: MIT Press.

Chan, J. C. K. (2009). When does retrieval induce forgetting and when does it induce facilitation? Implications for retrieval inhibition, testing effect, and text processing. *Journal of Memory and Language, 61,* 153–170. doi:10.1016/j.jml.2009.04.004

Chan, J. C. K., McDermott, K. B., & Roediger, H. L., III (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General, 135,* 553–571. doi:10.1037/0096-3445.135.4.553

Darley, C. F., & Murdock, B. B., Jr. (1971). Effects of prior free recall testing on final recall and recognition. *Journal of Experimental Psychology, 91,* 66–73. doi:10.1037/h0031836

Dunlosky, J., Hartwig, M. K., Rawson, K. A., & Lipko, A. R. (2011). Improving college students' evaluation of text learning using idea-unit standards. *Quarterly Journal of Experimental Psychology, 64,* 467–484. doi:10.1080/17470218.2010.502239

Glenberg, A. M., & Lehmann, T. S. (1980). Spacing repetitions over one week. *Memory & Cognition, 8,* 528–538. doi:10.3758/BF03213772

Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology, 81,* 392–399. doi:10.1037/0022-0663.81.3.392

Glover, J. A., Krug, D., Hannon, S., & Shine, A. (1990). The "testing" effect and restricted retrieval rehearsal. *Psychological Record, 40,* 215–226.

Hambrick, D. Z. (2003). Why are some people more knowledgeable than others? A longitudinal study of knowledge acquisition. *Memory & Cognition, 31,* 902–917. doi:10.3758/BF03196444

Karpicke, J. D. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology: General, 138,* 469–486. doi:10.1037/a0017341

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.

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review, 14,* 219–224. doi:10.3758/BF03194055

Kornell, N., & Bjork, R. A. (2008). Optimising self-regulated study: The benefits—and costs—of dropping flashcards. *Memory, 16,* 125–136. doi:10.1080/09658210701763899

Lipko, A. R., Dunlosky, J., Hartwig, M. K., Rawson, K. A., Swan, K., & Cook, D. (2009). Using standards to improve middle-school students' accuracy at evaluating the quality of their recall. *Journal of Experimental Psychology: Applied, 15,* 307–318. doi:10.1037/a0017599

Nelson, T. O. (1985). Ebbinghaus's contribution to the measurement of retention: Savings during relearning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11,* 472–479. doi:10.1037/0278-7393.11.3.472

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. doi:10.1037/0278-7393.8.4.279

Pyc, M. A., & Rawson, K. A. (2007). Examining the efficiency of schedules of distributed retrieval practice. *Memory & Cognition, 35,* 1917–1927. doi:10.3758/BF03192925

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. (2010, October 15). Why testing improves memory: Mediator effectiveness hypothesis. *Science, 330,* 335. doi:10.1126/science.1191465

Pyc, M. A., & Rawson, K. A. (2011). Costs and benefits of dropout schedules of test-restudy practice: Implications for student learning. *Applied Cognitive Psychology, 25,* 87–95.

Raffel, G. (1934). The effect of recall on forgetting. *Journal of Experimental Psychology, 17,* 828–838. doi:10.1037/h0075297

Rawson, K. A., & Dunlosky, J. (in press). Retrieval-monitoring-feedback (RMF) technique for producing efficient and durable student learning. To appear in R. Azevedo & V. Aleven (Eds.), *International handbook of metacognition and learning technologies.*

Rawson, K. A., & Kintsch, W. (2005). Rereading effects depend upon time of test. *Journal of Educational Psychology, 97,* 70–80. doi:10.1037/0022-0663.97.1.70

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.

Roediger, H. L., III., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1,* 181–210. doi:10.1111/j.1745-6916.2006.00012.x

Rose, R. J. (1992). Degree of learning, interpolated tests, and rate of forgetting. *Memory & Cognition, 20,* 621–632. doi:10.3758/BF03202712

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science, 3,* 207–217. doi:10.1111/j.1467-9280.1992.tb00029.x

Sones, A. M., & Stroud, J. B. (1940). Review, with special reference to temporal position. *Journal of Educational Psychology, 31,* 665–676. doi:10.1037/h0054178

Spitzer, H. F. (1939). Studies in retention. *Journal of Educational Psychology, 30,* 641–656. doi:10.1037/h0063404

Received October 19, 2010
Revision received March 22, 2011
Accepted March 24, 2011 ■