Educ Psychol Rev (2013) 25:523–548
DOI 10.1007/s10648-013-9240-4

REPLICATION

The Power of Successive Relearning: Improving Performance on Course Exams and Long-Term Retention

Katherine A. Rawson · John Dunlosky · Sharon M. Sciartelli

Published online: 24 September 2013

Abstract Practice tests and spaced study are both highly potent for enhancing learning and memory. Combining these two methods under the conditions in which they are most effective (i.e., practice tests that invoke successful retrieval from long-term memory and spacing study across days) yields a promising learning technique referred to as *successive relearning*. Given the obvious implications of successive relearning for promoting student learning and the voluminous literatures on testing and spacing more generally, surprisingly few studies have evaluated successive relearning, and none have done so in an authentic educational context. The two experiments reported here establish the potency of a successive relearning intervention for enhancing student learning by demonstrating meaningful improvements in course exam performance and on long-term retention tests.

Keywords Successive relearning · Testing effect · Spacing effect · Student learning · Retention

Testing effects and spacing effects are two of the oldest and most robust effects in cognitive psychology, with more than a century of research establishing the potency of each of these techniques for enhancing learning and memory (for reviews, see Cepeda et al. 2006; Dunlosky et al. 2013; Roediger and Karpicke 2006; Roediger et al. 2011). Both literatures have also revealed conditions in which each strategy is particularly efficacious (Dunlosky et al. 2013). Of particular interest here, spacing has larger and longer-lasting effects on retention when practice trials for a given item are distributed across days versus within a session (e.g., Cepeda et al. 2006; Kornell 2009). Concerning testing effects, practice test formats that evoke recall of target information from long-term memory (i.e., *retrieval practice*) versus target recognition are particularly effective for enhancing retention (Carpenter and DeLosh 2006; Glover 1989). Furthermore, retrieval practice is most

This research was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant #R305A080316 to Kent State University. Opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

K. A. Rawson · J. Dunlosky · S. M. Sciartelli
Kent State University, Kent, OH, USA

K. A. Rawson (✉)
Department of Psychology, Kent State University, P.O. Box 5190, Kent, OH 44242-0001, USA
e-mail: krawson1@kent.edu

effective when retrieval attempts are successful, with long-term retention increasing as the number of successful retrievals increases (e.g., Pyc and Rawson 2009; Vaughn and Rawson 2011).

Taken together, these outcomes suggest that testing and spacing will be most effective for enhancing long-term learning when practice involves *multiple successful retrievals* that are *distributed across days*. Bahrick (1979) has recommended this method, which he referred to as *successive relearning*. Successive relearning involves alternating retrieval practice with restudy opportunities during initial learning until each item is recalled correctly, followed by additional retrieval practice with restudy on one or more subsequent days until each item is again successfully recalled. Thus, successive relearning involves precisely those conditions in which testing and spacing are most effective. Bahrick (1979) reported compelling outcomes concerning the potency of successive relearning for long-term retention. During initial learning, participants practiced English–Spanish word pairs until each item had been correctly recalled once. Participants then relearned items (i.e., engaged in additional retrieval practice with restudy until each item was correctly recalled once) in either two or five subsequent sessions. Levels of performance on a final test 30 days after the last relearning session were impressive (56 % after two relearning sessions and 83 % after five relearning sessions).

Given the implications of successive relearning for promoting student learning and the voluminous literatures on testing and spacing more generally, it is surprising that only five other studies have examined successive relearning since Bahrick's seminal work in 1979 (Bahrick et al. 1993; Bahrick and Hall 2005; Pyc and Rawson 2011; Rawson and Dunlosky 2011, 2012a). Outcomes from two recent studies are particularly promising with respect to the possible educational benefits of successive relearning (Rawson and Dunlosky 2011, 2012a). In both studies, undergraduates who successively relearned key concept definitions showed relatively high levels of retention on cued recall tests administered in the lab 1 and 4 months after relearning (up to 68 % and 49 % respectively, depending on the particular schedule of relearning, as compared with around 11 % in a baseline control condition).

Although outcomes of these studies are promising, further research exploring the effects of successive relearning are clearly needed. Perhaps of greatest interest, none of the six prior studies of successive relearning involved authentic educational contexts. Some studies did involve representative educational material (foreign language vocabulary or key concept definitions from textbooks), but no prior research has involved a classroom study in which students learned actual course content and course outcome measures were examined. Furthermore, some of the prior studies involved timescales that are implausible for instantiation in authentic educational contexts (e.g., in Bahrick et al. 1993, successive relearning sessions were spread over as many as 5 years, and final retention tests were administered 1–5 years after relearning).

Overview of the Current Research

The current work significantly advances the literature on successive relearning by evaluating the potency of a successive relearning intervention in an authentic educational context, which is important given the paucity of research and the imperfect correspondence between the size (and sometimes even direction) of effects observed in laboratory versus field studies (Mitchell 2012). The primary goal was to examine the extent to which successive relearning enhances performance on authentic outcome measures, as compared with business as usual. To this end, we report two experiments involving students enrolled in Introductory Psychology who engaged in successive relearning via a virtual flashcard program to learn key concept definitions from actual content in their course. Practice sessions were scheduled to align with the course schedule, beginning when the instructor introduced the content in class and ending prior to the course exam over that content.

Primary outcome measures of interest include course exam performance for the target concepts. Importantly, for each student, half of the target concepts drawn from the course content were successively relearned in the context of the experiment whereas the other half were not. Performance for the latter concepts provides a baseline for how well students learned course concepts by themselves. The key prediction is that successive relearning will significantly improve performance over the business-as-usual baseline.

Of secondary interest, we also examined performance in two other comparison conditions. Experiment 1 included a restudy condition in which concepts were presented for restudy but with no retrieval practice, to demonstrate that retrieval practice is an important active ingredient in successive relearning (i.e., spaced study alone without testing is not as potent). Restudy is also the most common comparison condition included in basic research on testing effects (Dunlosky et al. 2013) and is a study strategy that most students report using in survey research on students' study behavior (e.g., Kornell and Bjork 2007; Karpicke et al. 2009; Hartwig and Dunlosky 2012).

Both experiments also included a self-regulated practice condition in which students controlled the amount and timing of retrieval practice and restudy in each session, rather than successive relearning being controlled by the program. Inclusion of this condition is of interest for exploring the extent to which students' self-regulated implementation of spaced testing can also enhance learning. In particular, many virtual flashcard programs are available for students to engage in spaced testing (e.g., iKnow, CueCard, StudyBlue, SuperMemo). Although these programs provide support for retrieval practice and restudy, they typically provide little to no guidance for how much to practice or when. Thus, we were interested in students' performance after using a program that permits retrieval practice but does not regulate the schedule of practice to ensure successive relearning.

Finally, a perennial concern of educators is that students use study strategies and study schedules that are just good enough to pass an exam (e.g., cramming the night before an exam involving mostly passive restudy) but then rapidly forget most of what they have learned after the exam. Thus, to evaluate longer-term retention, students completed final cued recall tests at two time points after the course exam in both experiments. To foreshadow, evidence from the cued recall tests not only suggests that this perennial concern is warranted but also establishes that successive relearning can protect against these losses.

Experiment 1

Methods

Participants and Design Participants included 79 undergraduates enrolled in large sections of Introductory Psychology taught by the third author. To minimize issues with coercion, the third author was not involved in recruitment of participants and was blind concerning which students participated in the study. Of the 59 students who provided demographic data, 76 % were female, 72 % were White, and mean age was 18.9 years (range, 17–35 years).

Functionally, the design included four conditions: successive relearning, self-regulated practice, restudy only, and baseline control. All participants experienced three of the four conditions, including successive relearning, baseline control, and one of the remaining two practice conditions (self-regulated practice or restudy only). Participants were randomly assigned to one of four groups, defined by the practice condition that students completed in the first and second halves of the experiment (see Table 1). The primary purpose of the complex design was to allow all students to experience successive relearning for one of the

Table 1 Summary of practice conditions involved in each group in the first and second half of Experiment 1

	n	First half	Second half
Group 1	20	Successive relearning	Self-regulated practice
Group 2	21	Successive relearning	Restudy only
Group 3	18	Self-regulated practice	Successive relearning
Group 4	20	Restudy only	Successive relearning

In each group, half of the concepts were practiced in the condition indicated in the table, and the other half of the concepts were not practiced in the context of the experiment (baseline control). See text for further details

two course exams, given uncertainty about whether self-regulated practice and restudy-only would produce any enhancement in exam performance. (To foreshadow, we simplified the design in "Experiment 2").

Materials Materials included 64 concepts, with eight concepts from each of eight units that the instructor covered in class. Each set included the concepts and definitions from the instructor's course materials that students were expected to learn for class. Table 2 includes a sample concept from the instructor's materials for each of the eight units.

To revisit, our primary question of interest concerned the extent to which successive relearning enhances performance as compared with business as usual. To this end, the eight unit sets were divided into four pairs of units that were introduced in class at similar times in the semester. In order of introduction within the class, the four pairs of units included classical and operant conditioning, encoding and retrieval from memory, thinking and intelligence, and cognitive and emotional development. Within each pair, the concept set from one unit was assigned to a practice condition, and the other was assigned to the baseline control condition, with assignment of concept set to condition counterbalanced across participants. Baseline control concepts were not practiced in the context of the experiment but were tested on the outcome measures, because students were expected to learn all of the concepts in the context of the actual course. Thus, performance for these concepts provides a baseline for how well students learned course concepts via "business as usual."

Table 2 Sample concepts from students' actual course materials

Higher-order conditioning: a strong conditioned stimulus is paired with a new neutral stimulus and the new one also then becomes a conditioned stimulus
Partial reinforcement effect: responses that are reinforced sometimes but not every time tend to be very resistant to extinction
Elaborative rehearsal: making information meaningful, which is easiest way to get information from short-term memory into long-term memory
State-dependent learning: memories formed during a particular physiological or psychological state will be easier to recall while in a similar state
Confirmation bias: tendency to search for or pay attention to information that supports what we already think, and to ignore or distort information that contradicts our beliefs
Intelligence quotient (IQ): a measure of how well you did compared with others your age on that particular test of knowledge and problem solving
Zone of proximal development (ZPD): the difference between what a child can do alone and what child can do with the help of a teacher, includes a child's emerging skills
Temperament: behavioral characteristics and emotional reaction patterns that appear in infancy, precursor to personality

One of the key outcome measures to be reported below concerns performance on course exams, and thus, we briefly describe these exam materials here. Across the exams used as outcome measures (i.e., Exam 2 and Exam 3 in the course), an average of 39 % of the questions directly tapped the concepts that were included in the experimental materials (22 of 51 for Exam 2 and 18 of 51 for Exam 3), with a similar number of questions tapping each concept set. Exam 2 covered classical and operant conditioning and encoding and retrieval from memory; Exam 3 covered thinking, intelligence, and cognitive and emotional development. All questions were multiple-choice format. (To foreshadow, further descriptive information about the level of learning tapped by the exam questions and exploratory analyses of this factor are reported after "Experiment 2" below.)

Procedure Students in all groups completed 14 sessions. A schematic of the schedule of practice and test sessions for each group is summarized in Table 3. The practice schedule was aligned with the course schedule, such that practice sessions for a given set of concepts began when the instructor introduced that content in class and ended prior to the course exam over that content. The particular practice condition that each participant completed within each practice session depended on group assignment. In each session, participants worked individually at computer carrels. All instructions and tasks were administered by computer.

Session 1 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. Participants were explicitly told that they would be asked to learn key concept definitions provided by their Introductory Psychology instructor from their actual course materials. Participants were encouraged to do their best to learn the concepts so that they would do well on memory tests that would be administered in later sessions and because it might help them do better on their actual course exams.

Each participant then received brief instructions about the tasks they would be completing that day, appropriate to their group assignment. Participants who were assigned to begin in the successive relearning condition (Groups 1 and 2) were told that they would be using three different tasks (studying, recall practice, and monitoring learning) and were given brief instructions for each task, including a sample item illustrating the format of the monitoring task. Participants were told that they would practice recalling items until they had recalled each one three times and that the computer would be scoring the accuracy of their responses (although in actuality, online recall accuracy was based on the participant's monitoring judgments, as described below). Participants assigned to begin with self-regulated practice (Group 3) were told that they would be given the option to use three different tasks (studying, recall practice, and monitoring learning) and were given brief instructions for each task, including an illustration of the monitoring task. Participants assigned to begin with restudy only (Group 4) were told that they would be able to study the concepts as many times as they wanted. Participants in all groups were told they would have up to 60 min for learning the concepts in the first session.

All participants were then presented with each of the eight concepts from the unit of the first pair that was assigned to the practice condition, one at a time for self-paced study. After the initial study trial for each concept, the procedure for the remainder of the session differed according to group. Participants in Groups 1 and 2 were then presented with the concepts one at a time for a self-paced *retrieval–monitoring–feedback* (RMF) trial. The sequence of tasks included in each RMF trial is shown in Fig. 1. Each RMF trial began with retrieval practice (top panel of Fig. 1), in which a concept was presented along with a text field and the participant was prompted to type in the definition. When participants clicked a button to

Table 3 Schematic of practice and test schedules for each group in Experiment 1

Session:	1	2	3	4	5	6		7	8	9	10	11	12		13	14
Day:	1	3	8	10	15	17	19	22	24	29	31	36	38	40	43	64
Unit:	A	A	AC	AC	C	C	A-D	A-D, E	E	EG	EG	G	G	E-H	A-D, E-H	E-H
Group 1:	SR	SR	SR	SR	SR	SR	Exam	CR1, Self	Self	Self	Self	Self	Self	Exam	CR2, CR1	CR2
Group 2:	SR	SR	SR	SR	SR	SR	Exam	CR1, RO	RO	RO	RO	RO	RO	Exam	CR2, CR1	CR2
Group 3:	Self	Self	Self	Self	Self	Self	Exam	CR1, SR	SR	SR	SR	SR	SR	Exam	CR2, CR1	CR2
Group 4:	RO	RO	RO	RO	RO	RO	Exam	CR1, SR	SR	SR	SR	SR	SR	Exam	CR2, CR1	CR2

Session=Numbered sessions in which students completed one or more tasks in the laboratory. Day=Calendar day on which a task took place (to show the interval between sessions). Days are numbered relative to the first session of the experiment. Unit=Materials included four pairs of units covered in the class, and A-H are arbitrary labels used to refer to the eight units. The unit labels in this row indicate which units were involved in the tasks completed on that day. Exam=A course exam was administered in class that day. The course exam included questions that tapped both practiced concepts and baseline control concepts. CR1=The first cued recall test for a given set of items was administered in the lab (approximately 3 days after course exam). Both practiced concepts and baseline control concepts were tested. CR2=The second cued recall test for a given set of items was administered in the lab (approximately 24 days after course exam). Both practiced concepts and baseline control concepts were tested. SR=Students completed the successive relearning procedure for items assigned to the practice condition. Self=Students completed the self-regulated practice procedure for items assigned to the practice condition. RO=Students completed the restudy-only procedure for items assigned to the practice condition. For example, in Session 4 on Day 10, students practiced concepts from Units A and C (via successive relearning in Groups 1–2, self-regulated practice in Group 3, and restudy only in Group 4). As another example, in Session 7 on Day 22, all students completed the first cued recall test for Units A, B, C, and D, and then practiced concepts from Unit E (via self-regulated practice in Group 1, restudy only in Group 2, or successive relearning in Groups 3–4)

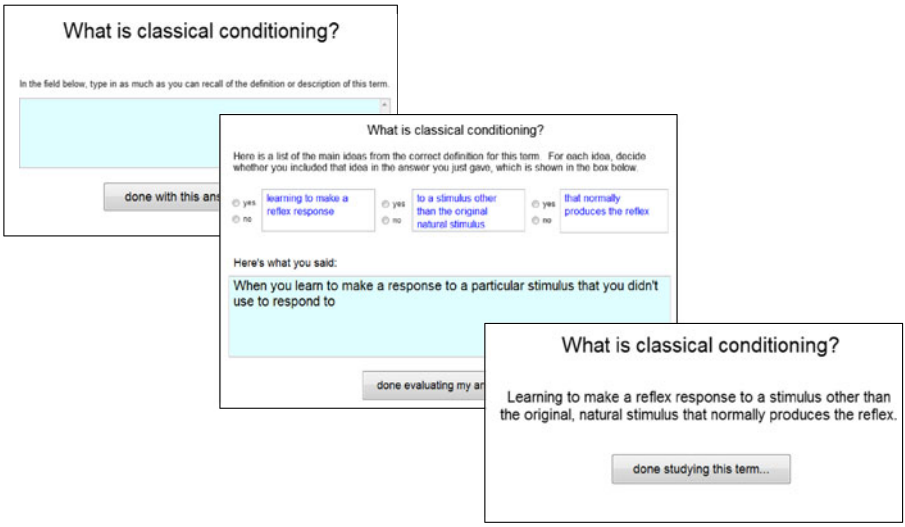

Fig. 1 Illustration of the sequence of tasks included in each retrieval–monitoring–feedback (RMF) trial in the successive relearning condition. The *top panel* illustrates the retrieval practice task; the *middle panel* illustrates the monitoring task, and the *bottom panel* illustrates the feedback task. See text for further procedural details

indicate they were done with recall, they were prompted to make a monitoring judgment. As illustrated in the middle panel of Fig. 1, the concept 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 a participant 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, as shown in the bottom panel of Fig. 1, a feedback screen presented the key term and intact definition for self-paced restudy.

The program tracked how many times the participant judged that the response for a given concept was completely correct (i.e., the participant judged that all of the idea units were contained in their response).[1] On a given RMF trial, if the participant did not judge the recall response as completely correct, that item was placed at the end of the list for another RMF trial later. If the participant judged the recall response as completely correct but it was not the third time, the item was placed at the end of the list for another RMF trial later. Once an item was judged as completely correct three times, practice for that item was discontinued for the remainder of the session. The practice phase in Session 1 ended once all items had reached the assigned criterion level (or once 60 min had elapsed).

[1] Based on our post-experiment scoring of participants' responses, the accuracy of the responses that participants judged as completely correct across all RMF trials in all sessions was 89 % (SE=2). The accuracy of responses participants judged as completely correct in "Experiment 2" was similar (M=80 %, SE=2).

After initial study, participants in Group 3 began the self-regulated practice phase in which they were shown a menu of the eight concepts along with three task options for each (see Fig. 2). When participants elected to study an item, the concept was shown on a separate screen along with the definition and a button to click when they were done studying to return to the menu. When participants elected to self-test for an item, the concept was presented on a separate screen along with a text field in which the participant was prompted to type the definition and a button to click when they were done to return to the menu. When participants elected to judge an item, they were shown a separate screen like that shown in Fig. 1 that included their most recent recall response for that item (clicking the "done evaluating" button returned them to the menu). For cases in which participants elected to judge an item that they had not yet self-tested, they were told, "You have not yet tested yourself on this item, so you do not currently have a response to evaluate." Participants were given up to 60 min to make as many choices as they wanted for as many of the concepts as they wanted. Participants could also terminate practice in Session 1 prior to 60 min by clicking a button below the menu.

After the first block of initial study trials, participants in Group 4 were presented with each of the concepts and their definitions for five blocks of self-paced restudy trials. To briefly explain the logic for this procedure, outcomes from our prior research supported the expectation that students in the successive relearning condition would require about five RMF trials per item on average to achieve a criterion of three correct recalls. Accordingly, the intent of presenting concepts for five restudy trials in the restudy-only condition was to increase the likelihood that participants in the restudy-only condition had at least as many trials as in the successive relearning condition. However, we did not want to unfairly disadvantage students in the restudy-only condition if they did not believe they had adequately learned the concepts at that point. Thus, after the fifth restudy trial for each

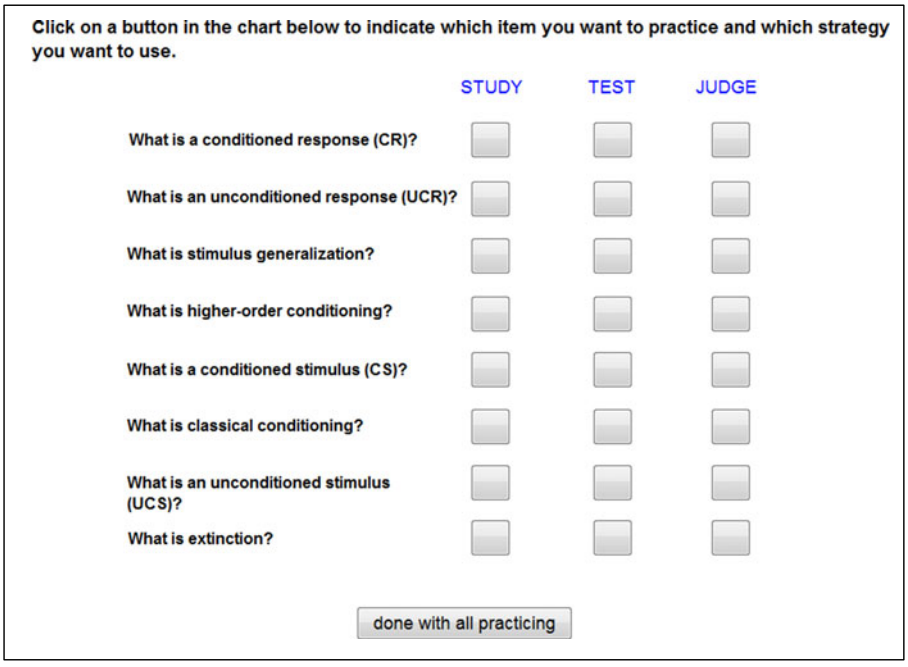

Fig. 2 Screen on which students in the self-regulated practice condition were able to select tasks to perform for each concept

concept, participants indicated whether they wanted to discontinue study of that concept for Session 1 or if they wanted to restudy it again later. If they elected to restudy it again later, it was placed at the end of the list for another restudy trial; otherwise, the concept was removed from the list for the remainder of Session 1. The practice phase in Session 1 ended once all items had been dropped from the list (or after 60 min). On average, participants completed Session 1 in 35 min (SE=2).

In all four groups, at the end of the practice phase, participants made a judgment of learning for each concept, in which they used a 0–100 scale to rate the likelihood that they would be able to remember the definition of the concept on a later memory test. Because these judgments are not relevant to the primary question of interest in the present report, we do not discuss them further.

Session 2 Participants returned 2 days later for Session 2 (see Table 3). The procedure for Groups 1 and 2 was the same as Session 1, except that initial study was omitted and participants instead began with RMF practice trials. Additionally, practice with a given concept was discontinued after one correct recall response (based on the participant's monitoring judgments, as in Session 1). The procedure for Group 3 was the same as in Session 1 except that the initial study was omitted and participants instead were taken directly to the SRL menu. The procedure for Group 4 was the same as in Session 1 except that initial study was omitted, and participants were given the option to discontinue restudy of each concept in the second block of restudy trials (based on outcomes of prior research suggesting that participants would use about two trials per item to achieve one correct recall in the successive relearning condition). On average, participants completed Session 2 in 19 min (SE=2).

Sessions 3–6 Participants returned the following week for Session 3 (see Table 3). For all four groups, the first part of Session 3 was identical to Session 2. Once the practice phase for the concepts from the first unit was complete, the concept set to be practiced from the second pair of units was then introduced. For all four groups, the procedure for the second unit was the same as for the first unit in Session 1 (initial study followed by the appropriate practice for that group). Session 4 took place approximately 2 days later. For all four groups, the first part of Session 4 was identical to Session 2. The second part of Session 4 was also identical to Session 2, except that practice involved the concepts from the second unit. Concepts from the second unit were then practiced again in Session 5 and 6. Session 6 took place approximately two days before a course exam was administered in the students' class (see Table 3). On average, participants completed Sessions 3–6 in 32, 16, 8, and 7 min, respectively (SEs=1–2).

Session 7 Session 7 took place the following week and began with a cued recall test over all 32 concepts from the first two pairs of units (including the two units that had been practiced and the other two units assigned to the baseline control condition). On each trial, the concept was presented along with a text field prompting students to type in the definition of the concept.

After the cued recall test, the concept set to be practiced from the third pair of units was introduced. First, participants were given instructions about the tasks they would be completing, as in Session 1. Note that each participant was switched to a different practice condition for Sessions 7–12 (see Table 1). Thus, the instructions and the remainder of the procedure for each participant in Session 7 were the same as for participants in that condition during Session 1, except that the materials included the concepts from the third and fourth pairs of units. On average, participants completed Session 7 in 30 min (SE=2).

Session 8–12 The instructions and the procedure for each participant in Session 8–12 were the same as for that condition during Sessions 2–6, except that the materials included the

concepts from the third and fourth pairs of units. Session 12 took place approximately 2 days before a course exam over the third and fourth pairs of units was administered in the students' class (Table 3). On average, participants completed Sessions 8–12 in 13, 35, 19, 11, and 9 min, respectively (SEs=1–2).

Sessions 13–14 Session 13 took place at the beginning of the following week and began with a cued recall test over all 32 concepts from the first two pairs of units. All participants then completed a cued recall test over all 32 concepts from the last two pairs of units. Approximately 3 weeks later, participants completed another cued recall test for the 32 concepts from the last two pairs of units. At the end of Session 14, all participants completed a short battery of auxiliary measures (these outcomes are not of primary interest here and thus will not be discussed further). On average, participants completed Sessions 13–14 in 31 and 19 min, respectively (SEs=1–2).

Scoring We describe the scoring procedure used for both experiments here. All recall responses generated during practice and final test sessions were scored, 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. Partial credit was given for responses that included some but not all of the correct meaning of the definition. Given the sizeable number of recall responses to be hand-scored across experiments (23,979 responses in Experiment 1 and 22,420 responses in Experiment 2), multiple raters were trained to complete the scoring. 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 (rs$\geq$0.88), each remaining protocol was then scored by one of the trained raters.

Participants also granted permission for the researchers to examine their course exam performance. For each of the two course exams, we identified the subset of questions that tapped learning of the 32 experimental concepts. For each participant, we then computed the percentage of questions answered correctly for the subset of items that had been assigned to the practice condition and for the subset of items assigned to the baseline control condition for that participant.

Results and Discussion

Data were excluded for eight participants who did not complete all experimental sessions and for three participants who failed to comply with task instructions. Course exam data were not available for two participants who missed the exam in class.

To revisit, the primary purpose of the complex design was to allow all students to experience the RMF condition for one of the two course exams. An unavoidable consequence of the design was that the structure of the data set did not afford conventional omnibus analyses of variance, so we report outcomes for paired comparisons of greatest interest.

Additionally, the rich data set included a large number of possible variables to analyze, given that we collected data for numerous behavioral measures (including performance and time measures for both cognitive and metacognitive indices) in each of the 14 sessions. To avoid overwhelming readers with extensive results and analyses, our general analytic approach involved

collapsing across groups when possible and reporting only the results of primary interest in the "Results and Discussion" section. Descriptive statistics separated by groups are reported in Table 6 in the Appendix.

Performance on Course Exams Mean performance on course exam questions tapping practiced concepts and baseline control concepts is reported in Fig. 3. Results for each condition are collapsed across the two exams and across groups. Thus, all participants contributed values in the successive relearning condition and the corresponding baseline control condition, and each participant also contributed values to either the self-regulated practice condition or the restudy-only condition (and the corresponding baseline control condition).

Of greatest interest, did successive relearning enhance performance on actual course measures? The answer is decisively yes: Successive relearning improved performance by more than a letter grade based on the grading metric used by the instructor (84 % versus 72 %) for course exam questions tapping practiced concepts versus baseline control concepts, $t(66)=3.92$, $p<0.001$, Cohen's $d=0.65$. This improvement is particularly noteworthy not only because it was obtained on an authentic outcome measure but also because it reflects transfer. Whereas the practice tests involved in successive relearning were cued recall for definitions, all exam questions were multiple choice. Furthermore, as will be described in the "Exploratory Analyses of Potential Moderators in Experiments 1–2" section after "Experiment 2," none of the exam questions involved verbatim statements of the definitions practiced during successive relearning.

Of secondary interest, outcomes in the restudy-only condition demonstrate that retrieval practice is an active ingredient in successive relearning (i.e., spacing without testing is not as potent). Exam performance for restudied versus baseline control concepts did not significantly differ [Fig. 1; $t(33)=1.09$, $d=0.28$], indicating that spaced restudy alone was not sufficient to produce meaningful gains. Furthermore, exam performance was greater for successively relearned concepts versus restudied concepts [$t(33)=2.31$, $p=0.014$, $d=0.45$], further confirming that retrieval practice is a critical ingredient of successive relearning.

Also of secondary interest, to what extent were students able to effectively regulate their own learning in a flashcard program that afforded but did not control successive relearning? The advantage for practiced versus baseline control concepts was not significant in the self-

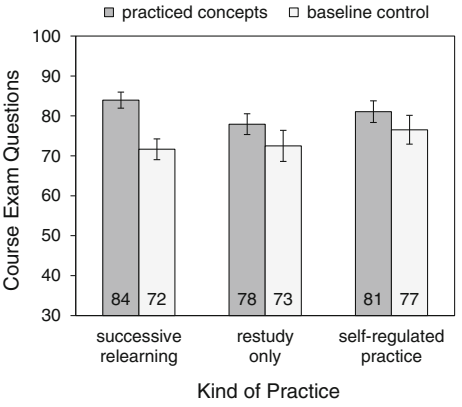

Fig. 3 Mean percent correct on course exam questions tapping practiced concepts versus baseline control concepts as a function of practice condition in Experiment 1. *Error bars* are between-subject standard errors of the means

regulated practice condition, $t(32)=0.93$, $d=0.25$, although the lack of a significant effect may be partly due to somewhat elevated exam performance for baseline control items in this condition (relative to performance for the baseline control items in the other two conditions).

Performance on Follow-Up Cued Recall Tests Mean performance on the cued recall tests administered 3 and 24 days after the course exams is reported in Fig. 4. Consistent with the concern of many instructors that students quickly forget what they have learned following an exam, performance for baseline control concepts was very poor after a lapse of only 3 days. Note that performance on the cued recall test is not directly comparable to performance on the multiple-choice course exam, and thus, the degree of forgetting cannot be estimated directly. Nonetheless, the absolute level of cued recall observed indicates limited accessibility of concept definitions that were learned via business as usual.

In contrast, successive relearning was effective for enhancing retention, with significantly greater performance for practiced versus baseline control concepts on the 3-day test [$t(67)=20.42$, $p<0.001$, $d=2.90$] and on the 24-day test [$t(67)=17.31$, $p<0.001$, $d=2.37$]. Performance on both tests was also greater for practiced versus baseline control concepts in the restudy-only condition [$t(33)=8.28$, $p<0.001$, $d=1.30$; $t(33)=7.59$, $p<0.001$, $d=1.07$]. However, at both time points, performance for practiced concepts was greater after successive relearning than after restudy only [$t(33)=4.08$, $p<0.001$, $d=0.83$; $t(33)=3.59$, $p<0.001$, $d=0.69$]. Finally, performance in the self-regulated practice condition was greater for practiced versus baseline control concepts on the 3- and 24-day tests [$t(34)=9.70$, $p<0.001$, $d=1.77$, and $t(33)=7.45$, $p<0.001$, $d=1.23$, respectively].

Experiment 2

Given that Experiment 1 provides the first empirical demonstration that successive relearning can produce meaningful improvements in an authentic educational context, an important goal of Experiment 2 was to replicate the key finding that successive relearning enhances course performance and long-term retention over business as usual (regarding the importance of replicating novel findings, see Pashler and Harris 2012).

Additionally, Experiment 2 was designed to provide two key extensions. First, in Experiment 1, the initial learning session for the successive relearning condition involved

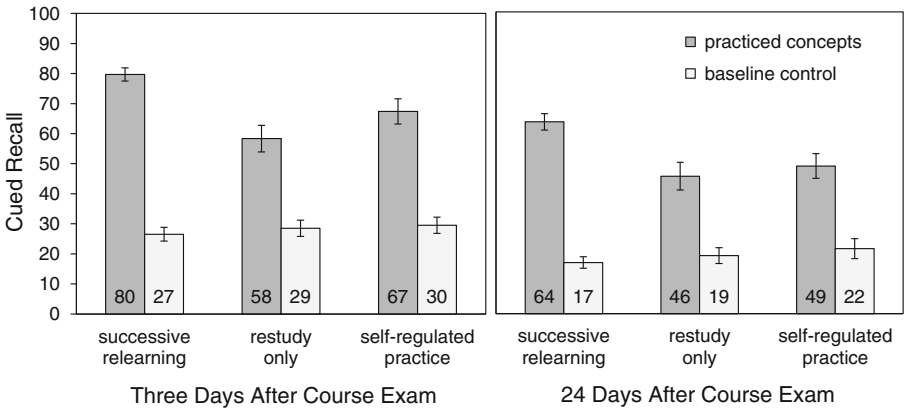

Fig. 4 Mean percent correct on cued recall tests for practiced concepts versus baseline control concepts as a function of practice condition in Experiment 1. *Error bars* are between-subject standard errors of the means

practicing until concept definitions were correctly recalled three times. However, recent research has shown that practicing beyond one correct recall during initial learning has minimal benefit for long-term retention if relearning sessions take place between initial learning and the final test (for a review, see Rawson and Dunlosky 2012b), as is the case with the successive relearning schedules of interest here. Furthermore, achieving a higher criterion level during initial learning (versus terminating practice after the first correct recall) requires a non-trivial amount of additional time, and this cost is not completely recouped via faster relearning in subsequent sessions. Thus, Experiment 2 included two successive relearning groups who only practiced to one correct recall during initial learning, to address the extent to which successive relearning yields meaningful improvements in course performance and long-term retention when the initial learning session involves a more efficient schedule of practice.

Importantly, one of these two groups did not complete their practice sessions under controlled laboratory conditions. Instead, students were given a flashdrive with the virtual flashcard program and were instructed to complete the learning sessions on their own outside of the lab. The purpose of this group was to examine the extent to which meaningful improvements in exam performance and retention could be achieved when successive relearning was likely completed with less fidelity than in the laboratory and hence under conditions that more realistically capture those in which students would be engaging in successive relearning while studying (e.g., in a noisy or distracting environment, with lapses in attention from talking or texting on cell phones, and so on).

Methods

Participants and Design Participants included 104 undergraduates at Kent State University who were enrolled in large sections of Introductory Psychology taught by the third author. All participants received monetary compensation for their participation. Of the 103 students who provided demographic data, 81 % were female, 82 % were White, and mean age was 18.7 (range, 16–34 years). Experiment 2 involved a 2(concept set)×4 (group) mixed-factor design. Concept set (practiced versus baseline control) was a within-participant manipulation. Students were randomly assigned to one of four groups (three different successive relearning groups and a self-regulated practice group, n=26 in each group), which are explained in more detail below.

Materials and Procedure Materials included four sets of eight concepts used in Experiment 1 (including concepts from the units on classical and operant conditioning, encoding, and retrieval from memory). Once again, these were the actual concepts and definitions from the instructor's course materials that students were expected to learn in the class, and performance on course exam questions tapping these concepts (i.e., from Exam 2) was used as an outcome measure. Questions on Exam 2 were the same as those in Experiment 1.

In all four groups, the schedule of practice and test sessions was the same as Experiment 1 for the first and second pairs of units (units A–D in Table 3). Thus, all participants completed six practice sessions (Sessions 1–6 in Table 3) and two test sessions (Sessions 7 and 13 in Table 3).[2] The procedure for practice sessions in the *self-regulated practice* group was the same as in Experiment 1, except that the idea-unit judgment option was removed from the

[2] Data collection for Experiment 1 occurred during a fall semester and in a spring semester for Experiment 2. Due to differences in the second half of the course schedule for these two semesters due to Spring Break and to some modifications the instructor made to course content later in the semester, we did not implement the full 14-session schedule in Experiment 2.

menu of task options (to more closely mimic available flashcard software, which does not include support for monitoring). The basic procedure for practice sessions in all three of the successive relearning groups was also the same as in Experiment 1. In one of these groups, each concept was dropped from further practice in the initial learning session after it had been correctly recalled three times, as in Experiment 1.

In the other two groups, each concept was dropped from the initial practice session after it had been correctly recalled once. One of these groups completed all practice sessions in the lab, whereas the other group completed all practice sessions outside of the lab. Specifically, after completing informed consent during Session 1, students who were assigned to the *unsupervised* group were given a packet that included an instruction sheet, a log sheet, and a flashdrive with the virtual flashcard program used in the other two successive relearning groups. The instruction sheet provided basic information about how to start the program and a brief recommendation to complete each session in a quiet place during a time in which they would not be distracted or interrupted. Participants were also informed that at the end of each session, the program would generate a text file that they were to send to the experimenter via email attachment (cf. submitting a homework assignment). The purpose of requesting that data be sent was to provide an ongoing indicator about whether participants were complying with the session schedule. Each participant's data were also saved internally within the program, not accessible to the participant but retrievable by the experimenter when the participant returned the flashdrive upon returning to the lab for Session 7, to ensure the integrity of the data submitted to analyses.[3] The log sheet in the packet indicated the date on which they should complete each of the six practice sessions (which conformed to the schedule shown for Sessions 1–6 in Table 3), along with boxes for students to fill in the date they actually completed each session and to check off that they sent each data file to the experimenter. The log sheet also included reminders about the two dates they would need to return to the lab to complete the two test sessions and instructed them to bring their packet and flashdrive back with them when they returned. The experimenter briefly reviewed the instruction sheet and the log sheet with participants before they left the lab.

As in Experiment 1, all participants completed final cued recall tests in the lab 3 and 24 days after the in-class exam (Exam 2), administered with the same procedure as in Experiment 1. On average, participants completed Sessions 1–8 in 40, 23, 37, 22, 11, 9, 18, and 18 min, respectively (SEs=1–2; mean values for Sessions 1–6 are based on times for the three groups who completed the sessions in the lab).

Results and Discussion

Data were excluded for five participants who did not complete all experimental sessions and for six participants who failed to comply with task instructions. Course exam data were not available for five participants who missed the exam in class.

Performance on Course Exams Mean performance on course exam questions tapping practiced concepts and baseline control concepts is reported in Fig. 5. A 2(concept set)×4

[3] Based on analysis of the session dates in the program-internal data files, participants showed high levels of compliance with the intended schedule for Sessions 1–6, with an average of 2.4, 4.9, 2.2, 4.7, and 2.2 days (SEs=0.2-0.4) between consecutive sessions (cf. the intended lags of 2, 5, 2, 5, and 2 days, as shown in Table 3).

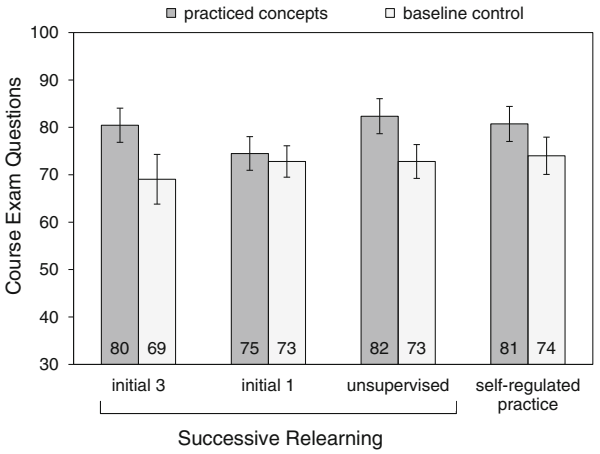

Fig. 5 Mean percent correct on course exam questions tapping practiced concepts versus baseline control concepts as a function of practice condition in Experiment 2. *Error bars* are between-subject standard errors of the means. Initial 3=criterion of three correct recalls during the initial practice session. Initial 1=criterion of one correct recall during the initial practice session

(group) ANOVA yielded only a significant main effect of concept set $F(1,84)=7.57$, $MSE=310.27$, $p=0.007$, $\eta_p^2=0.08$ (other Fs<1). Replicating the key outcome of Experiment 1, successive relearning involving three correct recalls during initial practice again improved exam performance over baseline by about one letter grade on the instructor's grading scale [80 versus 69 %, $t(18)=1.64$, $p=0.060$, $d=0.59$]. Importantly, a similar effect was obtained when successive relearning was unsupervised [$t(22)=2.01$, $p=0.029$, $d=0.55$], establishing the potency of successive relearning even when implemented under conditions with potentially lower fidelity than when administered in the lab. Although the interaction was not significant, the advantage of practiced concepts over baseline control concepts was numerically weaker in the supervised successive relearning group involving an initial learning criterion of one correct recall [$t(20)=0.39$, $d=0.11$]; this trend was unexpected, given that prior research has reported comparable levels of performance for successive relearning groups involving initial criterion levels of one versus three correct recalls (Rawson and Dunlosky 2011, 2012a, b). The current outcome may simply reflect measurement error, given that an effect was obtained in the unsupervised group (which also involved an initial learning criterion of one correct recall) and given that both groups showed sizeable benefits on the cued recall measures reported next. Nonetheless, additional investigation of initial criterion effects in future research may be useful. More important, the overall pattern across groups indicates that successive relearning can enhance performance on authentic course exams. Notably, the effect of successive relearning was obtained even when implemented under conditions with potentially lower fidelity than when administered with supervision in the lab. Finally, the advantage for practiced versus baseline control concepts in the self-regulated practice condition was again modest [$t(24)=1.29$, $p=0.105$, $d=0.35$].

Performance on Follow-Up Cued Recall Tests Mean performance on the cued recall tests is reported in Fig. 6. For cued recall 3 days after the exam, a 2 (concept set)×4 (group) ANOVA yielded only a significant main effect of concept set, $F(1,89)=377.78$, $MSE=265.06$, $p<0.001$, $\eta_p^2=0.81$ (other Fs<1). The performance for baseline control concepts shown in Fig. 6 again demonstrates that students quickly forget what they have learned following an

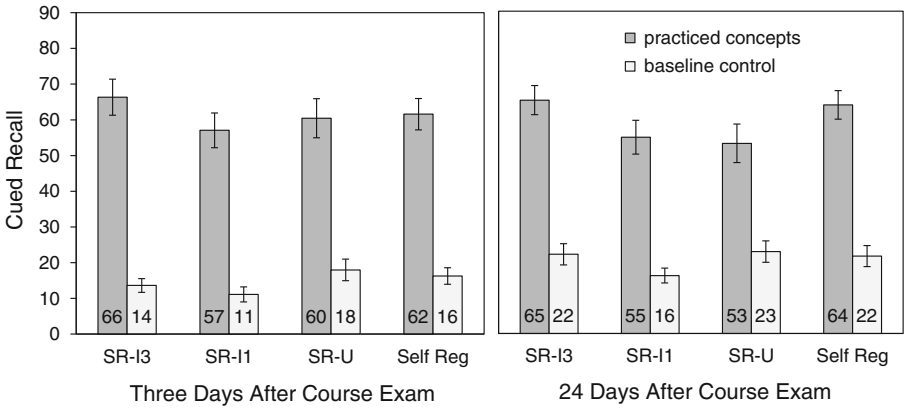

Fig. 6 Mean percent correct on cued recall tests for practiced concepts versus baseline control concepts as a function of practice group in Experiment 2. *Error bars* are between-subject standard errors of the means. SR-I3 and SR-I1=successive relearning with three or one correct recalls during the initial practice session, SR-U=unsupervised successive relearning, and Self Reg=self-regulated practice

exam. In contrast, successive relearning was effective for enhancing retention beyond the exam in all three groups (d=2.72 with an initial learning criterion of three, d=2.46 with an initial learning criterion of one supervised, or d=1.88 unsupervised). Students' self-regulation of spaced practice also enhanced retention (d=2.47). The pattern was highly similar 24 days after the course exam; a 2(concept set)×4 (group) ANOVA yielded only a significant main effect of concept set, $F(1,89)=309.87$, $MSE=220.84$, $p<0.001$, $\eta_p^2=0.78$ (other Fs<1.81). Successive relearning meaningfully enhanced retention over baseline control in all three groups (d=2.55 with an initial learning criterion of three, d=1.91 with an initial learning criterion of one supervised or d=1.35 unsupervised), as did students' self-regulation of spaced practice (d=2.31).

Exploratory Analyses of Potential Moderators in Experiments 1–2

The current research was not designed to systematically identify moderators of the effects of successive relearning, but it does afford exploratory analyses examining two potential moderators of educational relevance that have been underinvestigated in prior research. The first potential moderator concerns students' achievement level. None of the few previous studies of successive relearning examined effects as a function of individual differences in student achievement (nor on any other individual differences variable). More generally, in the sizeable research on testing effects (the parent literature to successive relearning), the most notable gap concerns the absence of examination of testing effects as a function of individual differences in student achievement (to our knowledge, the only prior study to do so was reported by Spitzer 1939).

To explore this issue, we accessed students' grades on the first course exam, which was completed prior to the beginning of the experiment. Given that the instructor used the same Exam 1 across semesters in Experiments 1–2, we combined samples for these analyses (i.e., including the successive relearning group in Experiment 1 and all three successive relearning groups in Experiment 2). From the overall sample of 132 students, we selected two extreme groups including the 48 students with the lowest grades on Exam 1 (range 45–77 %) and the 48 students with the highest grades on Exam 1 (range 89–105 %; note that the instructors'

exams included extra credit questions). Outcomes for these two subsets of lower-performing and higher-performing students are reported in Table 4. Validating the differences in achievement for these two subsets, higher performers significantly outscored lower performers on overall grade for Exam 2, $t(92)=7.69$, and on ACT Composite scores, $t(40)=4.04$ (ACT scores were only available for Experiment 2).

As is evident from inspection of the pattern of outcomes on the primary dependent variables, successive relearning was similarly profitable for both lower-performing and higher-performing students. We conducted a 2(pre-experimental performance level: higher versus lower)×2 (concept set: practiced versus baseline) ANOVA for each of the three outcome variables. Although the main effects of performance level and concept set were significant in all three analyses (all Fs>11.83), none of the interactions were significant (Fs<1.56). The amount of time required to engage in successive relearning also did not differ as a function of performance level (based on practice time per concept reported in Table 4), $t(91)=0.09$. With that said, although lower and higher performers required the same amount of time to achieve the same level of successive relearning, they obviously did not achieve the same level of exam performance or subsequent cued recall. One possibility is that increasing the amount of successive relearning for lower performers (by increasing the number of correct recalls within each session or the number of relearning sessions) may help them close the performance gap. Nonetheless, the overall pattern provides the first evidence suggesting that the added benefit of successive relearning over business as usual does not vary considerably as a function of achievement level.

The second potential moderator concerns the level of learning tapped by the instructor's questions on the course exams. Within the set of exam questions on the concepts included in the experimental materials, 42 % were primarily memory-based questions. Approximately half of these directly tapped memory for concept definitions, although none of these involved recognition of verbatim definitions. Rather, these questions involved identifying concept names based on paraphrases of the practiced definitions. The other half involved distant and/or incomplete paraphrases of the definitions. Examples of memory-based

Table 4 Descriptive statistics as a function of Exam 1 performance for students in the successive relearning groups, collapsed across Experiments 1–2

	Lower performers (Exam 1<78 %)	Higher performers (Exam 1>88 %)
Overall grade on Exam 1 (%)	66.7 (1.2)	94.8 (0.6)
Overall grade on Exam 2 (%)	74.9 (1.5)	90.3 (1.3)
ACT composite (Experiment 2 only)	20.0 (0.4)	23.3 (0.8)
Exam performance (%)		
Practiced concepts	75.1 (2.5)	87.8 (2.2)
Control concepts	66.4 (3.1)	78.9 (2.2)
Cued recall 3 days after exam (%):		
Practiced concepts	59.9 (3.6)	79.4 (2.6)
Control concepts	12.3 (1.7)	29.4 (2.5)
Cued recall 24 days after exam (%):		
Practiced concepts	51.6 (3.6)	69.6 (2.9)
Control concepts	13.3 (1.6)	25.6 (2.3)
Practice time per concept (in minutes)	7.5 (0.3)	7.5 (0.3)

questions are included in the top portion of Table 5, along with the corresponding cued recall prompts used during relearning to illustrate how practice and criterion tests differed in format, wording, and response requirements. The remaining 58 % of the exam questions primarily tapped comprehension of the concepts, requiring inferencing and/or application to answer correctly. Examples of comprehension-based exam questions are included in the bottom portion of Table 5.

To what extent might the effects of successive relearning depend on the degree of transfer involved in these different kinds of question? To explore this issue, we collapsed successive relearning groups across Experiments 1–2 (subdividing the relevant exam questions into memory-based and comprehension-based subsets resulted in a relatively small number of questions contributing to each cell for each participant, so we collapsed samples to increase the stability of the somewhat noisy estimates). A 2(concept set: practiced versus control)×2 (question set: memory-based versus comprehension-based) ANOVA revealed that both main effects and the interaction were significant [concept set, $F(1,128)=15.99$, $p<0.001$, $\eta_p^2=0.11$; question set, $F(1,128)=30.95$, $p<0.001$, $\eta_p^2=0.20$; interaction, $F(1,128)=5.81$, $p=0.017$, $\eta_p^2=0.04$]. The benefit of successive relearning over baseline was greater for memory-based questions [88 % versus 75 %, $t(128)=4.71$, $p<0.001$, $d=0.53$] than for comprehension-based questions [75 % versus 70 %, $t(128)=4.71$, $p<0.001$, $d=0.21$]. Although both kinds of exam questions involved some degree of transfer from the information overtly tested via cued recall during successive relearning, the degree of transfer is arguably greater for the comprehension-based questions. Nevertheless, this pattern should be interpreted with caution for various reasons, including because the number of questions contributing to each cell is relatively small, because most concepts were not equally evaluated by both kinds of questions, and because the similarity of the psychometric properties of the two subsets of questions is unknown. The outcomes do suggest that exploring the effects of successive relearning as a function of degree of transfer on criterion tests will be an important direction for further research.

General Discussion

Successive relearning combines two of the most potent learning strategies known to date—testing and spacing—under conditions in which they are particularly effective (i.e., testing that involves multiple successful retrievals and spacing of sessions across days). Despite the obvious implications for enhancing student learning, only six prior studies have examined successive relearning and none of these involved a successive relearning intervention implemented in an authentic educational context. The two experiments reported here provide the first demonstration that successive relearning can enhance performance on authentic course outcomes and can substantially improve retention after course exams. This work also contributes to successive relearning's two parent literatures on spacing and testing effects, in that only a handful of classroom studies exist among the several hundred experiments that have been reported in these two literatures.

Our conclusion that successive relearning enhances student learning is bolstered by converging evidence from more than one measure (multiple-choice exams and cued recall) and by including more than one comparison condition (baseline control and restudy only). The advantage of successive relearning over the baseline control condition is arguably of greatest interest, to the extent that the baseline condition reflects the level of learning that students achieve with whatever "business-as-usual" study strategies and study schedules they employ on their own. Of course, business as usual might involve successive relearning

Table 5 Sample questions from students' actual course exams

Memory-based questions
In the game show Jeopardy! Contestants are tested on general information. The type of memory used to answer these kinds of questions is _____.
a. Episodic
b. Procedural
c. Semantic
d. Working
(What is semantic memory? General knowledge, regardless of when or where you learned it.)
A stimulus presented to a person or animal that decreases the probability of a particular response is known as _____.
a. Punishment by removal
b. Positive reinforcement
c. Punishment by application
d. Negative reinforcement
(What is punishment? When an event or stimulus presented after a behavior makes the behavior less likely to occur again.)
The Flynn effect is the term used to describe the fact that
a. The average person today has more acquired knowledge than the average person of 20 years ago
b. The average person today is not as good at problem solving as the average person of 20 years ago
c. The average person today is a better problem solver than the average person of 20 years ago
d. The average person today has less acquired knowledge than the average person of 20 years ago
(What is the Flynn effect? The observed increase in raw scores on IQ tests across generations of test-takers.)
It has been shown time and time again that the way that data are presented to us has a significant impact on our ultimate opinions and decisions. This is called the
a. Framing effect
b. Representativeness effect
c. Belief perseverance effect
d. Drive-reduction theory effect
(What are the effects of framing? How an issue is posed can significantly affect decisions and judgments.)
Comprehension-based questions
The fact that most individuals feel even more strongly about their previously preferred candidate after watching a political debate is an example of
a. The confirmation bias
b. The representativeness heuristic
c. The availability heuristic
d. The reverse scored bias
You show your little sister a big chunk of fudge. While she watches, you cut the fudge into 6 pieces. She gets very excited because she believes now that there is more fudge. Which of Piaget's stages is your sister likely to be in?
a. Formal operation
b. Preoperational
c. Sensorimotor
d. Concrete operational
As an infant, Stephanie received many penicillin injections from her doctor. When she later saw a photographer in a white coat that was similar to the doctor's coat, she started to cry. This is an example of
a. Observational learning

Table 5 (continued)

Memory-based questions
b. Operant conditioning
c. Classical conditioning
d. Habituation
You start out using Netscape, then change to Explorer because your company demands that browsers be Microsoft products. If you have trouble with Explorer, it is most likely due to
a. Retroactive interference
b. Consolidation problems
c. Proactive interference
d. Anterograde interference

The information included in parentheses below each memory-based question includes the cued recall prompt and target response used during successive relearning that corresponds to the correct answer on the exam question, to illustrate how practice tests and criterion tests differed in format, the wording of the information explicitly stated, and the target response required

for some students at least some of the time, given that students frequently report engaging in retrieval practice when studying on their own (for recent survey studies, see Hartwig and Dunlosky 2012; Karpicke et al. 2009; Kornell and Son 2009; Wissman et al. 2012). To the extent that business as usual for student participants in the current experiments involved repeated retrieval practice, the comparison of successive relearning to baseline control here may somewhat underestimate the effects of successive relearning. However, the substantial differences in performance on the delayed cued recall tests for baseline control versus successive relearning would suggest that students do not spontaneously engage in much successive relearning on their own.

A related question concerns the extent to which successive relearning may have had spillover effects on learning of the baseline control items, given that these two conditions were manipulated within-participant (rather than including a "pure" control group that engaged in no experimental practice of any sort). For instance, perhaps retrieving targets definitions while successively relearning had a deleterious effect on memory for the baseline control items. Alternatively, successive relearning may have facilitated learning of the baseline control items, a possibility suggested by research showing that taking practice tests over subsets of material can enhance learning of related material under some conditions (e.g., Chan 2009, 2010; Szpunar et al. 2008; Wissman et al. 2011). However, neither negative nor positive spillover effects of retrieval practice seem likely here, given the minimal differences in performance for baseline control items in the successive relearning (which involved retrieval) and restudy-only conditions (which did not involve retrieval) in Experiment 1. Also, note that performance for practiced items was greater following successive relearning than restudy in Experiment 1; a relatively safe assumption is that a pure control group would have done no better than the restudy group (and likely would have performed worse).

In sum, the current research provides an important first step toward establishing the power of successive relearning over business-as-usual in authentic educational contexts. This work also lays a solid foundation for future research. One key direction will be to further explore the effects of successive relearning on various kinds of criterion measures. For example, the robust advantage of successive relearning over baseline on the delayed cued recall tests observed here were likely due in part to the overlap in the processing requirements of the tests used during practice and test (both involving cued recall for the same target information), consistent with the

transfer-appropriate processing framework (cites). By comparison, the effects were smaller on the non-identical course exam questions. Outcomes of the exploratory analyses also provide suggestive evidence that the effects of successive relearning may depend on the degree of transfer required on the criterion test, with stronger effects on the memory-based questions involving near transfer than on the comprehension-based questions involving farther transfer. This overall pattern is unsurprising, given that most educational interventions show weaker effects as the degree of transfer increases—indeed, far transfer is arguably the holy grail in educational research.

Other key directions for further research involve establishing generalizability across learners and materials. Concerning learners, outcomes of the exploratory analyses suggested that successive relearning will be similarly beneficial for learners of different achievement levels, but more systematic exploration of this pattern would be useful, as well as examination of other potentially important individual differences factors (e.g., level of prior knowledge, motivation). Additionally, the majority of participants in the current research were female. Given prior research showing that females are more conscientious and that conscientiousness is related to academic outcomes (e.g., Corker et al. 2012; Vecchione et al. 2012), examining the effects of successive relearning as a function of gender would be useful (particularly when successive relearning is unsupervised, as in Experiment 2). Extending the pattern to younger learners would also be of interest.

With respect to generalizing to other materials, a reasonably safe assumption is that the effects shown here with key concepts in psychology would generalize to key concepts from other topics or domains (e.g., biology, chemistry). Recent outcomes from laboratory research also suggest that the benefits of successive relearning will scale down to simpler kinds of material, such as foreign language vocabulary (Vaughn et al. 2013). By contrast, the extent to which successive relearning can be scaled up for use with materials that are more complex than key concepts is less clear. In related work, Fritz, Morris, Bjork, Gelman, and Wickens (2000; see also Howe 1970; Kay 1955) demonstrated *failure of further learning* across practice sessions for text material. Individuals completed another practice recall test with restudy in each of the next 2 weeks, and then completed a final recall test in Week 4. Recall performance showed minimal improvement across sessions. The key difference between this method and successive relearning is that individuals were given only one practice recall trial in each session versus practicing to criterion (i.e., continuing retrieval practice trials until all text information could be recalled). However, this highlights a potentially critical limitation of successive relearning for lengthy text material—namely, it is not feasible for students to engage in retrieval practice to criterion with lengthy text, given the time demands and the persistently low levels of recall typically observed for text material. One solution may be to extract the most important ideas, facts, or concepts from the text to submit to successive relearning. With that said, no learning strategy is applicable to every conceivable topic or kind of material, so there will necessarily be some limits to the situations in which successive relearning can be utilized. Nonetheless, successive relearning is a powerful, multi-purpose tool for many learning goals.

Successive relearning is an important technique not only because of its potent effects on student learning but also because it is a strategy that students can use on their own outside of class with minimal demands on instructors. By comparison, in prior studies showing positive effects of practice testing on course-related outcomes, the interventions involved practice tests that were developed by instructors, researchers, or textbook publishers (e.g., Daniel and Broida 2004; Lyle and Crawford 2011; McDaniel et al. 2012; Shapiro and Gordon 2012). Another practical constraint on the use of instructor-provided practice tests concerns the extent to which students receive timely feedback on response accuracy, which is an important component of test-enhanced learning. In part due to this constraint, most of the

practice tests in prior interventions involved multiple-choice tests. However, multiple-choice tests are often less effective than retrieval-based test formats for enhancing learning (e.g., Butler and Roediger 2007); McDaniel et al. 2007). A final practical constraint on the use of instructor-provided practice tests is that they are usually one-size-fits-all (e.g., involving a fixed number of questions) rather than tailored for individual students based on their level of learning, whereas successive learning as conducted in the present intervention studies inherently tailors learning for each student and for the difficulty of each item. Of course, instructor-provided practice tests and successive relearning are companion approaches and not mutually exclusive techniques. But not all instructors will provide students with practice tests, and even in cases in which they do, students will benefit from supplementing with successive relearning.

With that said, we note an important caveat to our statement above that students can use successive relearning on their own outside of class with minimal demands on instructors. Although students can use successive relearning on their own in principle, we suspect that many students do not use the strategy or may use it sub-optimally in practice (Wissman et al. 2012). On the latter point, outcomes from the self-regulated practice groups are relevant, as the self-regulated practice method instantiated here represents a less constrained variant of successive relearning in which the key elements were afforded but not required. Exhaustive analysis of the behavioral profiles of self-regulated practice in each session is beyond the scope and primary purpose of this paper, but we briefly touch on a couple summary outcomes. First, the majority of learners in the self-regulated practice group engaged in some amount of retrieval practice and restudy in each session. On average in Experiments 1 and 2, participants restudied one or more items in 97.8 % and 99.4 % of sessions, respectively, and self-tested on one or more items in 99.3 % and 94.4 % of sessions, respectively. However, self-regulated practice and successive relearning groups differed in the degree of self-testing and restudy. Across practice sessions in Experiment 1, successively relearned items were tested and restudied 10.4 times each on average, whereas participants elected to test 6.0 and 7.5 times per item during self-regulated practice. Additionally, learners in Experiment 1 spent less time practicing overall in the self-regulated versus successive relearning group (mean sum of minutes spent per item across all sessions, 4.3 versus 8.1, respectively). These outcomes suggest that learners sub-optimally regulated their successive relearning, which may partially explain why the effect sizes in Experiment 1 tended to be smaller with self-regulated practice than in the successive relearning group. By comparison, effect sizes were more comparable in Experiment 2, and behavioral profiles were also more comparable. Across practice sessions, successively relearned items were tested and restudied 9.8 times each on average, and participants elected to test 8.9 and restudy 13.6 times per item during self-regulated practice. Overall practice time was also more comparable for self-regulated practice versus the successive relearning groups (5.9 versus 6.6 min per item). Taken together, these patterns suggest that students may not always optimally regulate successive relearning, and the benefits for learning may depend on the extent to which they do.

Thus, instructors may still play a vital role by teaching students about successive relearning and coaching them on effective use of the strategy. To effectively engage in successive relearning, students need to identify the to-be-learned concepts, engage in full-blown retrieval attempts, accurately judge when their responses are correct, and manage the schedule of practice within and across sessions. Students will likely profit from guidance from instructors on each of these key components of successive relearning. First, instructors could spend a small amount of time in class to teach students how to effectively engage in successive relearning. Second, instructors could help students identify the most important concepts to submit to successive relearning (e.g., via study guides). Third, given that the implementation of successive relearning will require planning several study sessions in advance, many students would

likely profit from support or suggestions for time management (e.g., a handout with a recommended schedule of study sessions prior to a course exam; for an example, see Rawson and Dunlosky 2013).

Equipping students with effective learning strategies that they can implement outside of class is timely and important, particularly given increasing emphasis on the use of flipped classrooms and other pedagogical approaches in which classroom time is largely reserved for active processing activities that will engage students and promote critical thinking, problem solving, and comprehension. This pedagogical strategy can be quite effective (Deslauriers et al. 2011), but its potency relies heavily on students learning foundational facts and information on their own outside of class. Not only will successive relearning be compatible with this and other pedagogical approaches, it is also widely applicable to many different kinds of material, topics, and content domains and hence offers a powerful technique for supporting robust learning and enhancing student achievement.

Appendix

Table 6 Performance on outcome measures in each condition for each group, Experiment 1

	Group 1 (SR-Self)	Group 2 (SR-RO)	Group 3 (Self-SR)	Group 4 (RO-SR)
Course exam performance				
Successive relearning (SR)				
Practiced concepts	82 (5)	85 (3)	84 (4)	84 (4)
Control concepts	79 (5)	66 (5)	67 (5)	74 (5)
Self-regulated practice (self)				
Practiced concepts	80 (4)		82 (4)	
Control concepts	78 (5)		75 (5)	
Restudy only (RO)				
Practiced concepts		77 (4)		79 (4)
Control concepts		69 (7)		75 (5)
Cued recall 3 days after course exam				
Successive relearning (SR)				
Practiced concepts	80 (5)	75 (5)	84 (4)	79 (4)
Control concepts	28 (5)	30 (5)	24 (4)	25 (4)
Self-regulated practice (self)				
Practiced concepts	64 (6)		71 (6)	
Control concepts	26 (4)		33 (4)	
Restudy only (RO)				
Practiced concepts		48 (7)		67 (5)
Control concepts		21 (4)		35 (3)
Cued recall 24 days after course exam				
Successive relearning (SR)				
Practiced concepts	65 (6)	63 (6)	63 (5)	63 (6)
Control concepts	26 (5)	25 (4)	10 (2)	11 (2)

Table 6 (continued)

	Group 1 (SR-Self)	Group 2 (SR-RO)	Group 3 (Self-SR)	Group 4 (RO-SR)
Self-regulated practice (self)				
Practiced concepts	46 (6)		52 (6)	
Control concepts	12 (3)		30 (5)	
Restudy only (RO)				
Practiced concepts		32 (6)		56 (6)
Control concepts		8 (2)		28 (4)

Acronyms in parentheses below the group numbers indicate which two conditions were included in that group and the order in which they were completed; *SR*=successive relearning condition, *Self*=self-regulated practice condition, and *RO*=restudy only condition. Mean values are percentages, and values in parentheses are standard errors of the mean

For course exam performance in the SR condition, a 2(concept set: practiced vs. control)×4 (groups: 1, 2, 3, 4) ANOVA yielded a significant main effect of concept set but no main effect of group and no interaction (Fs<1.30). 2(concept set)×2 (group) ANOVAs for the Self and RO conditions yielded significant main effects of concept set but no group effects or interactions (Fs<1)

For cued recall after 3 days in the SR condition, a 2(concept set)×4 (group) ANOVA yielded a significant main effect of concept set but no group effect or interaction (Fs<1.58). The 2(concept set)×2 (group) ANOVAs for the Self and RO conditions yielded significant main effects of concept set but no other significant effects (Fs<1.51) except a main effect of group in the RO condition, $F(1,32)=7.69$, $p=0.009$

For cued recall after 24 days in the SR condition, a 2(concept set)×4 (group) ANOVA yielded a significant main effect of concept set but no group effect ($F=1.50$); the interaction was marginal, $F(1,64)=2.30$, $p=0.086$. For the Self condition, a 2(concept set)×2 (group) ANOVA yielded a significant main effect of concept set and a marginal effect of group [$F(1,32)=3.75$, $p=0.062$] but no interaction ($F=2.64$). For the RO condition, a 2(concept set)×2 (group) ANOVA yielded a significant main effect of concept set and a significant main effect of group [$F(1,32)=14.80$, $p=0.001$] but no interaction ($F=2.64$)

At both retention intervals, the effects involving group are most likely due to normative differences in the difficulty of the content for the units learned in the first half of the study (i.e., those evaluated on in-class Exam 2) versus in the second half of the study (i.e., those evaluated on in-class Exam 3). This explanation follows from comparing the level of performance for control concepts learned in the first versus second half of the study in each condition, because these items were never practiced in the context of the experiment and thus provide some estimate of baseline difficulty of the content. For example, control concept performance on the 24-day test in the Self condition was 30 % for the first half of the study (Self-SR group) versus 12 % for the second half (SR-Self group). All eight comparisons show the same numerical trend

References

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

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.

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.

Butler, A. C., & Roediger, H. L., III. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology, 19*, 514–527.

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

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

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.

Chan, J. C. K. (2010). Long-term effects of testing on the recall of nontested materials. *Memory, 18*, 49–57.

Corker, K. S., Oswald, F. L., & Donnellan, M. B. (2012). Conscientiousness in the classroom: A process explanation. *Journal of Personality, 80*, 995–1028.

Daniel, D. B., & Broida, J. (2004). Using web-based quizzing to improve exam performance: Lessons learned. *Teaching of Psychology, 31*, 207–208.

Deslauriers, L., Schelew, E., & Wieman, C. (2011). Improved learning in a large-enrollment physics class. *Science, 332*, 862–864.

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

Fritz, C. O., Morris, P. E., Bjork, R. A., Gelman, R., & Wickens, T. D. (2000). When further learning fails: Stability and change following repeated presentation of text. *British Journal of Psychology, 91*, 493–511.

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

Hartwig, M. K., & Dunlosky, J. D. (2012). Study strategies of college students: Are self-testing and scheduling related to achievement? *Psychonomic Bulletin & Review, 19*, 126–134.

Howe, M. J. (1970). Repeated presentation and recall of meaningful prose. *Journal of Educational Psychology, 61*, 214–219.

Karpicke, J. D., Butler, A. C., & Roediger, H. L., III. (2009). Metacognitive strategies in student learning: Do students practice retrieval when they study on their own? *Memory, 17*, 471–479.

Kay, H. (1955). Learning and retaining verbal material. *British Journal of Psychology, 46*, 81–100.

Kornell, N. (2009). Optimising learning using flashcards: Spacing is more effective than cramming. *Applied Cognitive Psychology, 23*, 1297–1317.

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

Kornell, N., & Son, L. K. (2009). Learners' choices and beliefs about self-testing. *Memory, 17*, 493–501.

Lyle, K. B., & Crawford, N. A. (2011). Retrieving essential material at the end of lectures improves performance on statistics exams. *Teaching of Psychology, 38*, 94–97.

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology, 19*, 494–513.

McDaniel, M. A., Wildman, K. M., & Anderson, J. L. (2012). Using quizzes to enhance summative-assessment performance in a web-based class: An experimental study. *Journal of Applied Research in Memory and Cognition, 1*, 18–26.

Mitchell, G. (2012). Revisiting truth or triviality: The external validity of research in the psychological laboratory. *Perspectives on Psychological Science, 7*, 109–117.

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

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

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.

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

Rawson, K. A., & Dunlosky, J. (2012a). Relearning attenuates the benefits and costs of spacing. *Journal of Experimental Psychology: General*. doi:10.1037/a0030498.

Rawson, K. A., & Dunlosky, J. (2013). Bang for the buck: Supporting durable and efficient student learning. *Submitted manuscript.*

Rawson, K. A., & Dunlosky, J. (2012b). When is practice testing most effective for improving the durability and efficiency of student learning? *Educational Psychology Review, 24*, 419–435.

Roediger, H. L. I. I. I., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*, 181–210.

Roediger, H. L., III, Putnam, A. L., & Smith, M. A. (2011). Ten benefits of testing and their applications to educational practice. *Psychology of Learning and Motivation, 44*, 1–36.

Shapiro, A. M., & Gordon, L. T. (2012). A controlled study of clicker-assisted memory enhancement in college classrooms. *Applied Cognitive Psychology, 26*, 635–643.

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

Szpunar, K. K., McDermott, K. B., & Roediger, H. L., III. (2008). Testing during study insulates against the buildup of proactive interference. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 34*, 1392–1399.

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

Vaughn, K.E., Rawson, K. A., & Dunlosky, J. (2013). *Bang for the buck: Successive relearning beats initial criterion level.* Paper presented at the 85th annual meeting of the Midwestern Psychological Association, Chicago, IL.

Vecchione, M., Alessandri, G., Barbaranelli, C., & Caprara, G. (2012). Gender differences in the Big Five personality development: A longitudinal investigation from late adolescence to emerging adulthood. *Personality and Individual Differences, 53*, 740–746.

Wissman, K. T., Rawson, K. A., & Pyc, M. A. (2011). The interim test effect: Testing prior material can facilitate the learning of new material. *Psychonomic Bulletin & Review, 18*, 1140–1147.

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